# BEYOND VERIFIER RELIABILITY: INFORMATION FOR POLICY UPDATES IN GRPO

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

## ABSTRACT

Reinforcement learning with verifiable rewards commonly uses binary correctness feedback, yet its accuracy on current responses need not identify which policy update most improves true correctness. Accuracy aggregates errors, whereas update value depends on their alignment with changes in response probability. To isolate this missing information, we fix the initial policy, verifier, and complete GRPO menu, then construct binary utility worlds with indistinguishable finite-precision reports but different preferred updates at equal KL. The separation persists under specified perturbations and finite-rollout construction. We also characterize the exact linear utility moments sufficient for finite candidate comparisons and give conditions for reusing an initial correctness contrast, with separate estimation and redistribution errors. Two controlled Qwen3-4B experiments examine the consequences for feedback evaluation. The first holds group confusion matrices and centered reward MSE fixed to test whether these summaries determine update outcomes: the resulting policies differ by 8.53 percentage points in average utility. Only a four-position prefix-KL statistic is matched; a complete-history diagnostic finds unequal sequence KL. The second tests whether targeted execution information improves training at a fixed test count: four diagnostic tests outperform four random tests by 3.9 and 7.0 percentage points in two seeds. Together, the results identify information omitted by reliability summaries and conditions under which decision-relevant information can be retained and reused.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) uses answer checks or code execution to guide language-model updates (Shao et al., 2024, Guo et al., 2025, Yu et al., 2025). These checks commonly return pass/fail outcomes, motivating our focus on binary verifier feedback $V$ and trusted correctness $U$, apart from auxiliary format or length rewards. A verifier can be evaluated by counting its errors on current responses. Choosing an update, however, requires knowing whether the responses it promotes are correct. The information sufficient for the first assessment may therefore be insufficient for the second.

GRPO makes this distinction explicit. Its advantages compare responses within a prompt, while shared parameters couple their effects across prompts. Sampling and normalization thus change how probability is redistributed (Liu et al., 2025, Thrampoulidis et al., 2026, Yang et al., 2026, Ge et al., 2026). The same number of reward errors can interact differently with this redistribution, depending on which responses carry them. This motivates two linked problems: identifying the utility information needed to compare updates, and determining when information acquired under the initial policy remains useful for later comparisons.

We address identification, GRPO realizability, and reuse under a common information model. The separation fixes the initial policy, verifier, and complete candidate menu while varying unknown binary utility, unlike a reference-policy intervention (Lamparth et al., 2026). Its finite-precision reporting restriction is explicit: raw response–label records permit additional statistics. Our contributions are as follows.

First, we establish that the information gap is realizable under GRPO's update constraints. Indistinguishable reports can require different choices over a complete menu of actual updates at equal KL



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.

from the initial policy. The construction excludes a third configuration that is good in both utility worlds, and persists under specified perturbations and finite-rollout construction.

Second, we characterize what resolves this gap and what limits reuse. The sufficient-moment criterion identifies the utility information required by a fixed menu. In the paired GRPO family, one initial correctness contrast supports later comparisons, with separate errors for estimating the contrast and changing the response composition within its blocks. A matching-order obstruction shows why acquiring more precise old means cannot by itself remove the latter error.

Third, two controlled Qwen3-4B interventions examine the empirical consequences of evaluating feedback for updates. The first holds error counts and centered MSE fixed to test whether they determine finite-step utility; the outcomes differ after prefix-KL calibration, although sequence KL remains unequal. The second tests whether diagnostic execution information improves training beyond simply adding tests, and finds gains over an equal-count random-test baseline. Together they probe the limitations of aggregate evaluation and a practical source of additional feedback, without claiming equal-sequence-KL neural separation or unrestricted reuse.

## 2 Related Work

*GRPO and normalization.* GRPO compares responses to each prompt without a learned critic (Shao et al., 2024). Changes to normalization, clipping, and handling silent groups alter how rewards drive learning (Liu et al., 2025, Yu et al., 2025). Analyses of effective objectives, advantage shaping, finite-group scores, and interactions across prompts explain these effects (Mroueh, 2025, Mroueh et al., 2026, Thrampoulidis et al., 2026, Bay & Yearick, 2026, Yang et al., 2026, Ge et al., 2026). We build on this distinction between reward values and their optimization effects, but ask a different question: whether a reward's reliability report identifies the better configuration under a separate trusted utility. Our construction jointly constrains the actual GRPO candidates, their KL divergence, and the information available for selecting one.

*Noisy rewards and robustness.* Flip-rate estimation can debias group-relative learning under query-dependent Bernoulli corruption (El Mansouri et al., 2026). Empirical studies find both gains under misleading supervision and degradation from incorrect targets, with effects depending on the noise model and training budget (Shao et al., 2026, Chen et al., 2026, Zhu & Kang, 2026, Mitsuhashi et al., 2026). Our uncertainty is different: verifier and utility are fixed functions, and their relationship is only partially observed. This lets us isolate what an unchanged reliability report leaves unknown about a particular update, rather than attribute every performance difference to a change in aggregate noise rate.

*Audits and policy improvement.* Reward overoptimization distinguishes proxy gains from intended improvement (Gao et al., 2023); Correlated Proxies relates base-policy reward correlation to policy regularization (Laidlaw et al., 2025). RewardBench 2 tests whether offline scores predict best-of-$N$ and PPO performance (Malik et al., 2026). Most closely, Reward Bias Substitution constructs identical audit observables with different mitigation outcomes by varying the reference policy (Lamparth et al., 2026). We instead fix the initial policy, verifier, and candidate menu and vary unknown binary utility locations. The resulting limitation explicitly accounts for reporting precision, point-label access, and every candidate in the menu.

*Retained information and reuse.* Aggregate-label learning studies information in label proportions (Busa-Fekete et al., 2025); successor features and abstract reward processes preserve policy-relevant reward information through factorizations or abstractions (Barreto et al., 2017, Chaudhari et al., 2024). Our characterization specifies the span needed for candidate contrasts using elementary linear identification. The disclosure bound likewise uses standard testing tools (Lattimore & Szepesvári, 2020), with alternatives constrained to actual GRPO updates. Reuse then requires controlling changes inside the cells whose initial utility means were retained. This information requirement complements PPO and TRPO, which limit updates made with a supplied advantage estimate (Schulman et al., 2015, 2017).

*Code feedback.* Internal representations can support code-correctness ranking (Ribeiro et al., 2026). UTRL rewards valid tests that expose generated-code errors (Lee et al., 2026); CVeDRL combines correctness, coverage, and difficulty to train a verifier (Shi et al., 2026). Our representation diagnostics

and fixed-test-count training comparison examine whether such response discrimination supplies useful group-relative feedback, while accounting for the interpreter that provides test answers.

