PROPER SCORING UNDER GROUP REWARD NORMALIZATION: TARGET DISTORTION AND EXACT FINITE-GROUP UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

GRPO trains language models from several answers to one prompt and standardizes their rewards to stabilize learning. For probability answers scored by one shared random outcome, prior work found overconfidence with two proper scores, which favor honest reports, in a large-group approximation, but left open whether this failure is general or persists in finite groups. We identify the key mechanism: the shared outcome determines both the rewards and their normalization scale, which can turn a small win and a large loss into equally strong updates. We precisely characterize the favored probability with group statistics fixed and prove that every such binary reward fails in some setting; even the best available answer can reverse. We derive the exact update when each sampled answer also changes its own denominator. Centering without division preserves the mean gradient direction at fixed group size; an independent scale shared across outcomes preserves each prompt's target. Qwen2.5-3B experiments connect exact updates to neural gradients and show higher scoring loss with standardization than with centering alone under plain SGD and complementary budget controls.

1 INTRODUCTION

Machine-learning systems increasingly need to report how likely an outcome is, not merely which outcome is more likely. A probability is useful only when it reflects how often the event actually occurs: if it rains in six out of ten comparable cases, a calibrated model should report 60% (Guo et al., 2017; Jiang et al., 2021). A proper score is a reward designed so that the true probability receives the highest average reward over repeated outcomes (Gneiting & Raftery, 2007). GRPO, meanwhile, trains language models by comparing several answers sampled for the same prompt. It subtracts their mean reward and divides by their group standard deviation, usually to make training more stable (Shao et al., 2024; Guo et al., 2025). In the setting studied here, all probability answers in a group are scored against one random outcome. That outcome therefore determines both their raw rewards and the scale used to train on them. Does the processed reward still teach the true probability?

The closest prior work shows that GRPO-style standardization can produce overconfident probability reports for Brier and logarithmic rewards under a large-group approximation (Bereket & Leskovec, 2025). Other work trains or prompts language models to express calibrated confidence (Lin et al., 2022; Tian et al., 2023; Band et al., 2024), including with proper-score rewards (Damani et al., 2026), or studies GRPO objectives, finite-sample bias, and group-level credit assignment (Vojnovic & Yun, 2025; Yang et al., 2026; Zhou et al., 2026; Mishra et al., 2026). However, the large-group approximation replaces the random group statistics used by training with their averages. It therefore leaves open whether another proper score can avoid the target change and what happens in an actual finite group, where every answer helps determine the standard deviation used to update itself. The latter is difficult because the answer's reward and its normalization scale are statistically dependent. It is also unknown whether the predicted change remains after neural parameterization and controls on gradient size and effective optimizer updates. These gaps matter because a model can train stably and still learn the wrong probability.

Our key insight is that correct probability learning depends not only on which answer wins after each outcome, but also on how much it wins. Suppose rain has probability 60%, and the model can answer

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

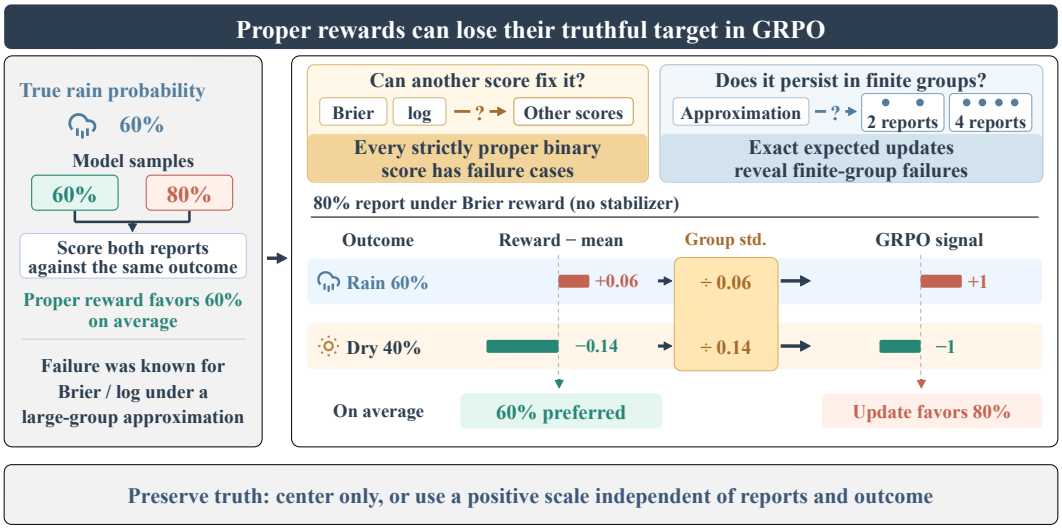


Figure 1: **Proper rewards can lose their truthful target in GRPO.** Every strictly proper binary score has failure cases under frozen population normalization; exact finite-group updates also reveal failures. The shared-outcome Brier example shows why: without a stabilizer, normalization equalizes unequal reward margins and favors 80% over the truthful 60%. Centering alone, or a common positive scale independent of reports and outcome, preserves the target.

60% or 80%. If rain occurs, 80% wins by a small amount; if it does not, 60% wins by a larger amount. These unequal margins make 60% optimal on average. Without a stabilizer, dividing by each group's standard deviation removes the magnitude difference, so each outcome contributes one equally strong vote. Because rain occurs more often, the normalized update favors 80%. More generally, a different divisor for each outcome changes the relative influence of the possible results. Centering alone keeps the reward margins, and one independent scale shared by both outcomes changes only that prompt's update strength. The central question is therefore not which proper score to choose, but whether the processing applied after scoring preserves the relative influence of the outcomes.

Our contributions are:

1. **A general theory of which probability is favored.** With group statistics fixed, we precisely characterize the favored probability and prove that every finite-valued binary proper score fails for some nontrivial report distribution, even reversing the best available answer. We show that failure extends beyond isolated examples, quantify the scoring loss, and generalize to multiclass outcomes.

2. **The exact update used by finite-group training.** We derive the exact expected update for any finite answer set and group size, retaining each answer's influence on its own scale. The two-report solution yields score-independent majority voting without stabilization and a group-size-dependent threshold with it. We also establish cases where the fresh update, with statistics fixed during back-propagation, is not the gradient of a scalar objective in ordinary logit coordinates.

3. **A safety rule and two corrections.** We prove that centering without division preserves the original mean gradient direction across prompts at fixed group size. An independent reference scale shared across outcomes preserves each prompt's target and direction; prompt-dependent scales can reweight the combined neural gradient.

4. **A theory-to-neural evidence chain.** We connect exact finite-group theory to sampled updates and Qwen LoRA gradients, and verify target changes across frozen policies and three proper scores. In controlled training, standardization has higher scoring loss than no division under plain SGD at validation-selected and fixed endpoints in all six paired seeds. The gap also persists under complementary AdamW controls that match minibatch counts or nonzero updates while fixing gradient norms.

2 Background and Motivation

A single result gives a direction, not a probability. On a rainy day, an 80% forecast is closer to the observed result than a 60% forecast, even when the true rain probability is 60%. A proper score makes the truthful report best on average across rainy and dry outcomes (Gneiting & Raftery, 2007). Brier uses squared prediction error (Brier, 1950); the logarithmic score penalizes confident mistakes more strongly (Good, 1952). These incentives also support probability-report training (Damani et al., 2026). Reward design also targets confidence ranking (Wang et al., 2026) and tool-use reliability (Xuan et al., 2026).

GRPO trains on processed rewards rather than the original scores. GRPO subtracts the group mean from each answer's reward and divides by the group standard deviation (Shao et al., 2024). Centering makes above-average rewards positive and below-average rewards negative; division makes update sizes comparable across groups. Positive signals increase an answer's probability and negative signals decrease it. GRPO analyses address objectives, sampling bias, credit assignment, and old-policy correction (Vojnovic & Yun, 2025; Mroueh, 2025; Yang et al., 2026; Zhou et al., 2026; Mishra et al., 2026; Pang et al., 2025). Normalization variants address reward variability, prompt difficulty, multiple reward components, and small rollout budgets (Liu et al., 2025; Yu et al., 2025; Salmani-Zarchi et al., 2026; Hernandez et al., 2026; Liu et al., 2026; Xiao et al., 2025; Kim, 2026); normalization can aid optimization with deterministic rewards (Ge et al., 2026). Here, one random outcome sets both the group's rewards and its scale. Each answer can also affect its own denominator, even when answers receive independent outcomes.

The division removes information needed to recover the probability. In the 60%-versus-80% example, Brier losses are 0.16 and 0.04 after rain, so 80% wins by 0.12; after no rain they are 0.36 and 0.64, so 60% wins by 0.28. This larger loss makes 80% worse on average. Centering keeps the unequal margins. Dividing each two-report group by its standard deviation, without a numerical stabilizer, turns both comparisons into equal-magnitude updates, leaving only which answer won. Such outcome-dependent weighting can change which forecast a score favors (Gneiting, 2011).

The remaining questions. Prior work finds overconfidence for Brier and logarithmic rewards under a large-group approximation (Bereket & Leskovec, 2025), leaving two questions. First, can another proper score prevent this target change? Second, what is the exact expected finite-group update when each answer affects the standard deviation used to update itself? These questions concern whether changing the score is a remedy and how actual training updates behave. The next two sections answer them in order.

3 How outcome-wise scaling changes the target

Consider a prompt asking for the probability of a binary event with truth $p \in (0, 1)$. A reward score $S : \mathcal{D} \times \{0, 1\} \to \mathbb{R}$ is defined on the continuous report domain $\mathcal{D} = (0, 1)$. Before normalization, report q has expected reward

$$U_p(q) = pS(q, 1) + (1 - p)S(q, 0). \tag{1}$$

The score is strictly proper when p uniquely maximizes U_p on $\mathcal{D}$ (Savage, 1971; Gneiting & Raftery, 2007). A model instead selects from a finite vocabulary $\mathcal{Q} = \{q_1, \ldots, q_M\} \subset \mathcal{D}$, whose operational target is the restricted Bayes set

$$\Gamma_{\mathcal{Q}}(p) := \arg\max_{q_j \in \mathcal{Q}} U_p(q_j). \tag{2}$$

When $p \notin \mathcal{Q}$, one or more tokens can maximize the expected score.

The policy draws reports $Q_1, \ldots, Q_n$ independently for a fixed prompt. One outcome $Y \sim \mathrm{Bernoulli}(p)$, independent of these reports, is shared by the group. After scoring every report against Y, training uses the detached advantage

$$\widehat{A}_i = \frac{S(Q_i, Y) - \widehat{\mu}_Y}{\widehat{\sigma}_Y + \tau}, \qquad \tau \geq 0, \tag{3}$$

where $\widehat{\mu}_Y$ and $\widehat{\sigma}_Y$ are the group mean and divisor-n standard deviation; τ is an additive stabilizer. Homogeneous groups at $\tau = 0$ receive zero update. We first freeze the current policy's population statistics to identify the target change. Section 4 restores each candidate's dependence on its own sampled center and scale.

3.1 EFFECTIVE TRUTH AND FINITE REPORTS

Let $Q \sim \nu$ be a report distribution with finite outcome-wise means μ_y and standard deviations σ_y, and put $a_y = \sigma_y + \tau > 0$. At this frozen policy, the normalized payoff is

$$G_{p,\nu}(q) = p\frac{S(q,1) - \mu_1}{a_1} + (1-p)\frac{S(q,0) - \mu_0}{a_0}. \tag{4}$$

Theorem 3.1 (Continuous effective truth)**.** For every finite-valued strictly proper binary score on $\mathcal{D}$,

$$G_{p,\nu}(q) = C_{p,\nu}U_{T_\nu(p)}(q) - K_{p,\nu}, \qquad C_{p,\nu} > 0, \tag{5}$$

where $K_{p,\nu}$ does not depend on q and

$$T_\nu(p) = \frac{pa_0}{pa_0 + (1-p)a_1}. \tag{6}$$

Hence $T_\nu(p)$ is the unique maximizer over $\mathcal{D}$. The continuous truth is preserved at an interior p if and only if $\sigma_0 = \sigma_1$.

Normalize the positive coefficients p/a_1 and $(1-p)/a_0$ in Equation (4): their outcome-one share is $T_\nu(p)$, while centering contributes only the constant $K_{p,\nu}$. Separate scales thus change the relative outcome weights, making the original proper score favor a different probability.

Corollary 3.2 (Finite-vocabulary effective target)**.** For every nonempty finite $\mathcal{Q} \subset \mathcal{D}$,

$$\arg\max_{q_j \in \mathcal{Q}} G_{p,\nu}(q_j) = \Gamma_{\mathcal{Q}}(T_\nu(p)). \tag{7}$$

If $\nu = \sum_j \pi_j \delta_{q_j}$, the frozen population softmax-logit field is given by Equation (27).

The best token changes when p and $T_\nu(p)$ fall in different decision cells. This logit identity is pointwise: both scales and $T_\pi(p)$ move with the policy. Appendix B.1 gives the derivation.

3.2 NO PROPER BINARY SCORE AVOIDS THE DISTORTION

Could another proper score equalize both scales for every policy? For two-point policies, this requires equal distances $|S(x,0) - S(z,0)|$ and $|S(x,1) - S(z,1)|$ for every pair. The score coordinates must then differ by a translation or reflection. The former makes the optimum independent of p; the latter makes all reports tie at $p = 1/2$. Both contradict strict propriety.

Theorem 3.3 (Universal calibration is impossible)**.** For every finite-valued strictly proper binary score and every $\tau \geq 0$, there is an interior two-report policy with positive, unequal outcome-wise variances. Its outcome-wise standardized population payoff has a unique continuous maximizer different from the true probability. No score in this class is calibrated for every nondegenerate finitely supported policy.

The same unequal-gap pair also reverses the best available token (Appendix B.2.1).

Corollary 3.4 (Operational two-token obstruction)**.** For every score covered by Theorem 3.3 and every finite $\tau \geq 0$, there exist a two-token vocabulary $\mathcal{Q}$, a nondegenerate policy, and an interior truth p for which $\Gamma_{\mathcal{Q}}(p)$ and $\Gamma_{\mathcal{Q}}(T_\nu(p))$ are opposite singletons.

Choose a truth between the distinct raw and standardized decision thresholds. The unique best token reverses. This failure is generic and has a measurable proper-score cost.

Proposition 3.5 (Generic and quantitative distortion)**.** For every score covered by Theorem 3.3, the equal-weight, zero-stabilizer report pairs that preserve calibration form a Borel set of two-dimensional Lebesgue measure zero. If the score is continuously differentiable, set $w(q) = \partial_q S(q,1)/(1-q)$. For $0 < \gamma < 1/2$, an equal-weight pair $1/2 + \gamma \leq x < z$ satisfies $T_\nu(1/2) - 1/2 \geq \gamma W_{x,z}/(W_{x,z} + 4\tau)$, where $W_{x,z} = \int_x^z w(s)\,ds$, with the reflected bound below $1/2$. If $w \geq m > 0$ between the truth and the displaced report, its regret under the original proper score is at least $m\{T_\nu(p) - p\}^2/2$.

Equal variances require two reports with the same expected score at truth $1/2$. Its strict monotonicity on either side leaves at most one partner for each report, giving the null-set result (Appendix B.3.1). Integrating the score gaps through w gives the displacement and regret bounds (Appendix B.3.2).