## 3 Policy Updates and Verifier Information

We evaluate an audit by the decision its retained information must support. Fix a prompt distribution, an initial policy $P_0$, and a nonempty finite menu of constructed candidate laws $\{P_c : c \in \mathcal{C}\}$ over $(X, Y)$. The verifier $V(X, Y) \in \{0, 1\}$ is available; trusted utility $U(X, Y) \in [0, 1]$ is fixed but unknown. The selector executes one candidate without further modification, with value and regret

$$J_U(P) = \mathbb{E}_P[U], \qquad \mathcal{R}_{\mathcal{C}}(\widehat{c}) = \max_{c \in \mathcal{C}} J_U(P_c) - J_U(P_{\widehat{c}}). \tag{1}$$

Base may be included. The general information criterion allows bounded utility; the GRPO construction and code experiments use binary utility. For code, success means passing the specified finite suite.

Even exact correctness rates need not determine this decision. Let $S = (X, V)$ identify verifier cells and, for $P_c, P_b \ll P_0$, put $f_{c,b} = (dP_c - dP_b)/dP_0$. This contrast records where $c$ places more probability than $b$. Conditioning its utility-weighted mean gives

$$J_U(P_c) - J_U(P_b) = \mathbb{E}_0[\mathbb{E}_0(U \mid S)\mathbb{E}_0(f_{c,b} \mid S)] + \mathbb{E}_0[\mathrm{Cov}_0(U, f_{c,b} \mid S)]. \tag{2}$$

Cell correctness rates determine the first term, but omit whether probability increases favor correct responses *within* a cell. The covariance is therefore the missing association, not further uncertainty about the same cell mean.

For example, let $P_0$ be uniform on four responses in one accepted cell. For $0 < \gamma < 1$, consider

$$f_{c,0} = (\gamma, \gamma, -\gamma, -\gamma), \qquad U_+ = (1, 1, 0, 0), \quad U_- = (0, 0, 1, 1).$$

The candidate shifts mass toward the first two responses. Both worlds have cell correctness $1/2$, yet $J_{U_+}(P_c) - J_{U_+}(P_0) = \gamma/2$ and $J_{U_-}(P_c) - J_{U_-}(P_0) = -\gamma/2$. The report loses which half is correct, precisely the information needed to decide whether to execute this update. This example isolates the information loss; realizing it under GRPO and equal-KL constraints requires the construction in Section 4.

To identify what must replace that summary, consider exact linear reports on a finite support $\Omega$ with $P_0(\omega) > 0$. Fix the candidates and reporting functions $h_1, \ldots, h_q$ before choosing $U \in \mathcal{U} := [0, 1]^{\Omega}$. Define the report, comparison vector, and optimal set by

$$M(U) = (\mathbb{E}_0[U h_j])_{j=1}^{q}, \quad D(U) = (J_U(P_c) - J_U(P_b))_{c,b \in \mathcal{C}}, \quad \mathcal{A}(U) = \arg\max_{c \in \mathcal{C}} J_U(P_c).$$

Write $\mathcal{H} = \mathrm{span}\{h_j\}_{j=1}^{q}$ and $\mathcal{F} = \mathrm{span}\{f_{c,b} : c, b \in \mathcal{C}\}$. The three conditions below respectively ask whether the report covers the contrasts, determines their values, and permits an optimal decision.

**Theorem 1** (Information sufficient for candidate comparisons)**.** *Under this finite-support setup, the following are equivalent:*

$$\begin{aligned} &\textit{(i)} \quad \mathcal{F} \subseteq \mathcal{H}; \\ &\textit{(ii)} \quad \forall U, U' \in \mathcal{U} : \quad M(U) = M(U') \implies D(U) = D(U'); \\ &\textit{(iii)} \quad \exists \chi : M(\mathcal{U}) \to \mathcal{C} \quad \forall U \in \mathcal{U} : \quad \chi(M(U)) \in \mathcal{A}(U). \end{aligned}$$

*Moreover, $\mathcal{F} \nsubseteq \mathcal{H}$ implies*

$$\exists U_+, U_- \in \mathcal{U} : \quad M(U_+) = M(U_-), \qquad \mathcal{A}(U_+) \cap \mathcal{A}(U_-) = \varnothing.$$

The proof is in Appendix D.7. A contrast inside the reported span can be reconstructed from the moments. If part lies outside, two utilities can agree on every report while making different candidates optimal. The theorem rules out even a common optimal third candidate, not merely recovery of one pair's value. For cell means, $\mathcal{H}$ consists of cellwise constant functions: sufficiency therefore requires that the contrasts have no unreported within-cell component. Appendix D.8 bounds the error when this is only approximate.

The criterion identifies information to retain, rather than a sample-efficient estimator: a basis of $\mathcal{F}$ minimizes the number of unrestricted exact moments, not their acquisition cost. Its necessity also uses the full cube $[0,1]^{\Omega}$. An exact real mean can encode a finite binary label vector, so a binary-utility limitation must specify reporting precision. Section 4 does so while also requiring the candidates to be actual GRPO updates.

There is a corresponding distinction when feedback constructs the update itself. Fix a group $g$ of $m_g$ responses, its centering matrix $C_g = I - \mathbf{1}\mathbf{1}^\top/m_g$, and score matrix $\Psi_g$. With detached, mean-centered advantages, supplied rewards $\widehat{\mathbf{r}}_g$ and trusted rewards $\mathbf{u}_g$ give unclipped surrogate gradients satisfying

$$\widehat{g}_g - g_{U,g} = \frac{1}{m_g}\Psi_g^\top C_g(\widehat{\mathbf{r}}_g - \mathbf{u}_g). \tag{3}$$

Centering removes a prompt-only prediction; the remaining errors are weighted by response scores. Thus identical error magnitudes need not imply identical updates. This fixed-batch identity, with any implemented length normalization included in $\Psi_g$, motivates the interventions in Section 6. It does not itself guarantee a finite-step utility difference. The next section establishes such a decision gap under explicit GRPO construction constraints.

## 4 When Reliability Does Not Determine the Better Update

Theorem 1 permits arbitrary candidate laws and bounded utilities. To make its information gap relevant to GRPO, we must realize it with binary correctness, actual mean-centered and standardized updates, and equal policy-change magnitude. We must also exclude a third configuration that works well in both utility worlds. The construction below meets these requirements within one fixed initial policy and verifier.

Candidates use expected or empirical on-policy group gradients with detached advantages, no clipping, and no gradient KL penalty. They access no trusted labels, and each nonbase candidate is calibrated to the same positive $\mathrm{KL}(P_0\|P_c)$. The selector knows the policy, verifier, features, candidates, and construction history. Its utility information consists of rounded initial cell correctness rates and queried labels; repeated queries return the same label. Raw response–label pairs are not released with the report. Additional disclosures are new acquisitions only if they require an oracle call.