For multiple outcomes, positive scales $a_y = \sigma_y + \tau$ similarly elicit weights proportional to p_y/a_y. At $\tau > 0$, no finite-valued strictly proper multiclass score equalizes these scales for every two-point policy (Appendix B.5).

3.3 MAJORITY COLLAPSE ON THE TWO-REPORT FACE

For reports $x < z$, strict propriety gives positive branch gaps

$$d_1 = S(z,1) - S(x,1), \qquad d_0 = S(x,0) - S(z,0). \tag{8}$$

Without standardization, z is preferred when $p > \theta$, where

$$\theta = \frac{d_0}{d_0 + d_1}, \qquad x < \theta < z. \tag{9}$$

Let the current policy place mass r on z and write $c(r) = \sqrt{r(1-r)}$. The two branch standard deviations are $d_1 c(r)$ and $d_0 c(r)$.

Theorem 3.6 (Two-report payoff gap and zero-stabilizer majority collapse)**.** For any finite-valued strictly proper binary score and any interior p, r, x, z, the standardized population payoff gap between z and x is $\Delta_\tau(r)$ in Equation (44). At zero stabilizer,

$$\Delta_0(r) = \frac{2p-1}{\sqrt{r(1-r)}}. \tag{10}$$

Thus outcome-wise standardization prefers z exactly when $p > 1/2$, independently of the score and report locations.

Division cancels the two score gaps, leaving equal-strength votes in this two-report, zero-stabilizer case. In Figure 1, the Brier reports 0.6 and 0.8 have raw threshold 0.7: at truth 0.6, the proper score selects 0.6 but Equation (10) selects 0.8. Positive stabilizers retain some gap information.

With $\tau > 0$, the decision threshold moves from θ at $c = 0$ toward $1/2$ as c reaches $1/2$, allowing two symmetric roots (Corollary B.5). Stabilization restores score geometry but leaves the threshold policy dependent; Appendix B.4 gives the exact formula.

4 EXACT FINITE GROUPS AND THE NEURAL FIELD

Each sampled report affects the mean and scale used to update it. We now retain this dependence for an arbitrary finite vocabulary, then solve the two-report case.

4.1 THE EXACT FIELD FOR ANY FINITE VOCABULARY

Let the available reports be $q_1 < \cdots < q_M$, with categorical policy $\pi \in \Delta_M^\circ$ and logits z. Write $s_{jy} = S(q_j, y)$. A group of size n has counts $N \sim \mathrm{Mult}(n, \pi)$ and empirical masses $u_j = N_j/n$. For outcome y, define

$$\bar{s}_y(u) = \sum_j u_j s_{jy}, \qquad \sigma_y(u) = \left\{ \sum_j u_j (s_{jy} - \bar{s}_y(u))^2 \right\}^{1/2}. \tag{11}$$

The score-function update (Williams, 1992) detaches these statistics. Zero-scale groups at $\tau = 0$ receive zero advantage. We use ascent signs: a positive field increases its logit, so the field is the negative gradient of the stopped-gradient loss.

Theorem 4.1 (Exact finite-vocabulary same-group field)**.** For every finite score matrix, $n \geq 2$, and $\tau \geq 0$, the expected update of logit z_ℓ is

$$g^{(M)}_{n,\tau,\ell} = \mathbb{E}_N \left[u_\ell \left\{ p \frac{s_{\ell 1} - \bar{s}_1(u)}{\sigma_1(u) + \tau} + (1-p) \frac{s_{\ell 0} - \bar{s}_0(u)}{\sigma_0(u) + \tau} \right\} \right], \tag{12}$$

with the zero-scale convention above. The field sums to zero and has the equivalent symmetric weighted-pair form in Appendix C.1, Equations (57)–(58).

Centering cancels the softmax baseline within each group. Expanding the remaining scores gives Equation (58): each score gap is divided by an outcome-specific random scale containing both candidates. Appendix C.1 derives this exact expectation.

For $n = 2$ and $\tau = 0$, strict propriety orders both branch scores and the field reduces to

$$g_\ell = (2p-1)\pi_\ell \left(\sum_{j<\ell} \pi_j - \sum_{j>\ell} \pi_j \right). \tag{13}$$

Two-sample groups erase score magnitudes, consistent with Vojnovic & Yun (2025). With $M \geq 3$ and $n \geq 3$, multiple gaps within a composition can make the direction score dependent; the universal majority rule instead holds on every two-report face. For three Brier reports, the exact field can even locally decrease the raw expected score (Appendix C.4).

Does the fresh field have a scalar objective?

Proposition 4.2 (No Euclidean-logit potential for the fresh rank field)**.** Let $n = 2$, $\tau = 0$, $M \geq 3$, and $p \neq 1/2$. At every interior policy, the rank field in Equation (13) is not locally the Euclidean logit gradient of a scalar potential. Adjacent-logit cross derivatives differ as shown in Equation (68). The conclusion holds on the softmax quotient and is not a redundant-logit artifact.

This separates the fresh update field from the scalar-objective interpretation of the frozen population payoff. This obstruction also holds for a three-report Brier family at group sizes two and three for every finite $\tau \geq 0$ (Appendix C.2).

4.2 The solved two-report finite-group face

Take reports $x < z$, upper-report mass r, and branch gaps d_1, d_0 from Equation (8). If K of the n reports equal z, write $u = K/n$ and $c_K = \sqrt{u(1-u)}$. Parameterize the upper-report mass by the scalar logit $\vartheta = \log\{r/(1-r)\}$. We write $g_{n,\tau}(r)$ for its expected ascent update, so a positive value increases r.

Theorem 4.3 (Exact two-report finite-group update)**.** For every $n \geq 2$, the expected stopped-gradient logit update is

$$g_{n,\tau}(r) = \mathbb{E}_{K\sim \mathrm{Bin}(n,r)} \left[c_K^2 \left\{ \frac{pd_1}{d_1 c_K + \tau} - \frac{(1-p)d_0}{d_0 c_K + \tau} \right\} \right]. \tag{14}$$

At zero stabilizer,

$$g_{n,0}(r) = (2p-1)h_n(r), \qquad h_n(r) = \mathbb{E}[c_K] > 0, \tag{15}$$

so every finite group size has the score-independent majority threshold.

At zero stabilizer, score gaps cancel, leaving $(2p-1)c_K$ per mixed group (Appendix C.3.1). For positive stabilization, define

$$B_{y,n}(r) = \mathbb{E}\left[\frac{d_y c_K^2}{d_y c_K + \tau} \right], \qquad q_n(r) = \frac{B_{0,n}(r)}{B_{0,n}(r) + B_{1,n}(r)}. \tag{16}$$

Then $g_{n,\tau} = (B_{0,n} + B_{1,n})\{p - q_n(r)\}$. For $n = 2, 3$, every nonhomogeneous composition has one scale, so q_n is constant. For $n \geq 4$ and unequal gaps, it is symmetric around $1/2$ and strictly monotone on $(0, 1/2)$, creating one unstable and one stable interior root for truths between its center and boundary values (Theorem C.4); Appendix C.3.2 proves both parity cases.

The proof pairs each composition with the one obtained by swapping the two report counts. As the policy approaches an equal mixture, more balanced compositions gain relative weight. With at least four samples there are multiple nonzero scales, so this ordering makes the threshold move monotonically and yields the stated root structure.

Numerical illustrations and checks. Across six group sizes, seven truths, 21 policy masses, and four scores, numerical evaluation reproduces the predicted fields and root structure. Direct sampling also agrees with the exact candidate-dependent update (Figure 2).

4.3 Preserving the restricted Bayes target

Preserve the relative outcome weights for each prompt: either omit division or use one positive candidate-independent scale shared across outcomes.

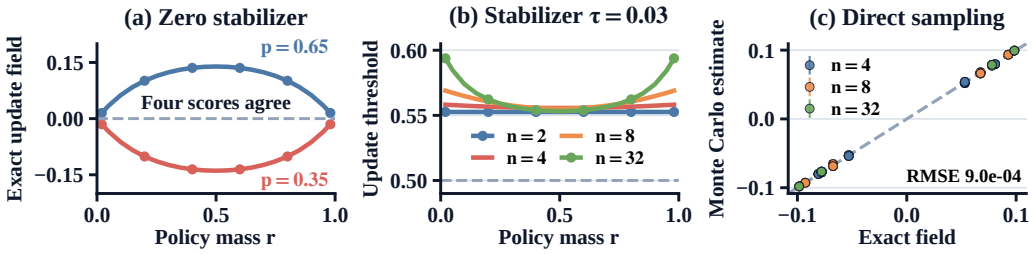


Figure 2: **Exact finite-group predictions survive direct checks.** (a) At zero stabilizer, four distinct strictly proper scores give the same two-report field, as Theorem 4.3 predicts. (b) A positive stabilizer restores score geometry and creates the group-size-dependent threshold of Theorem C.4. (c) Directly sampled groups agree with the exact expectation (RMSE 8.98×10^{-4}).

Without division, same-group centering preserves the target. Put $U_\ell = pS(q_\ell, 1) + (1-p)S(q_\ell, 0)$ and $\bar{U}_\pi = \sum_j \pi_j U_j$.

Proposition 4.4 (Exact no-standardization field)**.** Mean centering without standard-deviation division has expected field

$$g_\ell^{\text{no-std}} = \frac{n-1}{n}\pi_\ell(U_\ell - \bar{U}_\pi). \tag{17}$$

It is the ordinary categorical proper-score field times a positive constant and preserves $\Gamma_{\mathcal{Q}}(p)$. Moreover it is conservative, with potential $\Phi_{\text{no-std}}(z) = \frac{n-1}{n}\sum_j \pi_j(z)U_j$.

Multinomial second moments give $(n-1)/n$ (Appendix D.1). Each group's centered advantages equal $(n-1)/n$ times its RLOO advantages (Ahmadian et al., 2024). This gradient proportionality also appears in Bereket & Leskovec (2025).

For the second construction, let reference randomness R be candidate independent and evaluate the counterfactual branches with weights p and $1-p$. The implementation below also makes R independent of outcome Y. Consider

$$A(q, y; R) = \alpha_y(R)S(q, y) + \beta_y(R), \tag{18}$$

where $\alpha_y(R) > 0$ and $\beta_y(R)$ is report-independent.

Theorem 4.5 (Candidate-independent target preservation)**.** Under the positivity and candidate-independence conditions above, conditional on R and with the fixed counterfactual branch weights, Equation (18) preserves every continuous interior truth if and only if $\alpha_0(R) = \alpha_1(R) > 0$. On a finite vocabulary its maximizer set is $\Gamma_{\mathcal{Q}}(t_R)$, where

$$t_R = \frac{p\alpha_1(R)}{p\alpha_1(R) + (1-p)\alpha_0(R)}. \tag{19}$$

Equal coefficients therefore preserve $\Gamma_{\mathcal{Q}}(p)$ for every realized reference group.

The effective-truth reduction in Theorem 3.1 gives the proof: preserving every interior p forces equal positive branch coefficients (Appendix D.2). An independent fresh on-policy reference preserves each prompt's proper-score direction and zeros. Its positive multiplier can differ across prompts with shared parameters; Appendix D.2.1 treats this aggregation and its geometry.

Concretely, evaluate both outcome scores on a reference group independent of update reports and outcome. Use its branch means for centering and the RMS of its branch standard deviations plus a positive stabilizer as the shared divisor, without requiring p (Appendix D.2). Two-fold cross-fitting separates reference and update reports. Separate outcome scales reweight branches; candidate-dependent common scales can also change the target (Appendix D.3).

4.4 An Exact Bridge to the Neural Gradient

For the ten-token gradient check in Section 5, let $z_h(\phi) \in \mathbb{R}^{10}$ be the selected logits for prompt h, $\pi_h = \text{softmax}(z_h)$, and $J_h = \partial z_h/\partial\phi$. If L_h is the implemented mean stopped-gradient log-

probability loss, the shared-outcome protocol gives

$$\mathbb{E}[-\nabla_\phi L_h \mid h] = J_h^\top g^{(10)}_{16,0.03}(\pi_h, p_h). \tag{20}$$

The softmax identity gives the exact logit field and the chain rule supplies $J_h^\top$. This predicts the fresh raw gradient before clipping or AdamW for any finite report vocabulary, even when the effective truth is unavailable.

At fixed group size, the no-division factor $(n-1)/n$ is common across prompts. Under iid prompt draws, plain SGD with a positive global scale fixed before each minibatch therefore preserves the shared-parameter mean gradient (Corollary D.1). Dividing each gradient by its own norm can reweight random contributions (Appendix D.4), motivating the direct SGD and complementary unit-gradient tests below.

5 Experiments

We test the theory-to-neural chain, following the columns of Figure 3: the exact field predicts the model's gradient and the target shifts before training (Section 5.1); outcome-wise scaling then learns worse reports than centering alone (no division; Section 5.2); and the gap remains when gradient size and training budget are matched (Section 5.3). Qwen2.5-3B-Instruct reports the probability that a Civil Comments annotator judges a comment toxic (Qwen et al., 2024; Borkan et al., 2019). Annotation fractions define truths p_h; each update scores a prompt's reports against one shared draw $Y \sim \text{Bernoulli}(p_h)$. Restricted regret, the best available report's expected reward minus the policy's, measures scoring loss; lower is better. Appendix F gives full protocols and supporting experiments.

5.1 Mechanism checks before training

The exact field predicts the mean neural gradient. To check that the expected logit update, on which later predictions rest, is the model's mean gradient, we enumerate all size-16 groups over ten report tokens at ten fixed validation prompts and compare the exact field with an independently sampled one. Their cosine agreement is 0.999999 over the 100 selected logits and 0.999779 after both are mapped through the same fixed logit Jacobian into LoRA space (Figure 3a).

The target and the update direction shift before training. We next ask whether the distortion is active at the model's own policies. The learning experiments use reports $\{0.05, 0.25, 0.75, 0.95\}$ and truths in $[0.15, 0.21]$ or $[0.79, 0.85]$, so the best available report differs from the extreme nearest the more likely outcome. With stabilizer 0.03, frozen population statistics give a mean absolute shift in effective truth of 0.01657 and change the best report for 16.2% of 210 validation prompts; target shifts also occur at frozen Qwen policies under Brier, logarithmic, and asymmetric cubic scores. In finite groups of size 16, where each report enters its own normalization statistics, the mean estimated angle to the original proper-score logit field is $8.50°$ with outcome-wise scaling and $0.24°$ without division; outcome-wise scaling has the larger angle for all 210 prompts (Figure 3b).

5.2 Learning with plain SGD

Plain SGD gives an exact reference. To test whether the shifted target changes learning, we train with plain SGD, which applies each averaged gradient without clipping or rescaling; under centering alone, its expected step follows the proper-score gradient (Corollary D.1). Both arms share the supervised adapter, four-report task, Brier reward, and group size 16; outcome-wise scaling uses stabilizer 0.03. Two tuning seeds fix each arm's learning rate; six new paired seeds train for 320 iid minibatch attempts, including zero gradients. We report validation-selected checkpoints and step 320.

Outcome-wise scaling ends with higher regret. Selected test regret is 0.06784 without division and 0.09351 with outcome-wise scaling, a gap of 0.02568 (0.02391 at step 320). All six paired gaps are positive at both endpoints (Figure 3c). Both arms improve on the initial regret of 0.10340, but centering alone improves more.