The design separates the source of an update from the responses on which its utility is measured. Use three equally likely prompts and two shared parameters,

$$\pi_\theta(i \mid x) \propto \pi_0(i \mid x)e^{\theta^\top \phi_{x,i}}, \qquad \theta_0 = 0, \qquad v = (1,-1)/\sqrt{2}. \tag{4}$$

A probe prompt has accepted blocks $I = I_+ \cup I_-$, each with $R$ atoms. Write $w_i = P_0(i \mid I) > 0$, $\sum_I w_i = 1$, and $\sigma_i = \pm 1$ on $I_\pm$. Each accepted atom has feature $\phi_i = \sigma_i v + \xi_i$, $\|\xi_i\| \le \kappa$, and a rejected partner of equal initial probability and feature. Pairing keeps probe acceptance at $1/2$ for every $\theta$, while allowing redistribution between accepted blocks. Utility is binary on $I$ and zero elsewhere, so only false-positive verifier errors remain to be located.

The other prompts supply the gradient. Their differing pass rates cause mean centering and standardization to weight them differently. The explicit probabilities and bounded features in Appendix A yield mean-16 and standardized-16 gradients proportional to $(\lambda, 1)$ and $(1, \lambda)$, $\lambda > 1$. Their unit directions obey

$$v^\top d_M = a_*, \qquad v^\top d_S = -a_*, \qquad a_* > 1/20. \tag{5}$$

For $\kappa < a_*$, the mean update increases the conditional odds of $I_+$ and the standardized update decreases them. The probe has zero expected verifier gradient itself; the opposite movements arrive through shared parameters. Calibrating both to equal KL controls their size without deciding which movement improves utility.

Retain the complete nonbase menu $\mathcal{C}_{\mathrm{nb}} = \{(M,m),(S,m) : m \in \{4,8,16\}\}$, optionally adding base $0$. Under expected gradients, all three mean directions coincide, leaving four distinct calibrated laws but six configuration names. None is removed. To make their choice ambiguous, in world $z \in \{-1,1\}$ uniformly place exactly $R(1/2 + z\sigma\epsilon)$ positive labels in $I_\sigma$, with $0 < \epsilon \le 1/8$ and integer counts. Labels then stay fixed. At balanced, equal weights, both worlds have correctness $1/2$ among accepted responses, yet favor opposite block movements. At unequal weights, a report bin centered at $1/2$ with half-width $\tau$ can preserve this indistinguishability.

Two margins keep the construction informative away from symmetry: the report must stay in that bin, and the candidate gain must retain its sign. For $d_i = P_c(i) - P_b(i)$, put $A = \sum_I d_i$, $B = \sum_I \sigma_i d_i$, and $a = \sum_I \sigma_i w_i$. With concentration radii $r_\mu, r_\Gamma$ from Appendix B, sufficient conditions are

$$\tau > \epsilon|a| + r_\mu, \qquad \eta = \epsilon B - |A|/2 - r_\Gamma > 0. \tag{6}$$

They give a common report and opposite pair gains of magnitude at least $\eta$ with prior probability at least $1 - \alpha$ in each world. At the symmetric paired point, $a = A = r_\mu = r_\Gamma = 0$ and $B > 0$, so both margins are strict.

A pair gain alone would leave the possibility of a common good third choice. We eliminate it using simultaneous full-menu bounds $q_{z,c} \pm r_c$, where $q_{z,c} = A_c/2 + z\epsilon B_c$ compares $c$ with base. Define

$$L_z = \max_c (q_{z,c} - r_c), \qquad \mathcal{C}_z(t) = \{c : q_{z,c} + r_c \geq L_z - t\}. \tag{7}$$

Any choice with regret at most $t$ lies in this outer acceptable set. Disjoint sets therefore force a successful choice to identify the utility world. Let $\mathcal{G}_z$ contain exactly the fixed-count oracles with the central report $\mathsf{R}_\tau(U)$ and all gain bounds in Eq. (C.4). Assume $\alpha, \delta > 0$ and $\alpha + \delta < 1/4$.

**Theorem 2** (Separation for actual GRPO updates)**.** *There exist positive initial weights, bounded shared features, and $K_*, t > 0$ such that*

$$\mathrm{KL}(P_0 \| P_c) = K_* \quad (c \in \mathcal{C}_{\mathrm{nb}}), \qquad \mathcal{C}_+(t) \cap \mathcal{C}_-(t) = \varnothing,$$
$$\mathsf{R}_\tau(U) = \mathsf{R}_\tau(U') \quad (U, U' \in \mathcal{G}_+ \cup \mathcal{G}_-).$$

*For any selector restricted to the stated report-and-disclosure interface, with a deterministic cap of $n$ disclosures,*

$$\inf_{U \in \mathcal{G}_+ \cup \mathcal{G}_-} \Pr_U(\mathcal{R}_{\mathcal{C}}(\widehat{c}) \leq t) \geq 1 - \delta \quad \Longrightarrow \quad n \geq \min\left\{ \frac{R}{8}, \frac{\log(1/[4(\delta + \alpha)])}{32\epsilon^2} \right\}. \tag{8}$$

The proof is in Appendix C.2. The menu and initial observations are the same, but a useful redistribution depends on hidden correctness locations. Equal KL and access to the whole menu do not supply that information. A low-regret choice consequently distinguishes the worlds, which requires the additional labels in Eq. (8). This is the decision consequence of the missing association in Eq. (2).

The result does not rely solely on an exact symmetric table or expected gradients. Strict margins persist under small perturbations, including broken pairing. With bounded features, positive initial Fisher information, and an ideal menu margin, finite-rollout construction preserves separation with probability at least $1 - \rho$. Appendix C.4 gives explicit budgets and the corresponding joint construction/selection disclosure bound; Appendix C treats sampled reports.

Reporting resolution remains essential: at generic unequal weights, an exact real mean can encode every binary label, whereas more samples at fixed resolution may return the same bin. The limitation thus identifies what the stipulated summary discards, not an impossibility of learning useful information. A signed correctness contrast avoids the cancellation. We next study whether one initial estimate of that contrast can support multiple comparisons.

## 5 Directional Information and Its Reuse

The separation suggests a constructive response: retain which accepted block is more correct, rather than only their aggregate correctness. Reuse introduces a second question, however. An initial block mean need not describe its responses after the policy changes. We isolate these two issues in the paired family of Eq. (4), with fixed features, blocks, and arbitrary fixed binary utility on $I$, zero elsewhere. The fixed label counts used for the lower bound are no longer assumed.