The gap comes from less accurate mean reports. We split Brier regret into mean-report squared error and report variance within each prompt, minus the best available report's squared error, which cancels between arms. At the selected checkpoints, outcome-wise scaling has 0.02960 higher mean-report error and 0.00392 lower report variance than no division (Figure 3d). At step 320, all six

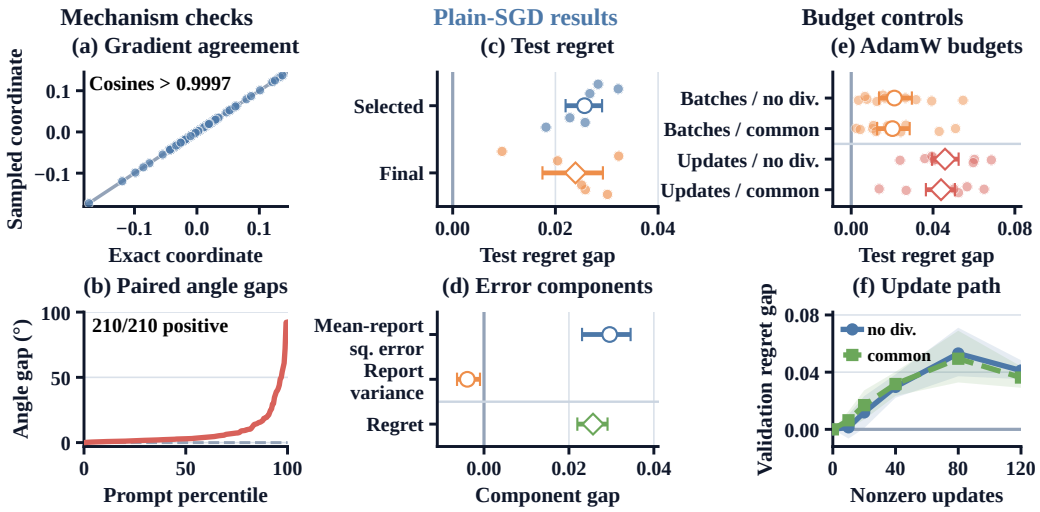


Figure 3: **Neural gradients and learning under group normalization.** (a) Exact versus sampled gradient coordinates. (b) Sorted paired differences in angle from the proper field across prompts. (c) Plain-SGD test regret at selected and final checkpoints. (d) Mean-report error and report variance decompose the selected regret gap. (e) AdamW controls match 320 minibatches (selected checkpoint) or 120 nonzero updates (final checkpoint). (f) Validation gaps along the matched-update path. Gaps subtract no division in (b,c,d) and each labeled control in (e,f); common denotes independent common scaling. In (c,d,e,f), large markers and curves show means, with bootstrap 95% intervals as bars or shading; small dots in (c,e) represent paired seeds.

outcome-wise policies are nearly constant at 0.25 across prompts (Table 7), failing to track the different truths. The regret gap therefore reflects biased probability reports rather than noisier ones.

5.3 Gradient-norm and budget controls

Unit-norm AdamW controls fix gradient size and training budget. The estimators have different raw gradient norms, and outcome-wise scaling produces more zero-gradient minibatches, so the gap could also reflect step size or training budget. We therefore rescale each nonzero gradient entering AdamW to unit norm and compare two training budgets. The controls are centering alone and independent common scaling, which uses one divisor shared across outcomes and independent of the scored reports and outcome.

The gap persists under both budgets. With 320 sampled minibatches, the selected test-regret gap against no division is 0.02110, positive in all 12 paired seeds. With exactly 120 nonzero updates, the corresponding gaps are 0.04586 at the fixed endpoint (all 12 pairs positive) and 0.01629 with validation selection (11 of 12 positive). Along the matched-update path, mean validation gaps against both controls stay positive at every measured checkpoint after initialization (Figure 3f). Independent common scaling, which also divides rewards, likewise has lower regret than outcome-wise scaling under both budgets (Figure 3e), consistent with Theorem 4.5.

6 Conclusion

When a shared outcome sets the group scale, normalization can change what a proper score elicits. We characterize this change, prove score-general impossibility, and derive the exact finite-group field. We identify cases where this field is not the gradient of a scalar objective in logit coordinates and give conditions that preserve the intended direction. Controlled Qwen experiments connect the theory to gradients and learned reports: the regret gap persists under plain SGD and complementary training-budget controls. Outcome-dependent normalization can therefore change the learning target, not just the update size.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and improving readability. The authors reviewed the AI-assisted revisions and take responsibility for the final content of this work.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2732–2778, 2024. URL https://proceedings.mlr.press/v235/band24a.html.

Yong Yi Bay and Kathleen A. Yearick. GRPO, Dr. GRPO, and DAPO are three operations on one number: The group-standard-deviation identity. *arXiv preprint arXiv:2607.00152*, 2026. URL https://arxiv.org/abs/2607.00152.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: GRPO induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*, 2025. URL https://arxiv.org/abs/2508.11800.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 491–500, 2019. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ASQ649zdHm.

Damek Davis and Benjamin Recht. What is the objective of reasoning with reinforcement learning? *arXiv preprint arXiv:2510.13651*, 2025. URL https://arxiv.org/abs/2510.13651.

Cheng Ge, Caitlyn Heqi Yin, Hao Liang, and Jiawei Zhang. Why GRPO needs normalization: A local-curvature perspective on adaptive gradients. *arXiv preprint arXiv:2601.23135*, 2026. URL https://arxiv.org/abs/2601.23135.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138. URL https://doi.org/10.1198/jasa.2011.r10138.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114, 1952. doi: 10.1111/j.2517-6161.1952.tb00104.x. URL https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1952.tb00104.x.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://jmlr.org/papers/v5/greensmith04a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Jefferson Hernandez, Jaywon Koo, Zilin Xiao, Chen Wei, and Vicente Ordonez. SoftmaxGRPO: Learning to reason using softmax advantage group estimation. *arXiv preprint arXiv:2608.09271*, 2026. URL https://arxiv.org/abs/2608.09271.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Youngeun Kim. MC-GRPO: Median-centered group relative policy optimization for small-rollout reinforcement learning. *arXiv preprint arXiv:2601.22582*, 2026. URL https://arxiv.org/abs/2601.22582.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. *arXiv preprint arXiv:2601.05242*, 2026. URL https://arxiv.org/abs/2601.05242.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Amritansh Mishra, Supriyo Chakraborty, and Berkcan Kapusuzoglu. On the policy gradient foundations of group relative policy optimization: Credit assignment, gradient sparsity, and rank collapse. *arXiv preprint arXiv:2606.29238*, 2026. URL https://arxiv.org/abs/2606.29238.

Youssef Mroueh. Reinforcement learning with verifiable rewards: GRPO's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025. URL https://arxiv.org/abs/2503.06639.

Lei Pang, Jun Luo, and Ruinan Jin. TIC-GRPO: Provable and efficient optimization for reinforcement learning from human feedback. *arXiv preprint arXiv:2508.02833*, 2025. URL https://arxiv.org/abs/2508.02833.

Qwen et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(83):2387–2422, 2010. URL https://jmlr.org/papers/v11/reid10a.html.

Mohammad Mahdi Salmani-Zarchi, Zahra Rahimi, Heshaam Faili, and Mohammad Javad Dousti. MDP-GRPO: Stabilized group relative policy optimization for multi-constraint instruction following. *arXiv preprint arXiv:2606.06058*, 2026. URL https://arxiv.org/abs/2606.06058.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. doi: 10.1080/01621459.1971.10482346. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482346.

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.

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. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Milan Vojnovic and Se-Young Yun. What is the alignment objective of GRPO? *arXiv preprint arXiv:2502.18548*, 2025. URL https://arxiv.org/abs/2502.18548.

Ziqi Wang, Xingzhou Lou, Meiqi Wu, Zhengqi Wen, and Junge Zhang. Calibration-aware policy optimization for reasoning LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18375–18390. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.836. URL https://aclanthology.org/2026.acl-long.836/.

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

Robert C. Williamson, Elodie Vernet, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(222):1–52, 2016. URL https://jmlr.org/papers/v17/14-294.html.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. BNPO: Beta normalization policy optimization. *arXiv preprint arXiv:2506.02864*, 2025. URL https://arxiv.org/abs/2506.02864.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Yunze Xiao, Junjue Wang, and Naoto Yokoya. The confidence dichotomy: Analyzing and mitigating miscalibration in tool-use agents. In *Proceedings of ACL*, pp. 11325–11349, 2026. doi: 10.18653/v1/2026.acl-long.520. URL https://aclanthology.org/2026.acl-long.520/.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*, 2026. URL https://arxiv.org/abs/2601.08521.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information*

Processing Systems, volume 38, pp. 113222–113244, 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Ying Yang, Shijin Gong, and Chengchun Shi. Demystifying group relative policy optimization: Its policy gradient is a U-statistic. *arXiv preprint arXiv:2603.01162*, 2026. URL https://arxiv.org/abs/2603.01162.

Guide to the appendix. Proofs cover population distortion (Appendix B), finite-group updates (C), and target preservation (D). Appendix A gives shared tools, E treats protocol variants, and F contains experimental details.

A SHARED MATHEMATICAL TOOLS

Write $f(q) = S(q, 0)$ and $g(q) = S(q, 1)$. We first record the only score-theoretic fact needed by the nonsmooth results.

Lemma A.1 (Opposite coordinate monotonicity)**.** If S is finite and strictly proper on $(0, 1)$, then for every $x < z$,

$$f(z) < f(x), \qquad g(z) > g(x). \tag{21}$$

Proof. Put $\Delta_f = f(z) - f(x)$ and $\Delta_g = g(z) - g(x)$. Strict propriety at truths x and z gives

$$\begin{aligned} x\Delta_g + (1-x)\Delta_f < 0, \\ z\Delta_g + (1-z)\Delta_f > 0. \end{aligned} \tag{22}$$

Subtracting shows $\Delta_g - \Delta_f > 0$. The affine function $t\Delta_g + (1-t)\Delta_f$ therefore has a unique zero $\theta = -\Delta_f/(\Delta_g - \Delta_f)$, and (22) places it in $(x, z) \subset (0, 1)$. Consequently $-\Delta_f > 0$ and $\Delta_g = (1-\theta)(\Delta_g - \Delta_f) > 0$. □

For any nondegenerate finitely supported report policy, both coordinate variances are therefore positive. The same lemma also gives $x < \theta < z$ in (9) without continuity or differentiability.

B POPULATION TARGET DISTORTION

B.1 EFFECTIVE TRUTH AND FINITE-VOCABULARY TARGETS

Let $a_y = \sigma_y + \tau$ and define

$$C_{p,\nu} = \frac{p}{a_1} + \frac{1-p}{a_0}, \qquad K_{p,\nu} = \frac{p\mu_1}{a_1} + \frac{(1-p)\mu_0}{a_0}. \tag{23}$$

Expanding (4) and collecting the terms that depend on q gives

$$\begin{aligned} G_{p,\nu}(q) &= \frac{p}{a_1} S(q, 1) + \frac{1-p}{a_0} S(q, 0) - K_{p,\nu} \\ &= C_{p,\nu} U_{T_\nu(p)}(q) - K_{p,\nu}, \end{aligned} \tag{24}$$

because

$$\frac{p/a_1}{p/a_1 + (1-p)/a_0} = \frac{pa_0}{pa_0 + (1-p)a_1} = T_\nu(p). \tag{25}$$

The coefficient $C_{p,\nu}$ is positive and $T_\nu(p) \in (0, 1)$, so strict propriety proves Theorem 3.1. Cross multiplication shows $T_\nu(p) = p$ at an interior p if and only if $a_0 = a_1$, equivalently $\sigma_0 = \sigma_1$.

Restricting (24) to $\mathcal{Q}$ preserves the argmax set because $C_{p,\nu} > 0$, proving (7), including ties. If $\nu = \sum_j \pi_j \delta_{q_j}$, branch centering gives $\sum_j \pi_j G_{p,\pi}(q_j) = 0$. The stopped-gradient softmax field is then

$$\begin{aligned} g_\ell^{\mathrm{pop}} &= \mathbb{E}[G_{p,\pi}(Q)(\mathbf{1}\{Q = q_\ell\} - \pi_\ell)] \\ &= \pi_\ell G_{p,\pi}(q_\ell). \end{aligned} \tag{26}$$

Averaging (24) under π gives $K_{p,\pi} = C_{p,\pi} \sum_j \pi_j U_{T_\pi(p)}(q_j)$, which proves

$$g_\ell^{\mathrm{pop}} = C_{p,\pi} \pi_\ell \left\{ U_{T_\pi(p)}(q_\ell) - \sum_j \pi_j U_{T_\pi(p)}(q_j) \right\}. \tag{27}$$

The effective truth has the useful odds and displacement forms

$$\frac{T_\nu(p)}{1-T_\nu(p)} = \frac{p}{1-p}\frac{a_0}{a_1}, \tag{28}$$

$$T_\nu(p) - p = \frac{p(1-p)(\sigma_0-\sigma_1)}{p(\sigma_0+\tau)+(1-p)(\sigma_1+\tau)}. \tag{29}$$

Thus a common finite additive stabilizer changes the magnitude but not the sign of an existing scale asymmetry.

B.2 Variance isometry and universal impossibility

Proposition B.1 (Variance-isometry characterization)**.** For arbitrary functions $f, g : (0,1) \to \mathbb{R}$, the following are equivalent:

(i) every nondegenerate two-point law has equal variances for $f(Q)$ and $g(Q)$;

(ii) $|f(x) - f(z)| = |g(x) - g(z)|$ for every x, z;

(iii) there are $\varepsilon \in \{-1, 1\}$ and $b \in \mathbb{R}$ such that $g(q) = \varepsilon f(q) + b$ for every q.

Either form in (iii) gives equal variances under every law for which they exist.

Proof. For $Q \sim w\delta_x + (1-w)\delta_z$,

$$\mathrm{Var}(f(Q)) = w(1-w)\{f(x) - f(z)\}^2, \tag{30}$$

and likewise for g, proving (i)⇔(ii). If f is constant, (ii) makes g constant and (iii) follows. Otherwise choose anchors a, b with $f(a) \neq f(b)$. Their signed difference fixes

$$\varepsilon = \frac{g(b) - g(a)}{f(b) - f(a)} \in \{-1, 1\}. \tag{31}$$

For any q, condition (ii) gives the distances from $g(q)$ to both real anchors $g(a), g(b)$. Two distinct real anchors determine a point uniquely, so $g(q) - g(a) = \varepsilon\{f(q) - f(a)\}$. Taking $b = g(a) - \varepsilon f(a)$ proves (iii). The converse follows because translations and reflections preserve variance. □

B.2.1 Universal impossibility and token reversal

Proof of Theorem 3.3. Assume toward contradiction that the two outcome-coordinate variances agree for every nondegenerate two-point policy. Proposition B.1 implies either

$$S(q,1) = S(q,0) + b \quad \text{or} \quad S(q,1) = -S(q,0) + b. \tag{32}$$

In the first case, $U_p(q) = S(q,0) + pb$, whose maximizer set is independent of p; it cannot have distinct unique maximizers at two different truths. In the second case, $U_{1/2}(q) = b/2$ for every report, contradicting uniqueness at truth $1/2$. Strict propriety excludes both forms.