Let $\mu_+, \mu_-$ denote initial within-block utility means. With $a = \sum_I \sigma_i w_i$, define the correctness contrast $b$, aggregate correctness $\mu$, and policy-dependent block allocation $s(\theta)$ by

$$b = \frac{\mu_+ - \mu_-}{2}, \qquad \mu = \frac{1 + a}{2}\mu_+ + \frac{1 - a}{2}\mu_-, \qquad s(\theta) = P_\theta(I_+ \mid I) - P_\theta(I_- \mid I). \tag{9}$$

The audit estimates $b$; a candidate specifies $s(\theta)$. To acquire $b$, choose a block uniformly and draw under its initial conditional weights. Then $Z = \sigma(U - 1/2)$ has mean $b$ and range $[-1/2, 1/2]$, so $n$

iid draws give

$$\Pr\Big(|\widehat{b} - b| \le e_n\Big) \ge 1 - \delta, \qquad e_n = \sqrt{\frac{\log(2/\delta)}{2n}}, \quad 0 < \delta < 1. \tag{10}$$

This proposal estimates the desired contrast even at unequal response weights; uniform sampling by identity would generally estimate a different one. Repeated labels may be cached, but $n$ counts draws.

We control reuse error by bounding how much correctness can change *within* each block. Fix $\|\theta\| \le H$ and $\|\xi_i\| \le \kappa$, keeping the oracle, prompt masses, features, and pairing fixed. Conditional block probabilities change only through $e^{\theta^\top \xi_i}$, whose log tilt has range at most $2H\kappa$. The normalized-tilt bound in Appendix D yields

$$J_U(P_\theta) = \frac{\mu}{6} + \frac{s(\theta) - a}{6} b + e(\theta), \qquad |e(\theta)| \le \frac{1}{6} \tanh(H\kappa/2). \tag{11}$$

The offset $\mu/6 - ab/6$ is common to all candidates; $b\, s(\theta)/6$ captures their utility difference from block redistribution. The residual $e(\theta)$ is what an old block mean cannot resolve. Substituting $\widehat{b}$ therefore gives a selection rule with two distinct error sources.

**Theorem 3** (Uniform approximate reuse)**.** *In the paired setting, let $\Theta_H = \{\theta : \|\theta\| \le H\}$ and $E_n = \{|\widehat{b} - b| \le e_n\}$, with $\Pr(E_n) \ge 1 - \delta$. On $E_n$, simultaneously for all $\varnothing \ne \mathcal{C} \subseteq \Theta_H$ with $|\mathcal{C}| < \infty$ and all $\widehat{\theta} \in \arg\max_{\theta \in \mathcal{C}} \widehat{b}\, s(\theta)$,*

$$\mathcal{R}_{\mathcal{C}}(\widehat{\theta}) \le \frac{e_n}{6} \left[\max_{\theta \in \mathcal{C}} s(\theta) - \min_{\theta \in \mathcal{C}} s(\theta)\right] + \frac{1}{3} \tanh(H\kappa/2). \tag{12}$$

The proof is in Appendix D.1. The audit's sampling error is weighted by how differently the candidates allocate mass between blocks. The second term bounds the unobserved change inside them. The same estimation event covers every finite menu in the fixed region, including menus chosen after the audit: this is the precise sense in which one contrast can be reused. More labels reduce the first term, but not the second. Pairwise cancellation and a nonzero contrast can sharpen the guarantees; Appendices D.4–D.6 give those refinements and their budgets.

To use this result at a phase boundary, fix the allowed region and a regret tolerance $\rho$ before acquiring labels. The drift allowance is $D_H = \tanh(H\kappa/2)/3$. If $\rho > D_H$, the remaining allowance can be assigned to estimating $b$; Appendix D.6 gives a sufficient draw budget. If $\rho \le D_H$, the bound cannot certify the tolerance regardless of label count; this is not a task-specific impossibility. Tighter drift analysis or new within-block information may still resolve the decision. The guarantee covers candidate comparisons inside the fixed family, not arbitrary optimization throughout a phase. Directly auditing frozen candidates remains an alternative when such a reuse region cannot be established.

To test whether the drift term reflects missing information rather than merely a loose bound, give the selector the old means exactly. Start with balanced blocks and split each into two equal subblocks indexed by $\zeta \in \{-1, 1\}$:

$$\phi_{\sigma,\zeta} = (\sigma + \kappa\zeta)v, \qquad U_z(\sigma, \zeta) = \frac{1 + z\zeta}{2}, \quad z \in \{-1, 1\}. \tag{13}$$

All four initial masses are equal, so both worlds have $\mu_+ = \mu_- = 1/2$ and $b = 0$. Their correct subblocks differ. Pairing leaves the GRPO directions unchanged, and symmetry gives a common radius $r$ at equal KL; their utility difference now reverses with $z$.

**Proposition 1** (Error from retaining only the old means)**.** *Let $\mathcal{C}_2 = \{rd_M, rd_S\}$ and let $\mathcal{I}_0$ denote the common information obtained without utility labels. For the observations*

$$O_z = (\mu_+(U_z), \mu_-(U_z), \mathcal{I}_0) = (1/2, 1/2, \mathcal{I}_0), \qquad z \in \{-1, 1\},$$

*the minimax regret over possibly randomized selectors $\chi(O_z) \in \mathcal{C}_2$ satisfies*

$$\inf_{\chi} \max_{z \in \{-1,1\}} \mathbb{E}_z[\mathcal{R}_{\mathcal{C}_2}(\chi(O_z))] \ge \frac{1}{12} \tanh(\kappa r a_*). \tag{14}$$

The proof is in Appendix D.2. Each old block is half correct in either world. The update nevertheless favors different halves, and the observed means cannot identify which half is correct. For small $\kappa r$ and fixed $a_* > 0$, the unavoidable regret is $\Omega(\kappa r)$, matching the order of Theorem 3's drift term at $H = r$. Improving the old estimates cannot resolve this ambiguity; a new statistic aligned with $\zeta$ can.

Together, the bounds explain why controlling within-cell change permits reuse, and why precise old means alone cannot guarantee it. In particular, $P_\theta(\cdot \mid S) = P_0(\cdot \mid S)$ makes old cell means exact for later comparisons. Cell-level reweighting or routing among frozen conditional generators can enforce this sufficient condition, with prompt conditioning when needed. Low adapter rank or sparse-prefix KL alone does not establish it for general neural updates. Appendix D.10 gives the multi-cell extension. The empirical study therefore tests the value of feedback information without presuming a neural reuse guarantee.

## 6 Empirical Study: Information and Update Outcomes

The theory identifies an information requirement for comparing updates; the experiments examine its consequences for feedback evaluation. The finite model first isolates what is lost by retaining a coarse summary of the same labels. Two neural interventions then separate summary sufficiency from feedback acquisition: one fixes error counts and changes their response locations, while the other tests whether targeted execution information improves training at a fixed test count. Neither neural study assumes the conditional stability needed for Theorem 3.