There must therefore be reports $x < z$ whose two coordinate distances differ. Lemma A.1 makes both distances positive. Under the equal-weight policy, the branch standard deviations are $d_0/2$ and $d_1/2$, hence positive and unequal. At truth $p = 1/2$, Theorem 3.1 gives the unique optimizer

$$T_\nu(1/2) = \frac{d_0/2 + \tau}{(d_0+d_1)/2 + 2\tau} \neq \frac{1}{2}. \tag{33}$$

This supplies an interior stochastic truth, two interior reports, a nondegenerate policy, and positive denominators for every finite τ. □

Thus every strictly proper score admits a policy that changes the elicited truth. Policies with equal branch variances remain calibrated.

Proof of Corollary 3.4. The proof above supplies $x < z$ with unequal positive branch gaps $d_0 \neq d_1$. Under their equal-weight policy, the raw two-token decision changes at

$$\theta = \frac{d_0}{d_0 + d_1}, \tag{34}$$

while the outcome-wise standardized decision changes at

$$p_c = \frac{\frac{1}{2}d_0 d_1 + \tau d_0}{d_0 d_1 + \tau(d_0 + d_1)}. \tag{35}$$

Equality $p_c = \theta$ would imply $d_0 = d_1$. Any interior truth strictly between the two thresholds therefore gives opposite singleton decisions for $\Gamma_{\{x,z\}}(p)$ and $\Gamma_{\{x,z\}}(T_\nu(p))$. □

B.3 Prevalence and magnitude of distortion

B.3.1 Almost-everywhere distortion

Corollary B.2 (Almost-everywhere distortion)**.** For every finite-valued strictly proper binary score, equal-weight, zero-stabilizer calibrated report pairs form a Borel set of two-dimensional Lebesgue measure zero in the ordered two-report triangle.

Let

$$H(q) = U_{1/2}(q) = \frac{f(q) + g(q)}{2}. \tag{36}$$

The coordinate functions are finite monotone functions, hence Borel measurable, and so is H. If $x < z \leq 1/2$, the restricted threshold between x and z lies in (x, z) and is below $1/2$. The upper report is therefore preferred at truth $1/2$, which gives $H(z) > H(x)$. Thus H is strictly increasing on $(0, 1/2]$. The symmetric argument shows that H is strictly decreasing on $[1/2, 1)$.

For the equal-weight two-report policy, the two branch standard deviations are equal exactly when

$$f(x) - f(z) = g(z) - g(x), \tag{37}$$

where Lemma A.1 fixes the signs. This is equivalent to $H(x) = H(z)$. Define

$$\mathcal{C} = \{(x, z) : 0 < x < z < 1,\ H(x) = H(z)\}. \tag{38}$$

The set is Borel. Its pairs must straddle $1/2$, and every vertical section contains at most one point because H is strictly decreasing on the right half. Every section thus has one-dimensional Lebesgue measure zero. Tonelli's theorem yields $\lambda_2(\mathcal{C}) = 0$, proving Corollary B.2.

B.3.2 Displacement and regret bounds

The nonsmooth results establish a universal obstruction. Differentiability adds explicit magnitude bounds through the standard weight-function representation of binary proper losses (Reid & Williamson, 2010).

Lemma B.3 (Binary weight identity)**.** If the coordinate functions of a strictly proper score are continuously differentiable, there is a continuous $w : (0, 1) \to [0, \infty)$ such that

$$\partial_q S(q, 1) = (1 - q)w(q), \qquad \partial_q S(q, 0) = -qw(q), \tag{39}$$

and $\int_x^z w(s)\, ds > 0$ for every $x < z$.

Proof. The first-order condition at the interior optimum $q = t$ gives $t\partial_q S(t, 1) + (1-t)\partial_q S(t, 0) = 0$. Coordinate monotonicity and continuous differentiability imply the two derivatives have the displayed signs, so they admit the common nonnegative factor w. Strict coordinate monotonicity makes every interval integral positive. □

Proposition B.4 (Exact pair displacement and lower bound)**.** For $0 < x < z < 1$, let $\nu = (\delta_x + \delta_z)/2$ and $W_{x,z} = \int_x^z w(s)\, ds$. At truth $1/2$,

$$T_\nu(1/2) - \frac{1}{2} = \frac{\int_x^z (2s - 1)w(s)\, ds}{2\{W_{x,z} + 4\tau\}}. \tag{40}$$

For $0 < \gamma < 1/2$, if $x \geq 1/2 + \gamma$, then

$$T_\nu(1/2) - \frac{1}{2} \geq \frac{\gamma W_{x,z}}{W_{x,z} + 4\tau}; \tag{41}$$

the reflected bound holds when $z \leq 1/2 - \gamma$. If additionally $w \geq m > 0$ on $[x, z]$ and $z - x \geq \ell > 0$, the right side is at least $\gamma m\ell/(m\ell + 4\tau)$.

Proof. Integrating (39) gives

$$\begin{aligned} d_1 &= \int_x^z (1 - s)w(s)\, ds, \\ d_0 &= \int_x^z sw(s)\, ds. \end{aligned} \tag{42}$$

Under the equal-weight policy, $\sigma_y = d_y/2$. Substitution into (6) and simplification yields (40). On a pair above $1/2 + \gamma$, $2s - 1 \geq 2\gamma$, which proves (41); the other claims follow identically. □

If $w \geq m > 0$ between p and $q_\star = T_\nu(p)$, the unstandardized proper-score regret satisfies

$$U_p(p) - U_p(q_\star) = \left| \int_p^{q_\star} (p - s)w(s)\, ds \right| \geq \frac{m}{2}(q_\star - p)^2. \tag{43}$$

Thus displacement translates directly into loss under the original proper objective.

B.4 Two-report population dynamics and score examples

B.4.1 Positive-stabilizer phase calculation

For the two-report population payoff in Theorem 3.6, subtracting the centered branch scores gives

$$\Delta_\tau(r) = \frac{pd_1}{d_1 c(r) + \tau} - \frac{(1 - p)d_0}{d_0 c(r) + \tau}. \tag{44}$$

At $\tau = 0$, cancellation gives Equation (10).

At $\tau > 0$, solving $\Delta_\tau(r) = 0$ gives the critical truth

$$p_c(c, \tau) = \frac{cd_0 d_1 + \tau d_0}{2cd_0 d_1 + \tau(d_0 + d_1)}. \tag{45}$$

Corollary B.5 (Positive-stabilizer population phase)**.** Suppose $d_0 > d_1$ and let $\theta_m = p_c(1/2, \tau)$, so $1/2 < \theta_m < \theta$. Truths $p \in (\theta_m, \theta)$ produce two symmetric interior roots under the probability flow: the lower is unstable and the upper stable, alongside a stable left boundary. Outside this interval the field has one direction throughout the interior. For $d_0 < d_1$, the portrait is reflected.

From (44), the zero of the payoff gap as a function of truth is (45), and

$$\Delta_\tau(r) = \left\{ \frac{d_1}{cd_1 + \tau} + \frac{d_0}{cd_0 + \tau} \right\} \{p - p_c(c, \tau)\}. \tag{46}$$

Direct differentiation gives

$$\frac{\partial p_c}{\partial c} = \frac{\tau d_0 d_1 (d_1 - d_0)}{\{2cd_0 d_1 + \tau(d_0 + d_1)\}^2}. \tag{47}$$

For $d_0 > d_1$, the critical truth decreases as c rises, from θ at $c = 0$ to θ_m at $c = 1/2$. Since $c(r)$ is symmetric and single-peaked, a horizontal truth in (θ_m, θ) intersects the critical curve twice. The signs in (46) yield the stability labels in Corollary B.5. The case $d_0 < d_1$ follows by reflection.

For fixed interior r, $p_c(c(r), \tau) \to 1/2$ as $\tau \downarrow 0$, but $p_c(0, \tau) = \theta$ for every positive τ. Setting $c = \lambda\tau$ gives

$$p_c(\lambda\tau, \tau) = \frac{\lambda d_0 d_1 + d_0}{2\lambda d_0 d_1 + d_0 + d_1}, \tag{48}$$

so the boundary layer has $c = O(\tau)$ and $r = O(\tau^2)$.

B.4.2 BRIER AND LOGARITHMIC REPORTS

For the negative Brier reward $S(q, y) = -(q - y)^2$ and a pair $x < z$,

$$d_0 = (z - x)(x + z), \qquad d_1 = (z - x)(2 - x - z). \tag{49}$$

The proper restricted threshold is $(x+z)/2$. With $(x, z) = (0.6, 0.8)$, it is 0.7, whereas zero-stabilizer outcome-wise standardization uses $1/2$.

For the logarithmic reward, the gaps are

$$d_0 = \log \frac{1 - x}{1 - z}, \qquad d_1 = \log \frac{z}{x}. \tag{50}$$

At $(0.6, 0.8)$, the unstandardized restricted threshold is approximately 0.706695. Equal branch distances occur exactly for symmetric pairs $x + z = 1$. These examples illustrate score-specific proper thresholds; the zero-stabilizer normalized threshold is $1/2$ for both.

Under the population natural/replicator flow and $\tau = 0$,

$$\dot{r} = (2p - 1)\sqrt{r(1 - r)}, \tag{51}$$

so $\arcsin(2r - 1)$ evolves linearly until a boundary. This is compatible with the known arcsine transform for standardized binary-reward GRPO (Davis & Recht, 2025). Here the same coordinate description follows for every two-report strictly proper score under a shared external stochastic truth.

B.5 MULTICLASS TARGETS AND IMPOSSIBILITY

Let $\Delta_k^\circ = \{q \in (0, 1)^k : \sum_y q_y = 1\}$ and let $S : \Delta_k^\circ \times \{1, \dots, k\} \to \mathbb{R}$ be a finite-valued strictly proper reward score, in the standard multiclass proper-score setting (Gneiting & Raftery, 2007; Williamson et al., 2016). For $Q \sim \nu$, define coordinate means μ_y, standard deviations σ_y, and positive scales $a_y = \sigma_y + \tau$. The effective-truth result requires only positivity of every a_y; the universal two-point impossibility below assumes $\tau > 0$, which guarantees that condition.

Theorem B.6 (Multiclass effective truth)**.** For $p \in \Delta_k^\circ$, the field

$$G_{p,\nu}^{(k)}(q) = \sum_{y=1}^{k} p_y \frac{S(q, y) - \mu_y}{a_y} \tag{52}$$

has the unique maximizer

$$T_{\nu,y}^{(k)}(p) = \frac{p_y / a_y}{\sum_{j=1}^{k} p_j / a_j}. \tag{53}$$

It is calibrated at an interior truth if and only if $a_1 = \cdots = a_k$; in that case it is calibrated at every interior truth.

Proof. With $C = \sum_y p_y / a_y > 0$ and $K = \sum_y p_y \mu_y / a_y$,

$$G_{p,\nu}^{(k)}(q) = C \sum_y T_{\nu,y}^{(k)}(p) S(q, y) - K. \tag{54}$$

Strict propriety gives (53). Since each $p_y > 0$, equality $T_\nu^{(k)}(p) = p$ is equivalent to $C a_y = 1$ for every coordinate. □

Theorem B.7 (Multiclass universal impossibility)**.** For $k \geq 2$, $\tau > 0$, and any finite-valued strictly proper score on Δ_k°, there is a nondegenerate two-point policy whose coordinate standard deviations are not all equal. Its outcome-wise standardized field is distorted at every interior truth.

Proof. If all coordinate variances agreed under every two-point policy, then for every report pair q, r and outcome y,

$$|S(q, y) - S(r, y)| = |S(q, 1) - S(r, 1)|. \tag{55}$$

The two-anchor argument from Proposition B.1 gives a real function h, constants b_y, and signs $\varepsilon_y \in \{-1, 1\}$ such that $S(q, y) = \varepsilon_y h(q) + b_y$. Expected score is then

$$\left(\sum_y p_y \varepsilon_y\right) h(q) + \sum_y p_y b_y. \tag{56}$$

For $k = 2$, equal signs make the report-dependent objective independent of the truth, while opposite signs make it constant at the uniform truth. For $k \geq 3$, two outcomes share a sign; moving probability mass between them produces distinct interior truths with the same coefficient of h and hence the same maximizer set. Every case contradicts strict propriety. Positive τ makes all denominators admissible, and Theorem B.6 completes the proof. □

The candidate-independent correction extends to this setting. Use a coordinate-specific reference mean and one positive candidate-independent denominator shared by all outcomes. Its conditional expectation is a positive affine transform of the original proper objective.

C Exact finite-group updates and dynamics

C.1 Exact many-report field

To prove Theorem 4.1, define

$$W_{\ell j}^{(y)} = \mathbb{E}_N\left[\frac{u_\ell u_j}{\sigma_y(u) + \tau} \mathbf{1}\{\sigma_y(u) + \tau > 0\}\right], \tag{57}$$

so the field can be written as

$$g_\ell = p \sum_{j \neq \ell} W_{\ell j}^{(1)} (s_{\ell 1} - s_{j1}) + (1 - p) \sum_{j \neq \ell} W_{\ell j}^{(0)} (s_{\ell 0} - s_{j0}), \tag{58}$$

with $\sum_\ell g_\ell = 0$.

For a realized composition and outcome, define $a_{j,y} = (s_{jy} - \bar{s}_y)/(\sigma_y + \tau)$, with the specified zero fallback. The empirical centered advantages satisfy

$$\sum_j u_j a_{j,y} = 0. \tag{59}$$

The group-average softmax score-function estimator therefore simplifies pointwise.

$$\begin{aligned} \widehat{g}_\ell &= \sum_j u_j a_{j,Y} (\mathbf{1}\{j = \ell\} - \pi_\ell) \\ &= u_\ell a_{\ell,Y}. \end{aligned} \tag{60}$$

The skipped case has both sides zero. Averaging the shared outcome and then the exact multinomial composition proves (12). Also,

$$u_\ell (s_{\ell y} - \bar{s}_y) = \sum_j u_\ell u_j (s_{\ell y} - s_{jy}), \tag{61}$$

which gives (58). Since $W_{\ell j}^{(y)} = W_{j\ell}^{(y)}$, pairwise cancellation gives $\sum_\ell g_\ell = 0$. The expectation is a finite sum over $\binom{n+M-1}{M-1}$ compositions and retains the candidate inside its own mean and denominator.

C.1.1 Rank-only groups of size two

For $n = 2$ and $\tau = 0$, a nonhomogeneous pair $q_j < q_\ell$ has standardized advantages $(-1, +1)$ under outcome one and $(+1, -1)$ under outcome zero. The unordered pair occurs with probability $2\pi_j \pi_\ell$, and each member has empirical mass $1/2$. Summing the upper-minus-lower contribution over all pairs incident to ℓ proves (13). This uses only the opposite monotonicity of the two branches of a strictly proper binary score. Every two-action face at arbitrary n similarly reduces to Theorem 4.3.

C.2 Objective geometry of the exact field

Let $\rho_1 = p$, $\rho_0 = 1 - p$, and write the detached contribution of a fixed composition as

$$b_i(N) = u_i \sum_{y\in\{0,1\}} \rho_y \frac{s_{iy} - \bar{s}_y(u)}{\sigma_y(u) + \tau}, \qquad g_i(z) = \mathbb{E}_{N\sim\mathrm{Mult}(n,\pi(z))} b_i(N), \tag{62}$$

with the specified zero-scale convention.

Proposition C.1 (Exact multinomial curl)**.** For every finite score matrix and finite group size,

$$\frac{\partial g_i}{\partial z_j} = \mathbb{E}[b_i(N)(N_j - n\pi_j)]\,,$$
$$C_{ij} := \frac{\partial g_i}{\partial z_j} - \frac{\partial g_j}{\partial z_i} = \mathbb{E}[b_i(N)(N_j - n\pi_j) - b_j(N)(N_i - n\pi_i)]\,. \tag{63}$$

The field is a scalar logit gradient exactly when every pairwise curl vanishes identically.

For fixed N, $b_i(N)$ is independent of π, whereas

$$\frac{\partial}{\partial z_j} \log \Pr_{\pi(z)}(N) = N_j - n\pi_j. \tag{64}$$

Differentiating the finite multinomial sum proves Equation (63). The field sums to zero and is invariant under $z \mapsto z + a\mathbf{1}$, so it defines a one-form on the simply connected softmax quotient. The last claim is the Poincaré criterion in any quotient chart.

For $n = 2$, define the skew-symmetric edge matrix

$$A_{ij} = \sum_{y\in\{0,1\}} \rho_y \frac{s_{iy} - s_{jy}}{|s_{iy} - s_{jy}| + 2\tau}. \tag{65}$$

Theorem C.2 (Complete two-sample characterization)**.** The exact two-sample field is $g_i = \pi_i \sum_j A_{ij}\pi_j$. It has a scalar logit potential if and only if every triangle circulation vanishes,

$$A_{ij} + A_{jk} + A_{ki} = 0 \qquad \text{for all } i, j, k. \tag{66}$$

Equivalently, $A_{ij} = c_i - c_j$ for constants c_i, in which case $\Phi(z) = \sum_i \pi_i(z)c_i$ satisfies $\nabla_z \Phi = g$.

Writing $h = A\pi$, direct softmax differentiation gives, for $i \neq j$,

$$\partial_{z_j} g_i - \partial_{z_i} g_j = 2\pi_i\pi_j(A_{ij} - h_i + h_j). \tag{67}$$

Vanishing throughout the interior is equivalent to $A_{ik} - A_{jk} = A_{ij}$ for all k, which is precisely Equation (66). Anchoring one index gives the difference representation, and differentiating the displayed potential proves the converse.

Proof of Proposition 4.2. For adjacent reports, the claimed cross-derivative difference is

$$\frac{\partial g_\ell}{\partial z_{\ell+1}} - \frac{\partial g_{\ell+1}}{\partial z_\ell} = -2(2p-1)\pi_\ell\pi_{\ell+1}(1 - \pi_\ell - \pi_{\ell+1}) \neq 0. \tag{68}$$

Put $c = 2p - 1$ and define the rank matrix $R_{\ell j} = 1$ for $j < \ell$, $R_{\ell j} = -1$ for $j > \ell$, and zero on the diagonal. With $h = R\pi$, Equation (13) is $g_\ell = c\pi_\ell h_\ell$. For $\ell < k$, softmax differentiation gives

$$\frac{\partial g_\ell}{\partial z_k} - \frac{\partial g_k}{\partial z_\ell} = 2c\pi_\ell\pi_k \left(\sum_{\ell<j<k} \pi_j - \sum_{j<\ell} \pi_j - \sum_{j>k} \pi_j \right). \tag{69}$$

Taking $k = \ell+1$ yields Equation (68). In the quotient chart $z_M = 0$, the $(1, 2)$ curl is nonzero at every interior policy because $\sum_{j=3}^{M} \pi_j > 0$. Hence no local potential exists. When $M = 2$ the quotient is one-dimensional and the potential is $(2p - 1)\pi_2$; when $p = 1/2$ the rank field vanishes. □

The obstruction persists after positive stabilization.

Corollary C.3 (An all-stabilizer Brier witness)**.** For $S(q, y) = -(q - y)^2$, reports $(1/4, 1/2, 3/4)$, and the uniform policy,

$$C_{12}^{(n=2)} = -\frac{10(2p-1)(8\tau+1)}{9(4\tau+1)(32\tau+3)(32\tau+5)}, \tag{70}$$

$$C_{12}^{(n=3)} = -\frac{40(2p-1)(6\sqrt{2}\tau+1)}{27(6\tau+\sqrt{2})(16\tau+\sqrt{2})(48\tau+5\sqrt{2})}. \tag{71}$$

Thus the exact stopped-gradient field is nonconservative for every finite $\tau \geq 0$ whenever $p \neq 1/2$, at both group sizes.

For $n = 2$, substituting the three Brier edges into Theorem C.2 gives Equation (70). For $n = 3$, substituting the ten compositions into Proposition C.1 gives Equation (71). Both expressions decay as $O(\tau^{-2})$. Stabilization attenuates the curl but does not generally restore an objective. The conclusion concerns Euclidean integrability of the exact expected field.

C.3 Two-report updates and stabilizer phase

C.3.1 Exact two-report update and boundary rate

Let $Z_i = \mathbf{1}\{Q_i = z\}$. Conditional on K, the outcome-one reward vector is an affine transform of $(Z_i)_{i=1}^n$ with positive gap d_1. Its centered entries and divisor-n standard deviation are $d_1(Z_i - u)$ and $d_1 c_K$. Under outcome zero the upper report has the lower reward, giving $-d_0(Z_i - u)$ and $d_0 c_K$. This gives the conditional advantages used in the main-text derivation.

For either outcome, the group-average Bernoulli-logit score-function term contains

$$\begin{aligned}\frac{1}{n}\sum_i (Z_i - u)(Z_i - r) &= \frac{1}{n}\sum_i (Z_i - u)Z_i - r\frac{1}{n}\sum_i (Z_i - u) \\ &= u - u^2 = c_K^2.\end{aligned} \tag{72}$$

Therefore

$$\mathbb{E}[\widehat{g} \mid K] = c_K^2 \left\{ \frac{p d_1}{d_1 c_K + \tau} - \frac{(1-p)d_0}{d_0 c_K + \tau} \right\}. \tag{73}$$

Averaging $K \sim \mathrm{Bin}(n, r)$ proves (14). Homogeneous groups have a zero centered numerator. At $\tau = 0$ we define their $0/0$ contribution as zero, matching the stated skip convention.

Let F_z and F_x be the expected detached advantage conditional on the focal report. Exchangeability gives

$$\begin{aligned}g_{n,\tau} &= rF_z(1-r) + (1-r)F_x(-r) \\ &= r(1-r)(F_z - F_x),\end{aligned} \tag{74}$$

Thus $F_z - F_x = g_{n,\tau}/\{r(1-r)\}$. At $\tau = 0$, every nonhomogeneous conditional term in (73) equals $(2p-1)c_K$, giving (15). For interior r and $n \geq 2$, the binomial law assigns positive probability to a nonhomogeneous composition, so $h_n(r) > 0$.

Two small-group examples are

$$h_2(r) = r(1-r), \qquad h_3(r) = \sqrt{2}\, r(1-r). \tag{75}$$

Near the right boundary, $h_n(r) = \sqrt{n-1}(1-r) + O((1-r)^2)$. Consider the deterministic expected update $\vartheta_{t+1} = \vartheta_t + \eta g_{n,0}(r_t)$, with $r_t = (1 + e^{-\vartheta_t})^{-1}$, fixed $n \geq 2$, constant $\eta > 0$, and $r_0 \in (0, 1)$. For $p > 1/2$, ϑ_t increases to infinity and $e^{\vartheta_{t+1}} - e^{\vartheta_t} \to \eta(2p-1)\sqrt{n-1}$. Consequently,

$$1 - r_t \sim \frac{1}{\eta(2p-1)\sqrt{n-1}\, t}. \tag{76}$$

C.3.2 Positive-stabilizer phase

Stability below refers to the continuous expected logit flow $\dot{\vartheta} = g_{n,\tau}(r)$, or equivalently $\dot{r} = r(1-r)g_{n,\tau}(r)$.

Theorem C.4 (Two-report finite-group stabilizer phase)**.** Assume $\tau > 0$. Every nonhomogeneous group has one scale and q_n is constant for $n = 2, 3$. For $n \geq 4$ and unequal gaps, q_n is symmetric around $1/2$ and strictly monotone on $(0, 1/2)$; truths between its center and boundary values yield one unstable and one stable interior root.

Put

$$b_y(c) = \frac{d_y c^2}{d_y c + \tau}, \qquad B_{y,n}(r) = \mathbb{E}[b_y(c_K)]. \tag{77}$$

Equation (14) gives $g_{n,\tau} = pB_{1,n} - (1-p)B_{0,n}$ and therefore (16).

Let $J = \min(K, n-K)$ and restrict first to $r \in (0, 1/2)$. With $t = r/(1-r) \in (0,1)$, the unnormalized folded mass at $j < n/2$ is

$$w_j(t) = \binom{n}{j} t^j \{1 + t^{n-2j}\}, \tag{78}$$

up to a factor common to all j; the even center has one additional constant factor independent of t. For adjacent states, put $m = n - 2j - 2 \geq 0$. Apart from a positive constant,

$$\frac{w_{j+1}(t)}{w_j(t)} \propto R_m(t) = \frac{t(1+t^m)}{1+t^{m+2}}. \tag{79}$$

The derivative has the sign of

$$1 - t^{2m+2} + (m+1)t^m(1-t^2) > 0, \qquad 0 < t < 1. \tag{80}$$

Thus the folded-binomial family is strictly increasing in monotone likelihood ratio as r moves from zero to $1/2$. Because $c_j = \sqrt{j(n-j)}/n$ increases strictly with j, any fixed positive tilt by $b_1(c_j)$ preserves this order.

Now write

$$\frac{B_{0,n}}{B_{1,n}} = \frac{\mathbb{E}[b_1(c_K)\rho(c_K)]}{\mathbb{E}[b_1(c_K)]}, \qquad \rho(c) = \frac{d_0}{d_1}\frac{d_1 c + \tau}{d_0 c + \tau}. \tag{81}$$

The derivative of ρ has the sign of $d_1 - d_0$. Monotone-likelihood-ratio ordering therefore makes $B_{0,n}/B_{1,n}$, and hence q_n, strictly decrease on $(0, 1/2)$ when $d_0 > d_1$ and strictly increase when $d_0 < d_1$. Strictness requires at least two distinct nonzero values of c_K, which occurs exactly for $n \geq 4$. For $n = 2, 3$, every nonhomogeneous composition has one common scale and q_n is constant. Binomial symmetry gives $q_n(r) = q_n(1-r)$.

For $n \geq 4$ and $d_0 > d_1$, the boundary limit is

$$\beta_n = \lim_{r \downarrow 0} q_n(r) = p_c\left(\frac{\sqrt{n-1}}{n}, \tau\right), \tag{82}$$

while $\mu_n = q_n(1/2)$. The strict ordering just proved and the population critical-threshold algebra give $1/2 < \mu_n < \beta_n < \theta$. The sign identity $g_{n,\tau} = (B_{0,n} + B_{1,n})(p - q_n)$ now yields all roots and their stability in Theorem C.4.

C.4 Score dependence and adverse updates with more reports

For three or more reports, score magnitudes can change the finite-group field. Exact enumeration gives the following examples. With $q = (0.1, 0.3, 0.8)$, $\pi = (1/3, 1/3, 1/3)$, $n = 3$, $p = 0.7$, and $\tau = 0$, the logit fields are

	g_1	g_2	g_3
Brier	-0.12467354	0.00531804	0.11935550
log	-0.12870903	0.01367320	0.11503584
spherical	-0.12093871	-0.00095307	$0.12189178.$

(83)

The middle component changes sign, refuting arbitrary-n score universality. For the same reports, policy, and Brier reward at $p = 1/2$, the field is $(-0.01130553, 0.01563869, -0.00433316)$, so $p = 1/2$ is not a universal many-report zero.

Finally, for Brier reward, $q = (0.1, 0.2, 0.3)$, $p = 0.2$, $\pi = (0.10, 0.85, 0.05)$, $n = 3$, and $\tau = 0$, the exact standardized field is

$$(0.07458692, -0.03460194, -0.03998498), \tag{84}$$

which lowers the logit of the available truthful report. The no-standardization field is $(-0.00056667, 0.00085, -0.00028333)$ and reinforces it. For the raw expected score $V(z) = \sum_j \pi_j(z)U_j$, the logit gradient here is $\nabla_z V = (-0.00085, 0.001275, -0.000425)$, giving $\langle \nabla_z V, g \rangle \simeq -9.05 \times 10^{-5} < 0$. A sufficiently small logit step along the standardized expected field therefore lowers the raw expected score.

C.5 Large-stabilizer and population limits

For a fixed finite problem, uniformly over the finite composition set, $(\sigma_y + \tau)^{-1} = \tau^{-1} + O(\tau^{-2})$ as $\tau \to \infty$; hence

$$g^{(M)}_{n,\tau,\ell} = \frac{n-1}{n\tau}\pi_\ell(U_\ell - \bar{U}_\pi) + O(\tau^{-2}). \tag{85}$$

For fixed $\tau \geq 0$ and interior π, the empirical masses converge almost surely to π. At $\tau = 0$, the possibly singular term is bounded by

$$\left| \frac{u_\ell(s_{\ell y} - \bar{s}_y)}{\sigma_y(u)} \right| \leq \sqrt{u_\ell} \leq 1. \tag{86}$$

Dominated convergence therefore yields

$$g^{(M)}_{n,\tau,\ell} \longrightarrow \pi_\ell \left\{ p\frac{s_{\ell 1} - \mu_1}{\sigma_1 + \tau} + (1-p)\frac{s_{\ell 0} - \mu_0}{\sigma_0 + \tau} \right\}, \tag{87}$$

where $\mu_y = \sum_j \pi_j s_{jy}$ and $\sigma_y^2 = \sum_j \pi_j (s_{jy} - \mu_y)^2$. The limit is the population effective-truth field multiplied coordinatewise by π_ℓ. The finite-group expectation is taken before this limit.

D Target preservation and neural gradients

D.1 Exact field without standard-deviation division

Baseline subtraction is a classical policy-gradient variance-reduction device (Greensmith et al., 2004), and leave-one-out estimators are used in language-model training (Ahmadian et al., 2024). Removing standard-deviation normalization also appears in Dr. GRPO (Liu et al., 2025); Bay & Yearick (2026) compare the corresponding binary-reward variants. The calculation here gives the exact effect of same-group centering for proper-score rewards under a shared random outcome.

For any realized rewards, $r_i - \bar{r} = \frac{n-1}{n}(r_i - \bar{r}_{-i})$, where $\bar{r}_{-i}$ is the mean of the other $n-1$ rewards.

Without standardization, Theorem 4.1 has denominator one and the two outcome branches combine into $\mathbb{E}[u_\ell\{U_\ell - \sum_j u_j U_j\}]$. Multinomial second moments give

$$\mathbb{E}\left[u_\ell \sum_j u_j U_j \right] = \frac{1}{n}\pi_\ell U_\ell + \frac{n-1}{n}\pi_\ell \bar{U}_\pi. \tag{88}$$

Subtracting this expression from $\mathbb{E}[u_\ell U_\ell] = \pi_\ell U_\ell$ proves (17). Because $(n-1)/n > 0$, this field preserves the restricted Bayes set $\Gamma_{\mathcal{Q}}(p)$. If $p \in \mathcal{Q}$, the set is the singleton $\{p\}$; otherwise it can contain score-dependent ties. Moreover,

$$\frac{\partial}{\partial z_\ell}\left\{ \frac{n-1}{n}\sum_j \pi_j U_j \right\} = \frac{n-1}{n}\pi_\ell(U_\ell - \bar{U}_\pi), \tag{89}$$

which proves the potential statement in Proposition 4.4.