### 6.1 Which summary preserves the decision?

The construction in Section 4 suggests a direct way to resolve its ambiguity: retain the signed correctness contrast rather than only aggregate correctness. To isolate this information choice from label quantity, two selectors receive summaries of the same 128 iid raw draws. One retains the correctness mean; the other retains $\sigma(U - 1/2)$. We vary the initial block imbalance $a$ at fixed family separation $\epsilon = 1/16$ and measure wrong choices over 20,000 repetitions with shared samples and random tie breaking.

Table 1: Wrong-choice rates (%) using two summaries of the same 128 raw draws. The audit samples are shared within each repetition.

| Initial block imbalance $a$ | 0.02 | 0.10 | 0.50 |
|---|---|---|---|
| Coarse correctness mean | 49.09 | 44.19 | 24.02 |
| Directional mean | 7.96 | 7.78 | 7.90 |

Table 1 shows the cancellation predicted by the construction: the coarse signal is $\epsilon a$, whereas the signed signal remains $\epsilon$. The directional summary consequently retains its decision accuracy as the blocks approach balance. Increasing coarse draws to $128/a^2$ brings their error to approximately $8\%$, illustrating how retaining an informative contrast can avoid this loss of signal. A direct method retaining the raw response–label pairs can compute the same contrast at the same acquisition cost; this is a comparison of summaries, not an advantage over that method.

Appendix E checks that the underlying menu separation survives specified unequal weights and actual rollout construction, and reports settings where its sufficient certificate fails. Those checks establish the finite model's scope. The neural studies below instead examine whether feedback information affects actual generated-program utility.

### 6.2 Do matched summaries determine update outcomes?

Equation (3) predicts that a reward error matters through its association with response scores, not only its magnitude. Theorem 2 realizes the analogous information gap for fixed GRPO candidates and varying hidden utility. Here we test a related neural question with utility held fixed: do feedback assignments with the same conventional error summaries produce the same outcome when passed through a prescribed update procedure? Actual generated-program utility is the endpoint, since a constructed gradient difference alone would not answer this question.

Using Qwen3-4B in a controlled record-transformation domain (Yang et al., 2025), we fix three batches of 64 groups, each with eight responses. In every group, $r^+$ and $r^-$ match TP/FP/TN/FN, reward mean, variance, and centered reward MSE exactly, while placing errors on different response identities. The assignments maximize or minimize projection onto a direction from a separate calibration batch; their names do not assert a utility ordering. Complete labels construct the assignments and a full-$U$ reference direction. This controls error quantity and retains a trusted-feedback comparator, but uses full information rather than a low-label audit.

To limit differences in update scale, all nine fixed directions are calibrated to the same four-position prefix-KL tolerance, $[0.00095, 0.00105]$ nat/token. Each frozen candidate then generates four stochastic responses on each of 128 shared evaluation tasks, without test-based reranking. The primary contrast averages $J_U(P_{b,+}) - J_U(P_{b,-})$ equally over the three batches; uncertainty uses paired task resampling. The complete protocol, earlier common-norm calibration stop, and secondary endpoints are in Appendix H.1.

Table 2: Utility after the matched-feedback intervention. Counts are successes among 512 stochastic responses per candidate; base is 317/512. Intervals are approximate paired-task bootstrap 95% intervals in pp. The equal-batch signed contrast is primary; individual intervals are not multiplicity-corrected.

| Batch | $r^+$ | $r^-$ | Full $U$ | $r^+ - r^-$ (pp) | Interval (pp) |
|---|---|---|---|---|---|
| 1 | 355 | 311 | 385 | 8.5938 | $[3.5156, 13.8672]$ |
| 2 | 342 | 312 | 359 | 5.8594 | $[1.9531, 9.9609]$ |
| 3 | 368 | 311 | 341 | 11.1328 | $[5.8594, 16.6016]$ |
| Equal-batch mean | | | | 8.5286 | $[4.9479, 12.3047]$ |

The equal-batch gap is 8.53 percentage points (pp), with interval $[4.95, 12.30]$ pp and positive contrasts in every batch (Table 2). Equal empirical error summaries therefore do not determine the observed outcome under this procedure. The result establishes neither harm from $r^-$, whose differences from base have intervals containing zero, nor superiority to full-$U$ feedback.

The extent of policy change remains coupled to error placement. Summing conditional KL over all positions of 16 frozen base trajectories gives higher sequence-KL estimates for $r^+$ in every batch despite the matched prefix statistic. These histories are independent of utility evaluation but overlap calibration. Thus the experiment does not reproduce the theorem's equal-KL comparison. We next examine the complementary practical problem of obtaining useful additional feedback.

### 6.3 Does diagnostic execution information improve training?

A limitation of reliability summaries does not itself identify a useful feedback source. We therefore compare ways to acquire execution information in the same controlled domain. All conditions use public tests; two add either four random or four diagnostic tests, and a fourth uses the complete-v4 suite. The random condition controls for simply adding more checks. Diagnostic inputs target empty cases, duplicate-key/filter conflicts, ties, and numerical boundaries using specification-based rules frozen before training. An interpreter supplies their expected outputs, so domain knowledge is an explicit input to the method.

Each reward condition trains the same Qwen3-4B base with rank-8 query/value LoRA (Hu et al., 2022), using group-mean-centered rewards for 64 steps and 1,024 fresh responses. Two paired seeds share initialization and prompt schedules within each run. The common evaluation set contains 64 in-domain and 64 compositional specifications; the latter span structures absent from training. The endpoint is greedy program success on the complete suite, without execution-based reranking. Appendix G specifies the optimizer, sampling, finite evaluator, and relatedness among specifications.

Diagnostic feedback outperforms the same number of random tests by 3.9 and 7.0 pp (Table 3). Test content therefore matters beyond test count in this domain. The complete-reward arm remains stronger, so the result does not establish that four tests replace full verification. The positive overall contrasts also survive removing any one structure signature, although the two seeds share the evaluation set and do not constitute independent-task confirmations.

Table 3: Qwen3-4B greedy programs passing the complete-v4 suite. Each evaluation subset has 64 specifications, with the same base output cache in both runs.

| | Seed 944190 | | | Seed 944192 | | |
|---|---|---|---|---|---|---|
| Training reward | In-domain | Comp. | Total | In-domain | Comp. | Total |
| Initial model | 35 | 38 | 73 | 35 | 38 | 73 |
| Public | 38 | 37 | 75 | 41 | 37 | 78 |
| Public + random four | 42 | 38 | 80 | 41 | 37 | 78 |
| Public + diagnostic four | 43 | 42 | 85 | 45 | 42 | 87 |
| Complete-v4 | 40 | 47 | 87 | 47 | 46 | 93 |

To connect this endpoint to the update mechanism, we inspect groups where public rewards are constant but trusted labels differ. On the diagnostic trajectories, the added tests restore reward variation in most such groups in both seeds (Appendix G). They expose response distinctions that public centering would remove in Eq. (3). The earlier prediction studies in Appendix F illustrate why this check is needed: good cross-task prediction alone did not ensure useful within-group feedback. The trajectory observations are consistent with the update-information explanation, but differing trajectories and improved label accuracy prevent isolating it as the sole cause of the gains.