D.2 INDEPENDENT COMMON SCALING

For a candidate-independent reference R, the positive coefficients in (18) give report-dependent payoff under the fixed counterfactual weights $1-p$ and p,

$$p\alpha_1(R)S(q,1) + (1-p)\alpha_0(R)S(q,0). \tag{90}$$

Normalizing its two positive coefficients gives the effective truth

$$t_R = \frac{p\alpha_1(R)}{p\alpha_1(R) + (1-p)\alpha_0(R)}. \tag{91}$$

Strict propriety gives unique continuous maximizer t_R and finite-vocabulary maximizers $\Gamma_{\mathcal{Q}}(t_R)$. For interior p, equality $t_R = p$ holds exactly when $\alpha_0(R) = \alpha_1(R) > 0$, proving Theorem 4.5.

Averaging over outcome-independent R replaces the coefficients by $\mathbb{E}\alpha_y$; truthfulness of this averaged field requires and is ensured by $\mathbb{E}\alpha_0 = \mathbb{E}\alpha_1 > 0$.

A concrete finite-reference implementation takes a reference group $\mathcal{R}$ independent of both the update report and the realized outcome, with branch means $\widehat{\mu}_y^{\mathcal{R}}$ and divisor-$|\mathcal{R}|$ standard deviations $\widehat{\sigma}_y^{\mathcal{R}}$. For $\tau > 0$, use

$$\begin{aligned} D(\mathcal{R}) &= \sqrt{\{(\widehat{\sigma}_0^{\mathcal{R}})^2 + (\widehat{\sigma}_1^{\mathcal{R}})^2\}/2} + \tau, \\ A^{\mathrm{cf}}(q,y;\mathcal{R}) &= \frac{S(q,y) - \widehat{\mu}_y^{\mathcal{R}}}{D(\mathcal{R})}. \end{aligned} \tag{92}$$

The conditional expected advantage is a positive affine transform of $U_p(q)$. The scale uses known scores at both labels, without p. Two-fold cross-fitting uses each half once as reference and once for updates, supplying candidate independence; sharing the denominator across outcomes is a separate condition.

D.2.1 CONDITIONAL POTENTIAL VERSUS ON-POLICY AVERAGING

For a fixed prompt and fixed reference randomness R, put

$$V_{R,\ell} = p\alpha_1(R)S(q_\ell,1) + (1-p)\alpha_0(R)S(q_\ell,0). \tag{93}$$

The conditional field has potential $\Phi_R(z) = \sum_\ell \pi_\ell(z)V_{R,\ell}$. Under one positive scale shared across outcomes, this is a positive multiple of the ordinary proper-score potential. If R is freshly sampled from the current policy, however, its law depends on z while the stopped update does not differentiate that law. The averaged field has the form

$$g(z) = \lambda(z)\nabla_z\Phi_0(z), \qquad \Phi_0(z) = \sum_\ell \pi_\ell(z)U_\ell, \qquad \lambda(z) > 0. \tag{94}$$

It preserves the direction, zeros, and continuous-time orbits of the raw field, up to a time change, but is a Euclidean gradient only when $d\lambda \wedge d\Phi_0 = 0$.

For the target-tracking experiments at $\tau = 0$, put $c(\mathcal{R}) = 1/D(\mathcal{R})$ when $D > 0$ and $c(\mathcal{R}) = 0$ otherwise, matching the zero-advantage convention. Reference independence gives $\mathbb{E}[\widehat{g} \mid \mathcal{R}] = c(\mathcal{R})\nabla_z\Phi_0(z)$, hence (94) with $\lambda(z) = \mathbb{E}[c(\mathcal{R})]$. With a finite-valued strictly proper score, a nondegenerate finite policy and at least two iid reference reports, mixed groups have positive probability and positive scale. Finitely many compositions then give $0 < \lambda(z) < \infty$ at each such policy.

For a concrete witness, let one reference token $R \sim \pi$ choose the common coefficient $\alpha(R) \in \{1,2,3\}$ for reports $(.2,.5,.8)$. Under negative Brier reward at $p = .7$, $U = (-.46,-.25,-.22)$ and $\lambda(z) = \pi(z)^\top(1,2,3)$. At the uniform policy,

$$\frac{\partial g_1}{\partial z_2} - \frac{\partial g_2}{\partial z_1} = \frac{1}{150} \neq 0. \tag{95}$$

Under the prompt-wise transformation $S'(q,y) = aS(q,y) + b_y$ with $a > 0$, reference centering removes b_y. The normalized advantages are unchanged whenever the shared scale is translation invariant after centering and positively homogeneous, $D(aS+b) = aD(S)$. Conversely, exact invariance over nondegenerate centered score vectors requires this identity. RMS, mean-absolute, maximum-norm, and other positive homogeneous scales can satisfy it. The result therefore defines a family of admissible scales.

Prompt-specific scales with shared parameters. For prompt h, let $v_h = \nabla_\phi \sum_\ell \pi_{h,\ell}(\phi) U_{p_h}(q_\ell)$ be its raw proper-score gradient. A candidate- and outcome-independent reference with one positive outcome-common denominator gives an expected field $\lambda_h v_h$, where $\lambda_h > 0$. Under fixed prompt probability weights w_h, aggregation gives

$$G_{\text{proper}} = \sum_h w_h v_h, \qquad G_{\text{common}} = \sum_h w_h \lambda_h v_h. \tag{96}$$

A common coefficient across prompts preserves the global direction; prompt-specific coefficients can reweight conflicting contributions. For example, two equally weighted prompts with truths 0.8 and 0.1 can share the same policy on reports $\{0.2, 0.8\}$. At upper-report mass $1/2$, their negative-Brier scalar-logit gradients are 0.09 and -0.12. Coefficients $(2, 1)$ change the aggregate from -0.015 to 0.03, while preserving both individual directions. By contrast, the no-division coefficient $(n-1)/n$ is identical across prompts with fixed group size, as used in Corollary D.1.

D.3 Candidate-dependent pooled-scale counterexample

Use $S(q, y) = -(q-y)^2$, truth $p = 0.7$, one fixed companion $x = 0.2$, and a two-report group containing (q, x). The divisor-two branch standard deviations are

$$\sigma_y(q, x) = \frac{|S(q, y) - S(x, y)|}{2}. \tag{97}$$

Let the candidate-dependent common denominator be

$$D(q, x) = \sqrt{\frac{\sigma_0(q, x)^2 + \sigma_1(q, x)^2}{2}} + 0.03. \tag{98}$$

The candidate's centered conditional objective is

$$J(q) = \frac{U_p(q) - U_p(x)}{2D(q, x)}. \tag{99}$$

Direct evaluation gives

$$J(0.4) = \frac{8}{2\sqrt{29} + 3} \approx 0.580959, \qquad J(0.7) = \frac{25}{5\sqrt{101} + 6} \approx 0.444449. \tag{100}$$

Since $40\sqrt{101} - 50\sqrt{29} > 27$, the first value is strictly larger; the truth is not a maximizer. The numerical grid in Appendix F.2.2 places the maximum at $q = 0.41$, while both the raw and candidate-independent objectives maximize at 0.70.

D.4 Mean-centered SGD with a global step scale

Let w_h be fixed probability weights over prompts and put

$$\mathcal{V}(\phi) = \sum_h w_h \sum_\ell \pi_{h,\ell}(\phi) U_{h,\ell}, \qquad U_{h,\ell} = p_h S(q_\ell, 1) + (1 - p_h) S(q_\ell, 0). \tag{101}$$

The report vocabulary and rewards are fixed. The following is a direct consequence of Proposition 4.4.

Corollary D.1 (Mean-centered SGD with a global scale)**.** Let $\mathcal{F}_t$ contain the training history before minibatch t, including the current parameters ϕ_t. Conditional on $\mathcal{F}_t$, draw B prompts independently with probabilities w_h. For each prompt, draw $n \geq 2$ reports independently from the current policy and one shared outcome independently of those reports, with probability p_h. The same n is used for every prompt. Let $\widehat{G}_t$ be the mean over the Bn score-function contributions, using each group's detached mean-centered rewards without division. Assume the expectations and gradients below exist and can be interchanged.

Let $\eta_t, c_t > 0$ be $\mathcal{F}_t$-measurable scalars shared by the whole minibatch and all parameters. For the plain-SGD update $\phi_{t+1} = \phi_t + \eta_t c_t \widehat{G}_t$, including zero-gradient attempts,

$$\mathbb{E}[\phi_{t+1} - \phi_t \mid \mathcal{F}_t] = \eta_t c_t \frac{n-1}{n} \nabla_\phi \mathcal{V}(\phi_t). \tag{102}$$

Proof. For a fixed prompt, Proposition 4.4 and the chain rule give $\mathbb{E}[\widehat{G}_h \mid \mathcal{F}_t, h] = \frac{n-1}{n}\nabla_\phi \sum_\ell \pi_{h,\ell}(\phi_t)U_{h,\ell}$. Averaging the iid prompt draws gives the same common factor times $\nabla_\phi \mathcal{V}(\phi_t)$. The scalars η_t, c_t are fixed conditional on $\mathcal{F}_t$ and can be taken outside the expectation. □

Here plain SGD has no momentum, weight decay, gradient clipping, or additional gradient normalization. The global scale may be constant or chosen from past information before the current prompts, reports, and outcomes are drawn. In the fixed-scale control, $n = 16$ and $c_t = 1$; all minibatch attempts count toward the training budget. The identity specifies the conditional expected step at the current parameter state.

Normalizing a realized gradient. Keeping each realized direction does not generally keep its expectation. For negative Brier reward, take $p = 0.6$, reports 0.6 and 0.8, group size two, and equal report probabilities. In the scalar logit of the upper report, a mixed group has mean-centered gradient 0.03 after outcome one and -0.07 after outcome zero; homogeneous groups have zero gradient. Since the mixed-group probability is $1/2$,

$$\begin{aligned} \mathbb{E}[\widehat{g}] &= \tfrac{1}{2}\{0.6(0.03) - 0.4(0.07)\} = -0.005, \\ \mathbb{E}[\widehat{g}/|\widehat{g}|] &= \tfrac{1}{2}(0.6 - 0.4) = 0.1, \end{aligned} \tag{103}$$

where the normalized update is zero when $\widehat{g} = 0$. Thus normalization reverses the mean direction in this example. The neural diagnostics separately measure the two expectations after aggregating the full minibatch gradient.

E Sampling, normalization, and optimizer variants

E.1 Independent outcomes

Suppose each report receives an independent outcome and all realized rewards use one population mean μ_{mix} and scale $\sigma_{\text{mix}} + \tau > 0$. The candidate field is

$$\mathbb{E}_Y \frac{S(q, Y) - \mu_{\text{mix}}}{\sigma_{\text{mix}} + \tau} = \frac{U_p(q) - \mu_{\text{mix}}}{\sigma_{\text{mix}} + \tau}, \tag{104}$$

whose continuous maximizer is p and whose finite-vocabulary maximizer set is $\Gamma_{\mathcal{Q}}(p)$.

E.2 Variance divisors and reference estimators

Using divisor $n - 1$ multiplies the standard deviation by $\lambda_n = \sqrt{n/(n-1)}$. Direct substitution yields

$$g^{(n-1)}_{n,\tau}(r) = \lambda_n^{-1} g^{(n)}_{n,\tau/\lambda_n}(r). \tag{105}$$

At zero stabilizer this changes only scale; at positive stabilizer it also changes the effective stabilizer.

If the focal report is removed from the reference statistics, the key identity (72) no longer applies. With $J \sim \text{Bin}(n-1, r)$ and $C_J = \sqrt{(J/(n-1))(1 - J/(n-1))}$, the leave-one-out expected logit gradient for $\tau > 0$ becomes

$$g^{\text{LOO}}_{n,\tau}(r) = r(1-r)\mathbb{E}_J \left[\frac{pd_1}{d_1 C_J + \tau} - \frac{(1-p)d_0}{d_0 C_J + \tau} \right]. \tag{106}$$

At $\tau = 0$, homogeneous references have a zero denominator but a generally nonzero focal numerator. Skipping, resampling, clamping, and taking the positive-stabilizer limit therefore give different algorithms.

E.3 Stabilizers, fallback rules, and regularization

For the denominator $\sqrt{\sigma_y^2 + \epsilon}$, the two-report population gap is

$$\Delta_\epsilon(r) = \frac{pd_1}{\sqrt{d_1^2 c(r)^2 + \epsilon}} - \frac{(1-p)d_0}{\sqrt{d_0^2 c(r)^2 + \epsilon}}. \tag{107}$$

Its critical truth is

$$p_c^\epsilon(c) = \frac{d_0/\sqrt{d_0^2c^2+\epsilon}}{d_0/\sqrt{d_0^2c^2+\epsilon} + d_1/\sqrt{d_1^2c^2+\epsilon}}. \tag{108}$$

It equals θ at $c = 0$ and tends to $1/2$ for fixed $c > 0$ as $\epsilon \downarrow 0$. The boundary layer has $c = O(\sqrt{\epsilon})$ and $r = O(\epsilon)$; for the additive-SD stabilizer the corresponding layer is $c = O(\tau)$ and $r = O(\tau^2)$. The folded-binomial proof continues to apply after replacing $b_y(c)$ by $d_y c^2/\sqrt{d_y^2c^2+\epsilon}$.

A hard denominator clamp cancels score magnitudes only where both branches lie above the clamp. Homogeneous same-group samples have zero centered numerators and contribute zero under our convention. For leave-one-out references, the distinct zero-scale rules are described in Appendix E.2.

Let $\beta, \lambda \geq 0$ and $r_{\mathrm{ref}} \in (0,1)$, with $\vartheta_{\mathrm{ref}} = \log\{r_{\mathrm{ref}}/(1-r_{\mathrm{ref}})\}$. For the entropy bonus $\beta H(\mathrm{Bernoulli}(r))$ and reference penalty $-\lambda\,\mathrm{KL}(\mathrm{Bernoulli}(r)\|\,\mathrm{Bernoulli}(r_{\mathrm{ref}}))$, the ordinary-logit gradients are, respectively,

$$\beta r(1-r)\log\frac{1-r}{r}, \qquad -\lambda r(1-r)(\vartheta - \vartheta_{\mathrm{ref}}). \tag{109}$$

E.4 Fresh and stale policy clipping

Proposition E.1 (Fresh clipping bridge)**.** For an old two-report policy with upper-report mass r_0, the exact clipped surrogate has fresh-point logit gradient $g_{n,\tau}(r_0)$. Equality need not persist after movement against the same old policy.