The benefit comes from acquiring fresh execution feedback; only the test assets are reused. The cost records show no consistent trainer-time advantage over complete verification. This distinguishes improved feedback acquisition from the reuse of old moments in Section 5; Appendix H.2 examines the further obstacle of natural-code transfer.

## 7 Discussion

Feedback evaluation should reflect the updates it will guide. A reliability report counts errors; a candidate comparison additionally depends on where probability moves. Theorems 1 and 2 make this distinction precise and realizable under GRPO. The neural studies examine its empirical relevance through matched summaries and equal-count test suites. Aggregate accuracy remains informative, but does not specify the whole update decision.

For infrequent audits, the same perspective makes stability a second requirement. A moment aligned with one comparison may lose relevance as response composition changes. Theorem 3 controls that drift within a fixed region, while Proposition 1 shows why more precise old means cannot remove it. A reusable audit must therefore identify both a small set of useful contrasts and a region where they remain valid; low dimension alone establishes neither stability nor a cost advantage over direct use of the same records.

The experiments leave these requirements open. Matched-feedback v2 does not control sequence KL, and diagnostic tests acquire fresh feedback using domain knowledge. The natural-code checks further show that apparent gains can disappear once public information and evaluator validity are addressed. A practical extension must establish both an information increment beyond a strong baseline and its validity under the intended policy changes. These are the missing steps from the present controlled results to a useful reuse method.

## 8 Conclusion

Choosing a useful GRPO update requires knowing how correctness aligns with probability redistribution, information that a reliability summary may omit. We characterize sufficient moments, realize the omission among actual equal-KL updates, and bound when an initial contrast remains informative for later comparisons. The neural studies show why this distinction matters for feedback evaluation, while retaining the limits of prefix-KL control and domain-specific execution tests. Together, the results motivate evaluating an audit by the decisions it supports and the policy changes over which it remains valid, rather than by its predictive accuracy or dimension alone.

## AI USE STATEMENT

We used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools for provide critical ingredients for proving mathematical claims, generate synthetic data sets, propose or refine hypotheses, assist with translation. We have reviewed all AI-assisted work, including checking ideas, checking mathematical claims, testing code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/350db081a661525235354dd3e19b8c05-Abstract.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/2607.00152v1.

Robert Istvan Busa-Fekete, Travis Dick, Claudio Gentile, Haim Kaplan, Tomer Koren, and Uri Stemmer. Nearly optimal sample complexity for learning with label proportions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 5943–5976. PMLR, 2025. URL https://proceedings.mlr.press/v267/busa-fekete25a.html.

Shreyas Chaudhari, Ameet Deshpande, Bruno Castro da Silva, and Philip S. Thomas. Abstract reward processes: Leveraging state abstraction for consistent off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 17069–17105, 2024. doi: 10.52202/079017-0543. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1e6dcc16ffa7ced2228d1f2fdc8b5adf-Abstract-Conference.html.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. Exploration vs exploitation: Rethinking RLVR through clipping, entropy, and spurious reward. In *International Conference on Learning Representations*, pp. 134506–134540, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d99e8e80a6c41e148db686918dd7eab3-Abstract-Conference.html.

Samuel N. Cohen and Eliana Fausti. Hyperbolic contractivity and the hilbert metric on probability measures, 2024. URL https://arxiv.org/abs/2309.02413v2. arXiv:2309.02413v2.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected GRPO: From noisy rewards to unbiased gradients. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61804.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

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/2601.23135v1.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

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.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

Max Lamparth, Daniel Fein, Andreas Haupt, Marcel Hussing, and Mykel J. Kochenderfer. Reward bias substitution: Single-axis bias mitigations redirect optimization pressure. arXiv preprint arXiv:2605.27996, 2026. URL https://arxiv.org/abs/2605.27996v2.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401. URL https://doi.org/10.1017/9781108571401.

Dongjun Lee, Changho Hwang, and Kimin Lee. Learning to generate unit test via adversarial reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VqjNYF9nbP.

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

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572, 2023. doi: 10.52202/075280-0943. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

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.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fb0G86Dewb.

Ryo Mitsuhashi, Patrick Chen, Isabelle Tseng, Jasin Cekinmez, and Addison J. Wu. Quantifying empirical compute-supervision tradeoffs in RLVR. arXiv preprint arXiv:2605.25252, 2026. URL https://arxiv.org/abs/2605.25252.

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/2503.06639v4.

Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, Jiri Navratil, Jarret Ross, and Jesus Rios. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XXfOf22o3K.

Francisco Ribeiro, Claudio Spiess, Premkumar Devanbu, and Sarah Nadi. On LLMs' internal representation of code correctness. In *Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering*, pp. 1160–1172. ACM, 2026. doi: 10.1145/3744916.3787846. URL https://doi.org/10.1145/3744916.3787846.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

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.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61082.

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/2402.03300v3.

Ji Shi, Peiming Guo, Meishan Zhang, Miao Zhang, Xuebo Liu, Min Zhang, and Weili Guan. CVeDRL: An efficient code verifier via difficulty-aware reinforcement learning. arXiv preprint arXiv:2601.22803, 2026. URL https://arxiv.org/abs/2601.22803.

Christos Thrampoulidis, Sadegh Mahdavi, and Wenlong Deng. Advantage shaping as surrogate reward maximization: Unifying Pass@K policy gradients. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=R1RhBFUk8t.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

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/2601.08521v2.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, et al. 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.

Yuxuan Zhu and Daniel Kang. Noisy data is destructive to reinforcement learning with verifiable rewards. arXiv preprint arXiv:2603.16140, 2026. URL https://arxiv.org/abs/2603.16140.

# A FULL SETTING AND EXPLICIT GRPO CONSTRUCTION

This appendix constructs the candidates required by Theorem 2. The argument has four parts. We first make the probe's verifier rate invariant while allowing its response probabilities to change. We then compute the expected group gradients, choose two other prompts that give opposite probe directions, and calibrate the resulting policies to the same KL. Appendix B will place unknown utility labels on these already constructed responses. Thus no trusted label is used to manufacture a candidate.