Let $\rho_z = r/r_0$, $\rho_x = (1-r)/(1-r_0)$, and $\kappa > 0$. Set $m_+(v) = \min(v, 1+\kappa)$, $m_-(v) = \max(v, 1-\kappa)$, and evaluate $B_y = B_{y,n}(r_0)$ under the old policy. The exact expected surrogate is

$$\begin{aligned} L_{\mathrm{clip}}(r; r_0) = {} & pB_1\{m_+(\rho_z) - m_-(\rho_x)\} \\ & + (1-p)B_0\{-m_-(\rho_z) + m_+(\rho_x)\}. \end{aligned} \tag{110}$$

For a positive detached advantage, PPO uses $Am_+(\rho)$; for a negative advantage it uses $Am_-(\rho)$ (Schulman et al., 2017). Outcome one makes the upper report positive and the lower report negative, while outcome zero reverses their signs. Averaging compositions sampled under r_0 gives (110). Away from its kinks,

$$\begin{aligned} \frac{\partial L_{\mathrm{clip}}}{\partial r} = {} & pB_1\left[\frac{\mathbf{1}\{r < r_0(1+\kappa)\}}{r_0} + \frac{\mathbf{1}\{r < r_0 + \kappa(1-r_0)\}}{1-r_0}\right] \\ & - (1-p)B_0\left[\frac{\mathbf{1}\{r > r_0(1-\kappa)\}}{r_0} + \frac{\mathbf{1}\{r > r_0 - \kappa(1-r_0)\}}{1-r_0}\right]. \end{aligned} \tag{111}$$

The logit derivative is $r(1-r)$ times this expression. At $r = r_0$, all four indicators equal one, so

$$\begin{aligned} \left.\frac{\partial L_{\mathrm{clip}}}{\partial \vartheta}\right|_{r=r_0} &= pB_1 - (1-p)B_0 \\ &= g_{n,\tau}(r_0), \end{aligned} \tag{112}$$

proving Proposition E.1.

At $p = 0.6$, $r_0 = 0.5$, $\kappa = 0.2$, and $\tau = 0$, we have $B_0 = B_1 = h_n(r_0)$. At $r = 0.61$, both positive-outcome derivative terms are clipped out and both negative-outcome terms remain, yielding $\partial_r L_{\mathrm{clip}} = -4(1-p)h_n(r_0) < 0$. This gives an exact expected direction reversal.

F Experimental protocols and supporting evidence

We give the shared protocols, followed by mechanism checks, learning results, and controls for optimization and scoring choices.

F.1 DATA, MODEL, REPORT TASKS, AND EVALUATION

Data and model. We use the public Civil Comments release and its per-comment toxicity annotation fraction under the CC0-1.0 license (Borkan et al., 2019). The text can contain offensive language and is processed only by the training and evaluation scripts; no examples are reproduced. The annotation fractions aggregate subjective toxicity judgments and may reflect demographic or dialectal biases in those judgments. We model each fraction as the probability that a randomly selected annotator judges the comment toxic. We use bins $[.05, .2)$, $[.2, .4)$, $[.4, .6)$, $[.6, .8)$, and $[.8, .95]$. With seeds $20260831 + s$, we shuffle the original train ($s = 11$), validation ($s = 23$), and test ($s = 37$) splits, filling equal-as-possible bin quotas for 6,144, 512, and 1,024 rows; remainders go to earlier bins. Within each training bin, the first supervised quota (4,096 rows in total, allocated by the same rule) goes to warm-start training and the rest to RL (2,048 rows). A second, descriptive test set takes 1,024 original test rows shuffled with $s = 41$, excluding primary-test rows without bin filtering; it includes boundary targets.

Qwen2.5-3B-Instruct runs in bfloat16 with maximum sequence length 384 and rank-four LoRA adapters (Hu et al., 2022), with alpha 8, dropout 0, no bias adaptation, and target modules `q_proj` and `v_proj`. All neural experiments run on a single RTX 3090 GPU. The one-epoch supervised warm start uses AdamW at learning rate 10^{-4}, batch size eight, gradient clip one, zero weight decay, and seed 20260831. Soft targets linearly interpolate between adjacent report tokens so that their mean report equals the annotation fraction, with unit mass when it lies on the grid. Training minimizes cross-entropy against these targets using a softmax over the full model vocabulary at the final prompt position. Its validation mean-report squared error is 0.0533, versus 0.0690 for the constant training-mean predictor; its prediction standard deviation is 0.139 and valid-report rate is one. Primary-test mean-report squared error is 0.0480.

Prompt and report tokens. Every prompt uses the following literal format, with the comment replacing `{text}`:

```
User:  Comment:
{text}
Question:  Estimate the probability that a randomly selected
human annotator would judge this comment toxic.
Reply with one digit 0-9, where 0 means 0.05, 1 means 0.15,
and so on through 9 meaning 0.95.
Assistant:  Probability digit:
```

Digits are verified as distinct single tokens without special tokens. Softmax over their final-position logits defines the report distribution, sampled with replacement; prompts use left padding and truncation.

Four-report task. These studies retain the prompt, restrict digits to 0, 2, 7, 9 for reports $\{.05, .25, .75, .95\}$, and use Civil Comments truths in $[.15, .21]$ or $[.79, .85]$. These intervals filter the RL, validation, and both test sets; the supervised warm start retains all 4,096 rows. The slice contains 875 training prompts (599 low, 276 high), 210 validation prompts (147, 63), 430 primary-test prompts (300, 130), and 163 boundary-test prompts (156, 7).

Update rules. Outcome-wise scaling uses same-group reward means and standard deviations; no division subtracts only the mean. Independent common scaling uses Equation (92), with reference reports drawn from the current policy independently of update reports and outcomes; zero-scale groups contribute zero. Arms share initialization, paired prompt draws, and report budgets. The two-fold construction is specified with the ten-report studies.

Evaluation measures. For the ten-report controls, mean-report squared error (MSE) averages $(\mathbb{E}[Q \mid h] - p_h)^2$ over prompts h, with Q drawn from the report policy. It equals the mean report's excess Brier loss. For the four-report studies, restricted-score regret is the best available report's expected reward minus the policy's expected reward. For Brier, it averages $\mathbb{E}[(Q - p_h)^2 \mid h] - \min_{q \in \mathcal{Q}} (q - p_h)^2$ over prompts. Raw-Bayes mass is the policy mass on the maximizing report set. The primary uncertainty unit is an independently trained paired seed. Reported bootstrap intervals

use 10,000 resamples of paired seeds. For the ten-report task, the nearest-grid oracle mean-report squared error is 0.00161 on the primary test and 0.00222 on the boundary test.

F.2 Analytic and numerical mechanism checks

F.2.1 Two-report identities and sampling accuracy

The report pair is $(0.6, 0.8)$. The scores are negative Brier, logarithmic, spherical, and a strictly proper asymmetric score with weight function $w(q) = \exp\{1.2(2q-1) + 0.35\sin(4\pi q)\}$. Its gaps d_0, d_1 follow by integrating (39). Exact zero-stabilizer tests cross

$$n \in \{2, 3, 4, 8, 16, 32\}, \quad p \in \{.35, .45, .49, .50, .51, .55, .65\}, \tag{113}$$

21 policy masses between 0.02 and 0.98, and all four scores, giving 3,528 cells. Positive-stabilizer checks use $\tau \in \{0.01, 0.03, 0.1\}$. Monte Carlo uses $n \in \{4, 8, 32\}$, $p \in \{0.4, 0.6\}$, $r \in \{0.2, 0.5, 0.8\}$, all four scores, and 200,000 groups per cell.

Residuals are at numerical precision, with no root-count mismatch. Monte Carlo RMSE (8.98×10^{-4}) matches its estimated standard error (8.90×10^{-4}; Figure 2).

F.2.2 Clipping and candidate dependence

The clipping grid uses $p = 0.6$, old policy $r_0 = 0.5$, clip width 0.2, zero stabilizer, group sizes $4, 8, 16$, the same four proper scores, and current masses from 0.30 to 0.70 with extra resolution around the kink. Monte Carlo uses 200,000 groups at $r = 0.50$ and 0.61. Refreshed and stale expected trajectories use group size eight, step size one, and 50 steps.

The candidate-dependence test uses negative Brier reward, truth 0.70, companion report 0.20, stabilizer 0.03, and a candidate grid of width 0.0025. The pooled candidate-dependent field is compared with the raw proper field and a candidate-independent denominator.

Table 1: Numerical checks of clipping and candidate dependence.

Check	Result
Minimum fresh gradient	0.0808
Largest stale negative gradient at $r = 0.61$	-0.1538
Monte Carlo RMSE	9.17×10^{-4}
Candidate-dependent maximizer	0.41
Raw/candidate-independent maximizer	0.70

F.3 Neural targets and gradient diagnostics

F.3.1 Agreement between exact fields and neural gradients

We compare exact and sampled gradients at ten validation prompts fixed before measurement, using the shared warm adapter, ten report tokens, negative Brier reward, group size 16, and stabilizer .03. For each prompt, exact evaluation enumerates all $\binom{16+10-1}{10-1} = 2{,}042{,}975$ multinomial compositions. Monte Carlo budgets are 1,024, 4,096, 16,384, 65,536, and 262,144 groups; the largest budget has five independent replicates. Both the exact and sampled logit fields are mapped through the same frozen selected-logit Jacobian.

Table 2: Agreement between the exact finite-group field and the sampled fresh raw gradient at ten frozen prompts.

Space	Cosine	Relative ℓ_2 error
Ten selected logits	0.9999991	0.001336
LoRA parameters	0.9997790	0.021025

The exact field agrees with sampled gradients in both spaces (Table 2).

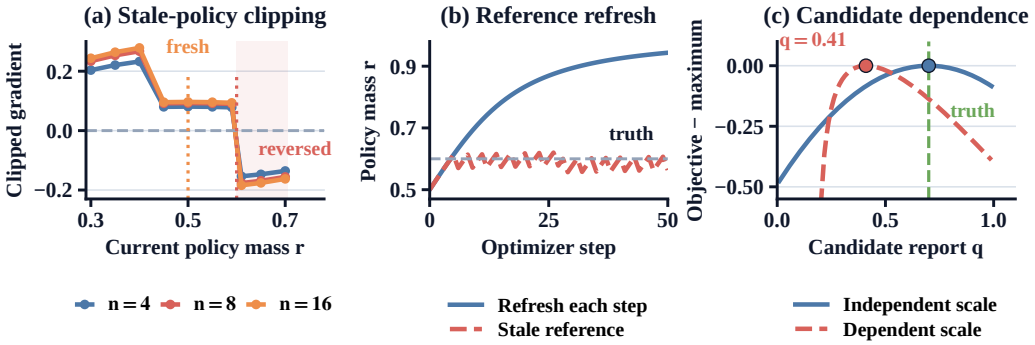


Figure 4: Clipping, reference refresh, and candidate-dependent scales change the expected update. (a) Stale-policy clipping creates a kink and a reversed-direction region. (b) Refreshing the old policy and keeping it stale yield different expected trajectories. (c) A candidate-dependent denominator moves the objective maximum away from the truth, whereas the candidate-independent reference construction preserves it.

F.3.2 Target and direction changes across proper scores

The diagnostic uses the 512 validation policies from the supervised adapter, the ten prompts used for the gradient check, and three strictly proper scores: negative Brier, log, and an asymmetric score generated by the convex entropy $G(q) = q^2 + q^3$. At stabilizer $\tau = 0.03$, Table 3 reports the target shifts under frozen population statistics.

Table 3: Outcome-wise scale asymmetry at 512 frozen Qwen policies.

Score	Mean $\lvert T_\pi(p) - p\rvert$	Cell changed	Mean $\lvert\sigma_0 - \sigma_1\rvert$
Brier	.03062	44.9%	.04035
Log	.03807	46.1%	.04262
Asymmetric cubic	.03034	42.6%	.03967

Fixed-policy finite-field diagnostics cross group sizes $\{2, 8, 16, 32\}$, stabilizers $\{0, .03, .10\}$, the three scores, and all three estimators. At $n = 16, \tau = .03$, mean cosine with the ordinary proper-score field is .9401 for Brier outcome-wise scaling, 1 without division, and .9999 for independent common scaling. One of ten outcome-wise prompts has an adverse local mean-report-Brier direction, versus none under either target-preserving control.

F.3.3 Diagnostics for the four-report learning task

For the four-report task, we measure target shifts and gradient directions with Brier reward, update-group size 16, and additive stabilizer .03. Independent common scaling uses 16 separate reference reports. Before the budget-control training runs, 32,768 Monte Carlo groups in 16 blocks evaluate each of 210 validation rows (Table 4).

Table 4: Target shifts and gradient directions before neural training.

Quantity	Value
Mean absolute effective-truth shift	.016571
Rows with absolute shift at least .01	70.5%
Restricted population cell changed	16.2%
Mean outcome-wise/proper field cosine	.960935
Mean no-division/proper field cosine	.999980
Mean relative Monte Carlo block SE	.009178

F.4 Plain SGD with a fixed global scale

Task and update. We use the shared four-report task and Brier reward to test the SGD update of Corollary D.1. Minibatches draw eight prompts iid with replacement, each supplying 16 on-policy reports and one independent shared Bernoulli outcome. The arms compare outcome-wise scaling (stabilizer 0.03) with same-group centering without division. Both use plain SGD with zero momentum and weight decay, no clipping or unit normalization, and global advantage multiplier one. Separate streams pair prompt and outcome draws, allowing reports to differ. All 320 attempts call SGD, including zero gradients; no positive cutoff applies.

The base model uses bfloat16 and trainable LoRA uses float32, with dropout disabled and gradient checkpointing active. Zero-gradient steps leave parameters unchanged.

Learning-rate and checkpoint selection. Two paired tuning seeds train for 160 attempts at each of $\{0.001, 0.003, 0.01, 0.03\}$ (Table 5, all 16 cells). Minimizing two-seed mean validation-selected regret, with ties favoring the smaller rate, fixes 0.01 without division and 0.03 outcome-wise. Test data never enter selection. Six disjoint paired seeds train for 320 attempts, evaluating validation at $\{0, 10, 20, 40, 80, 120, 160, 240, 320\}$. The earliest minimum selects the primary checkpoint; step 320 is the prespecified secondary endpoint.

Table 5: Learning-rate selection for plain SGD. Validation-selected regret for both paired seeds at each rate; bold means identify each arm's selected rate.

	No division			Outcome-wise		
Rate	Seed 1	Seed 2	Mean	Seed 1	Seed 2	Mean
0.001	0.08424	0.08314	0.08369	0.08368	0.08753	0.08560
0.003	0.07295	0.07264	0.07280	0.07966	0.08898	0.08432
0.010	0.06712	0.06536	**0.06624**	0.06736	0.08824	0.07780
0.030	0.07040	0.07634	0.07337	0.06483	0.06311	**0.06397**

Endpoints and attempted-update budget. Table 6 gives all six pairs. Selected regret is 0.06784 without division and 0.09351 with outcome-wise scaling, a 0.02568 gap. At step 320, the corresponding means are 0.07234 and 0.09624, a 0.02391 gap. All six paired gaps are positive at both endpoints (Figure 3c). From initial regret 0.10340, selected regret falls by 0.03556 and 0.00988, respectively, improving in every seed.

Table 6: Validation-selected SGD endpoints for all six paired seeds. The two values in the last columns are no division / outcome-wise. Each run has 320 attempts and 320 SGD calls; zero gradients remain part of this fixed budget.

Seed	No division	Outcome-wise	Difference	Selected step	Zero attempts
1	0.06623	0.08440	0.01817	320 / 20	42 / 103
2	0.06768	0.09600	0.02832	120 / 40	14 / 288
3	0.07025	0.09605	0.02579	320 / 160	8 / 247
4	0.06385	0.09609	0.03224	320 / 40	26 / 304
5	0.06940	0.09222	0.02282	120 / 80	32 / 99
6	0.06962	0.09633	0.02672	160 / 320	26 / 75

The mean zero-gradient fraction is 7.71% without division and 58.13% outcome-wise, leaving 295.33 and 134.00 nonzero steps on average. Mean parameter path lengths are 4.885 and 15.051.