## A.1 POLICY, PAIRING, AND ACTUAL CANDIDATES

The three prompts $x \in \{0, 1, 2\}$ each have probability $1/3$. On the finite response support of prompt $x$, write the policy explicitly as

$$\pi_\theta(i \mid x) = \frac{\pi_0(i \mid x) e^{\theta^\top \phi_{x,i}}}{Z_x(\theta)}, \qquad Z_x(\theta) = \sum_j \pi_0(j \mid x) e^{\theta^\top \phi_{x,j}}. \tag{A.1}$$

All initial probabilities are positive. In particular, differentiation of this finite sum is legitimate, and every finite-parameter policy has the same support. The initial parameter is $\theta = 0$, for which $Z_x(0) = 1$. Joint response probabilities include the prompt mass, so $P_\theta(x, i) = \pi_\theta(i \mid x)/3$.

Prompt 0 is the probe. Its accepted set is $I = I_+ \cup I_-$, with $R$ atoms in each block. Set $\sigma_i = 1$ on $I_+$ and $\sigma_i = -1$ on $I_-$. Positive weights satisfy $\sum_{i \in I} w_i = 1$. Accepted atom $i$ has initial conditional-on-prompt probability $w_i/2$ and feature

$$\phi_i = \sigma_i v + \xi_i, \qquad v = (1, -1)/\sqrt{2}, \qquad \|\xi_i\| \le \kappa. \tag{A.2}$$

Its rejected partner has exactly the same probability and feature. Each pair therefore receives the same exponential multiplier in Eq. (A.1). The accepted and rejected unnormalized masses are equal at every parameter, which proves

$$P_\theta(V = 1 \mid X = 0) = \frac{1}{2}, \qquad P_\theta(X = 0, V = 1) = \frac{1}{6}. \tag{A.3}$$

Pairing fixes only this total mass. It does not fix the relative masses of $I_+$ and $I_-$, or the conditional distribution inside either block. This distinction is what permits another prompt's parameter update to change probe utility without changing probe verifier success.

Trusted utility is an arbitrary fixed binary function on $I$ and is known to be zero on all other responses. Hence $U = 1$ implies $V = 1$, and all differences in trusted utility come from the accepted probe. In the separation construction, the weights, features, verifier, and candidates are fixed before choosing a utility oracle.

Given a nonzero expected or empirical gradient $g_c$, normalize its direction to $d_c = g_c/\|g_c\|$ and choose a positive radius $r_c$ satisfying $\mathrm{KL}(P_0 \| P_{r_c d_c}) = k_0$, where $k_0 > 0$ is common to the menu. The $r_c$ in this paragraph is a calibration radius, not the concentration radius denoted $r_c$ in Appendix C. We prove existence and uniqueness below. When discussing a finite-rollout history, we condition on its label-free realization and use the actual resulting candidates. Expected and empirical gradients are not identified with one another. Appendix C.4 also supplies a probability guarantee before the history is drawn.

## A.2 FINITE-GROUP GRADIENT IDENTITIES

Fix a prompt and a parameter, suppress their indices, and draw $m$ independent responses. Write $V_i \in \{0, 1\}$ for their verifier rewards, $p = \mathbb{E}V_i \in (0, 1)$, and $\ell_i = \nabla_\theta \log \pi_\theta(Y_i \mid x)$ for their score vectors. Normalization and differentiation give

$$\mathbb{E}\ell_i = 0, \qquad \mathbb{E}[V_i \ell_i] = \nabla p. \tag{A.4}$$

The gradient considered here is the average $m^{-1} \sum_i A_i \ell_i$, with the computed advantages detached during differentiation. It has no clipping or KL penalty.

For mean centering, $A_i = V_i - \overline{V}$ and $\overline{V} = m^{-1}\sum_j V_j$. Independence makes $\mathbb{E}[V_j \ell_i] = 0$ whenever $j \neq i$. Only the self term remains in the baseline contribution, so

$$\mathbb{E}\left[\frac{1}{m}\sum_i (V_i - \overline{V})\ell_i\right] = \nabla p - \frac{1}{m^2}\sum_{i,j}\mathbb{E}[V_j \ell_i] \tag{A.5}$$

$$= \left(1 - \frac{1}{m}\right)\nabla p. \tag{A.6}$$

This calculation also explains why changing the mean-centered group size only changes a positive scalar in the expected gradient.

For standardization, use the empirical standard deviation with denominator $m$. Let $K = \sum_i V_i$. If $0 < K = k < m$, that deviation is $\sqrt{k(m-k)}/m$. The advantage is $\sqrt{(m-k)/k}$ on an accepted response and $-\sqrt{k/(m-k)}$ on a rejected one. Equations (A.4) imply

$$\mathbb{E}[\ell_i \mid V_i = 1] = \frac{\nabla p}{p}, \qquad \mathbb{E}[\ell_i \mid V_i = 0] = -\frac{\nabla p}{1-p}. \tag{A.7}$$

Conditional on the reward indicators, each response still has its corresponding accepted or rejected conditional law. Averaging over indicator locations with $K = k$ therefore gives

$$\mathbb{E}\left[\frac{1}{m}\sum_i A_i \ell_i \,\middle|\, K = k\right] = \frac{\sqrt{k(m-k)}}{m}\left(\frac{\nabla p}{p} + \frac{\nabla p}{1-p}\right) \tag{A.8}$$

$$= \frac{\sqrt{k(m-k)}}{m}\frac{\nabla p}{p(1-p)}. \tag{A.9}$$

For $k = 0$ or $k = m$, we define every advantage to be zero. The same formula then holds with zero scalar coefficient. Since $K \sim \mathrm{Bin}(m, p)$, the prompt-averaged identities are

$$g_c = \mathbb{E}_X\left[\frac{L_c(p_X)}{p_X(1-p_X)}\nabla p_X\right], \tag{A.10}$$

where

$$L_{M,m}(p) = \left(1 - \frac{1}{m}\right)p(1-p), \qquad L_{S,m}(p) = \mathbb{E}_{K\sim\mathrm{Bin}(m,p)}\frac{\sqrt{K(m-K)}}{m}. \tag{A.11}$$

These are finite-group identities, not a replacement of the vector gradient by a scalar utility proxy. Related normalization identities are studied by Bay & Yearick (2026). We have derived the form needed here to specify exactly the group convention and the zero-variance case.