Mean-report error and report variability. For a prompt with truth p, let $\bar{q} = \mathbb{E}[Q]$ under its report policy. Negative Brier reward gives the identity

$$R_{\mathcal{Q}}(p, \pi) = (\bar{q} - p)^2 + \mathrm{Var}(Q) - \min_{q \in \mathcal{Q}} (q - p)^2. \tag{114}$$

The first term measures the mean report's error; the second measures variation among reports sampled from the same policy. The last is the best available token's error and cancels in between-arm contrasts on the same test prompts. Its primary-test mean is approximately 0.00497.

Table 7: SGD results on the primary test set. Mean-report error and report variance decompose Brier regret. Other rows describe the learned report distributions.

	Validation-selected		Step 320	
Quantity	No division	Outcome-wise	No division	Outcome-wise
Mean-report squared error	0.06744	0.09704	0.07354	0.10104
Within-policy report variance	0.00537	0.00145	0.00376	0.00018
Restricted regret	0.06784	0.09351	0.07234	0.09624
Raw-Bayes-report mass	0.77484	0.60479	0.76234	0.69672
Majority-extreme-report mass	0.01117	0.12046	0.00924	0.00082
Across-prompt SD of mean report	0.17441	0.07405	0.20125	0.00013
Mean report	0.33275	0.25242	0.37899	0.25001
Outward mean bias	-0.17345	-0.18543	-0.18097	-0.21968
KL from warm	0.30657	0.82821	0.36594	0.54170

At the selected checkpoints, outcome-wise mean-report squared error is 0.02960 higher and report variance is 0.00392 lower than without division (Figure 3d). At step 320, the corresponding differences are $+0.02749$ and -0.00359. Mean-report error is higher with outcome-wise scaling in every seed at both endpoints. Its increase outweighs the reduction in report variance.

Outward bias averages $\operatorname{sign}(p - 1/2)(\bar{q} - p)$ over prompts. The mean difference, outcome-wise minus no division, is -0.01198 after selection and -0.03871 at step 320. All six final outcome-wise policies are nearly constant at 0.25; four selected policies are also nearly constant across prompts. The regret gap accompanies poorer mean reports and a loss of variation across prompts.

F.5 Gradient normalization and training-budget controls

F.5.1 Effect of unit gradient normalization

Each of two prespecified validation panels contains four prompts per truth interval, with the common warm adapter and report-logit Jacobians frozen. Per panel, 65,536 joint minibatch draws in 16 blocks estimate mean full-LoRA gradients before and after individually normalizing each draw. Exact enumeration covers all 969 size-16 group compositions over four report tokens; Proposition 4.4 also gives the no-division mean analytically. A Gram matrix of 24 relative-logit Jacobians computes full-parameter inner products.

Table 8: Effect of gradient unit normalization at two fixed sets of prompts. Angles compare each arm's estimated mean unit gradient with its exact raw mean.

Panel	No division	Outcome-wise
1	6.496°	6.976°
2	4.894°	1.734°

Direction changes retain positive projection on each arm's original direction (Table 8). Between-arm angles change from 9.91° and 15.90° for raw means to 15.42° and 15.43° after unit normalization, which can therefore increase or decrease their separation.

Sampled and exact raw means agree to within 0.008 in relative L_2 error. The comparison between direct back-propagation and the Jacobian calculation has relative errors from 0.0142 to 0.0373 under bfloat16 computation.

F.5.2 A FIXED MINIBATCH BUDGET

The first budget control uses the four-report task, Brier reward, update-group size 16, additive stabilizer .03, and 16 separate reference reports for independent common scaling. The pretraining diagnostics are in Appendix F.3.3.

Learning rates are selected separately by arm on paired seeds 20261111–20261112 from $\{2.5\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$. Mean best validation restricted-score regret selects 10^{-4} for outcome-wise, 2.5×10^{-5} for no division, and 5×10^{-5} for independent common scaling. Evaluation uses 12 disjoint paired seeds 20261121–20261132, 320 minibatch attempts, and validation checkpoints at attempts $\{0, 10, 20, 40, 80, 120, 160, 240, 320\}$.

Each nonzero global LoRA gradient is divided by its norm before AdamW, and clipping is inactive. Every minibatch calls the optimizer, including those with a zero raw gradient. This fixes sampling exposure while the number of nonzero gradients can differ between arms.

Table 9: Validation-selected endpoints with 320 minibatches. Change is relative to the common warm checkpoint; KL and zero-gradient fractions are finite-time diagnostics.

Arm	Regret	Change	Raw-Bayes mass	KL	Zero fraction
Outcome-wise	.08991	−.01349	.445	1.081	.417
No division	.06881	−.03459	.748	.280	.0357
Independent common	.06973	−.03367	.722	.351	.0326

Outcome-wise selected regret exceeds no division by .02110 and independent common scaling by .02018. All 12 paired differences are positive in both comparisons (Figure 3e). Its mass on the raw Bayes reports is lower by .3029 and .2763, respectively. Each arm improves mean selected regret over initialization.

Raw gradient norms differ between estimators, with arm means 10.076, 1.963, and 19.083. The unit-norm transform removes those magnitudes before AdamW; the raw fields remain estimator-dependent. Outcome-wise training has more zero-gradient batches and larger selected KL. The regret gap remains with equal minibatch exposure and unit-norm gradients entering AdamW. The next study fixes the number of nonzero updates, allowing minibatch counts to differ.

F.5.3 A FIXED NUMBER OF NONZERO UPDATES

The second control reuses the first control's model, data, warm checkpoint, estimator settings, and selected learning rates. Evaluation uses 12 disjoint paired seeds, 20261221–20261232, without retuning. The endpoint of 120 nonzero updates was fixed before evaluation. It is the largest prespecified validation checkpoint in the study with a fixed minibatch budget reached by every outcome-wise evaluation seed; the next checkpoint, 160, was not reached by all.

Each attempted minibatch draws one $Y \sim \mathrm{Bernoulli}(p_h)$ per prompt and shares it across that prompt's report group. A raw global gradient with norm at most 10^{-12} is recorded but does not apply an optimizer step or advance AdamW's moments. Every applied gradient is rescaled to norm one, and each run stops after exactly 120 applied updates, with a hard cap of 480 attempted minibatches. Validation checkpoints are indexed by applied updates at $\{0, 10, 20, 40, 80, 120\}$. All 36 evaluation runs reach the endpoint, with clipping inactive.

Table 10: Endpoints with 120 nonzero updates. Selected columns use validation-selected checkpoints; the exact column uses 120 applied updates; attempts/update records completion cost.

	Validation-selected checkpoint			Exact 120	Completion cost
Arm	Test regret	Raw-Bayes mass	KL	Test regret	Attempts/update
Outcome-wise	.08731	.491	.799	.11752	1.872
No division	.07101	.752	.226	.07166	1.004
Indep. common	.07093	.739	.285	.07361	1.021

At exactly 120 updates, outcome-wise test regret exceeds no division by .04586 and independent common scaling by .04391. Both contrasts are positive in all 12 seeds. At that horizon, outcome-wise regret is .01412 above initialization, while both controls remain below it. At the validation-selected checkpoints, the corresponding contrasts are .01629 and .01638, each positive in 11 of 12 seeds. All three arms improve selected regret over the common initial value .10340. Figure 3(e,f) shows the paired results and the validation paths.

Selected raw-Bayes mass differs by $-.2610$ against no division and $-.2484$ against independent common scaling. These masses and the table's KL values are measured at the validation-selected checkpoints. Outcome-wise runs require a mean 224.58 attempted minibatches to obtain 120 updates, versus 120.50 and 122.50 for the two controls. Matching nonzero updates therefore requires more sampled minibatches for outcome-wise scaling. The preceding study fixes minibatch exposure, and the gap persists under both budget definitions.

F.6 Learning across proper scores

On the four-report task, the outcome-wise update group has size two and $\tau = 0$, where Equation (13) gives the exact all-score rank field. On the retained rows, Brier and log have raw restricted Bayes report .25 but majority report .05 at lower truths; upper-truth rows analogously separate .75 from .95. These Bayes reports are unique for all three scores on the 210 validation, 430 primary-test, and 163 boundary-test rows. Among 875 training rows, cubic favors 0.05 for one row with truth 0.15217. All regrets and Bayes masses use each row's score-specific maximizer set.

All three arms draw two update and 16 separate reference reports from paired streams. Outcome-wise scaling and no division use the update pair and ignore references; independent common scaling uses the reference group. Group size, report budget, prompt order, and neural forward/backward work are matched. Common scaling preserves the raw per-prompt mean-field direction (Appendix D.2.1).

All arms use AdamW with gradient clipping at norm one. Learning rates are tuned separately by arm on two seeds over the grid $\{2.5\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$. Selection by best mean validation restricted-score regret chooses, respectively, 2.5×10^{-5}, 5×10^{-5}, and 10^{-4} for outcome-wise, no division, and independent common scaling. Tuning and evaluation seeds are disjoint. Brier uses 12 evaluation seeds and 320 minibatch attempts; log and asymmetric cubic use six seeds each and 240 attempts. Each run selects its checkpoint by validation restricted-score regret.

Table 11: Brier learning results after arm-specific tuning and validation selection. Changes are relative to the common warm regret .10340.

Arm	Selected regret	Change	Mean step	Final raw/majority mass
Outcome-wise	.09463	$-.00876$	20.0	.026/.755
No division	.06941	$-.03398$	210.0	.764/.008
Independent common	.07304	$-.03036$	186.7	.308/.479

Selected Brier regret is higher with outcome-wise scaling than without division by .02522. Against independent common scaling, the gap is .02160. Both comparisons are positive in all 12 paired seeds. Both controls improve selected regret over initialization. At the final checkpoint, outcome-wise scaling places .7473 more mass on the majority extreme than no division.

After one frozen positive global factor aligns each score's largest pairwise gap with Brier, log and asymmetric cubic reuse the Brier-selected arm-specific rates. Against no division, outcome-wise excess selected regret is .02200 for log and .02115 for cubic. Against independent common scaling, the gaps are .01650 for log and .02106 for cubic. All four gaps are positive in all six paired seeds.

F.7 Additional ten-report studies

Training with ten report tokens. The ten-report studies compare three arms starting from the shared warm adapter. They use report tokens $\{0.05, 0.15, \ldots, 0.95\}$, negative Brier reward, prompt batch size eight, group size 16, 160 minibatch attempts, gradient clip one, zero weight decay, and no extra reference samples. AdamW uses learning rate 2×10^{-4} except in the learning-rate sweep. A run therefore uses 20,480 sampled and gradient-bearing reports. The outcome-wise arm uses additive-SD

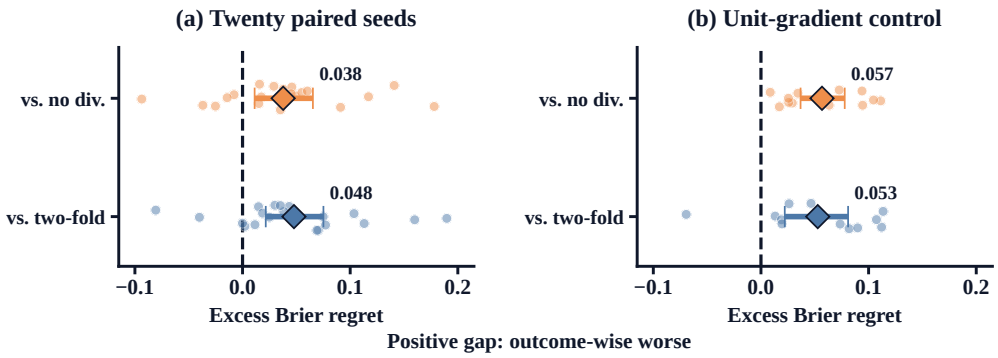


Figure 5: Fixed-horizon mechanism controls. The plotted excess Brier regret is the difference in mean-report squared error between arms. (a) The 20-seed study shows a fixed-horizon separation. (b) That separation remains when every nonzero gradient entering AdamW has unit norm. Dots are paired seeds; diamonds and bars are means and bootstrap 95% intervals.

stabilizer 0.03; the no-division arm centers only; the two-fold arm divides each size-eight update fold by a pooled RMS scale computed from the other fold. Groups are fresh and used once, with no importance ratio, PPO clip, KL/entropy term, or minibatch reuse. Across the 20-seed robustness runs, each arm takes about 167 seconds on average and uses 6.71 GiB of peak memory.

F.7.1 Endpoint differences across twenty seeds

All 20 seeds use the same configuration and appear in Table 12.

Table 12: Twenty paired seeds at learning rate 2×10^{-4}. Contrasts are outcome-wise minus comparator.

Arm or contrast	MSE/difference	Positive seeds
Outcome-wise	0.11202	–
No division	0.07427	–
Two-fold common	0.06424	–
Outcome-wise − no division	0.03775	15/20
Outcome-wise − two-fold	0.04778	17/20

Figure 5a shows the paired differences and their uncertainty. The positive mean differences persist after removing any single seed.

F.7.2 Mean-report error at comparable KL divergence

Each learning-rate cell in Table 13 contains six paired seeds. The common-support comparison is piecewise linear in the arm means and does not extrapolate beyond either arm's observed KL range.

On common KL support, outcome-wise minus no division has minimum/mean excess MSE $0.02077/0.03033$ over $[1.6853, 1.9798]$; outcome-wise minus two-fold has $0.02026/0.02677$ over $[1.4344, 2.0578]$.

F.7.3 Fixed-horizon and selected endpoints with unit gradients

With the same ten-report setup, 12 paired seeds use learning rate 2×10^{-4} and validation at steps $\{0, 20, 40, 80, 120, 160\}$. Each nonzero gradient is normalized to unit norm before AdamW (Loshchilov & Hutter, 2019). Table 14 compares the 160-step and validation-selected endpoints.

At the fixed horizon, outcome-wise MSE exceeds no division by 0.05671 and two-fold common scaling by 0.05275. Validation selection returns the warm checkpoint (MSE 0.048027) in 31 of 36

Table 13: Raw arm means across the frozen learning-rate sweep.

Arm	LR	MSE	KL	Clip rate	Pre-clip norm
Outcome-wise	10^{-4}	.067312	1.1981	.9865	14.2898
Outcome-wise	2×10^{-4}	.112688	2.0207	.7687	12.6400
Outcome-wise	4×10^{-4}	.126886	2.1041	.6365	7.3270
No division	10^{-4}	.055274	1.6853	.3458	.8831
No division	2×10^{-4}	.066562	1.9798	.2479	.7017
No division	4×10^{-4}	.088178	1.9420	.1323	.8807
Two-fold	10^{-4}	.050143	1.4344	.9990	14.2523
Two-fold	2×10^{-4}	.068329	1.5770	.8594	13.0387
Two-fold	4×10^{-4}	.081238	2.0578	.5833	8.6950

Table 14: Arm means and validation-selected steps with unit gradients.

Arm or contrast	Step-160 MSE	Selected MSE	Selected step counts
Outcome-wise	0.121769	0.048027	0: 12
No division	0.065056	0.047699	0: 9; 120: 2; 160: 1
Two-fold common	0.069020	0.048213	0: 10; 20: 1; 80: 1

runs, so the selected endpoints mostly coincide with initialization. The fractions of minibatches with a zero raw gradient are 20.16%, 2.08%, and 6.67% for outcome-wise, no-division, and two-fold scaling, respectively.