*The standardized coefficient is difficulty-dependent.* Define $\alpha_m(p) = L_{S,m}(p)/[p(1-p)]$. Expanding the binomial expectation and using $\binom{m}{k}k(m-k) = m(m-1)\binom{m-2}{k-1}$ gives

$$\alpha_m(p) = (m-1)\sum_{z=0}^{m-2}\binom{m-2}{z}p^z(1-p)^{m-2-z}\psi_m(z), \tag{A.12}$$

$$\psi_m(z) = [(z+1)(m-1-z)]^{-1/2}. \tag{A.13}$$

Thus $\alpha_m$ is a Bernstein polynomial of degree $m-2$, multiplied by $m-1$. Extend $\psi_m$ to the real interval $[0, m-2]$, and write $u = z + 1 - m/2$ and $c = m^2/4$. Then

$$\psi_m(z) = (c - u^2)^{-1/2}, \qquad \psi_m''(z) = \frac{c + 2u^2}{(c-u^2)^{5/2}} \geq \frac{8}{m^3}. \tag{A.14}$$

For $m \geq 4$, differentiating the Bernstein basis twice yields

$$\alpha_m''(p) = (m-1)(m-2)(m-3)\mathbb{E}_{Z\sim\mathrm{Bin}(m-4,p)}[\psi_m(Z+2) - 2\psi_m(Z+1) + \psi_m(Z)]. \tag{A.15}$$

The second difference equals $\int_0^1\int_0^1 \psi_m''(Z+s+t)\,ds\,dt$, so

$$\alpha_m''(p) \geq c_m := \frac{8(m-1)(m-2)(m-3)}{m^3} > 0. \tag{A.16}$$

Symmetry under $p \mapsto 1-p$ gives $\alpha'_m(1/2) = 0$. Strict convexity then implies $\alpha'_m(p) < 0$ for $0 < p < 1/2$. This argument deliberately excludes $m = 2, 3$, where direct evaluation gives the constants $\alpha_2 = 1$ and $\alpha_3 = \sqrt{2}$.

*Numerical constants used in the construction.* Set $A_{16} = \alpha_{16}(1/2)$ and $B_{16} = \alpha_{16}(1/16)$. Integrating Eq. (A.16) twice from the symmetry point gives

$$B_{16} - A_{16} \geq \frac{c_{16}}{2}\left(\frac{1}{2} - \frac{1}{16}\right)^2 = \frac{66885}{131072} > \frac{1}{2}. \tag{A.17}$$

The elementary bounds $2/m \leq \psi_m(z) \leq 1/\sqrt{m-1}$ imply $A_{16} \geq 15/8$ and $B_{16} \leq \sqrt{15} < 4$. Also $\sqrt{K(m-K)}/m \leq 1/2$, so $L_{S,16}(1/2) \leq 1/2$ and $A_{16} \leq 2$. Consequently

$$\frac{5}{4} < \frac{B_{16}}{A_{16}} < \frac{32}{15} < \frac{9}{4}, \qquad \sqrt{5/4} < \lambda := \sqrt{B_{16}/A_{16}} < \frac{3}{2}. \tag{A.18}$$

The function $(\lambda-1)/\sqrt{2(\lambda^2+1)}$ is increasing for $\lambda > 1$. At $\lambda = \sqrt{5}/2$ it equals $(\sqrt{5}-2)/(3\sqrt{2}) > 1/20$. Hence

$$a_* := \frac{\lambda - 1}{\sqrt{2(\lambda^2+1)}} > \frac{1}{20}. \tag{A.19}$$

This lower bound ensures a nonzero geometric separation, rather than a reversal obtained by canceling nearly identical gradient components.

### A.3 The two training prompts

Let $p_* = 1/16$ and $k_1 = p_*\lambda$. Table A.1 specifies the other two prompts. All listed probabilities are strictly positive because $k_1 < 3/32 < 1/2$. For each prompt, the total mass on each of the two

Table A.1: Two training prompts in the explicit construction. Entries within each row are ordered by response identity.

| Prompt | Initial probabilities | Features | Accepted |
|---|---|---|---|
| $x = 1$ | $\big((1/2+k_1)/2, (1/2-k_1)/2,$ $(1/2-k_1)/2, (1/2+k_1)/2\big)$ | $(e_1, -e_1, e_1, -e_1)$ | First two |
| $x = 2$ | $(p_*, 1/2-p_*, 1/2)$ | $(e_2, e_2, -e_2)$ | First one |

opposite features is $1/2$. Its initial feature mean is zero, and its partition function is respectively $\cosh\theta_1$ or $\cosh\theta_2$. At zero, differentiating the accepted mass gives

$$p_1 = \frac{1}{2}, \qquad \nabla p_1 = \left(\frac{1/2+k_1}{2} - \frac{1/2-k_1}{2}\right)e_1 = p_*\lambda e_1, \tag{A.20}$$

$$p_2 = p_*, \qquad \nabla p_2 = p_* e_2. \tag{A.21}$$

The probe contributes zero because its pass rate is constant. For group size 16, Eq. (A.10) therefore becomes

$$g_M = \frac{15p_*}{48}(\lambda, 1), \qquad g_S = \frac{p_*}{3}(A_{16}\lambda, B_{16}) = \frac{p_* A_{16}\lambda}{3}(1, \lambda). \tag{A.22}$$

Both components are positive. In particular, $\|g_M\| > 15\sqrt{2}/768 > 1/40$, and the displayed lower bounds on $A_{16}$ and $\lambda$ give $\|g_S\| > 1/40$ as well. Normalizing gives

$$d_M = \frac{(\lambda, 1)}{\sqrt{\lambda^2+1}}, \qquad d_S = \frac{(1, \lambda)}{\sqrt{\lambda^2+1}}, \qquad v^\top d_M = a_*, \quad v^\top d_S = -a_*. \tag{A.23}$$

These are genuinely different shared-parameter directions, produced using only the common verifier.

*Existence and uniqueness of actual KL calibration.* For a unit direction $d$, let $F_d(r) = \mathrm{KL}(P_0 \| P_{rd})$. Using Eq. (A.1) gives

$$F_d(r) = \frac{1}{3}\sum_x \left[\log Z_x(rd) - r\mathbb{E}_{\pi_0(\cdot|x)} d^\top \phi_{x,Y}\right], \tag{A.24}$$

$$F_d(0) = F'_d(0) = 0, \qquad F''_d(r) = \frac{1}{3}\sum_x \mathrm{Var}_{\pi_{rd}(\cdot|x)}(d^\top \phi_{x,Y}). \tag{A.25}$$

At least one coordinate of $d$ is nonzero. The corresponding training prompt puts positive probability on both opposite coordinate features for every finite $r$, so $F_d''(r) > 0$. Integrating from zero proves $F_d'(r) > 0$ for $r > 0$.

The same training prompt contributes $\log\cosh(rd_j)/3$, which diverges as $r \to \infty$. Each remaining prompt's KL contribution is nonnegative. Thus $F_d$ is continuous, strictly increasing from zero to infinity on the positive ray. Every $k_0 > 0$ has exactly one positive solution. Since the derivative is nonzero at that solution, the root also varies continuously under small changes in probabilities, features, and nonzero directions.

For mean centering at any $m \in \{4, 8, 16\}$, Eq. (A.6) multiplies the same direction by $1 - 1/m > 0$. Radial calibration removes this scalar difference. The three mean configurations therefore yield one ideal policy, while the standardized candidates must be computed with their respective $\alpha_m$.

### A.4 Block redistribution

Write $a = \sum_I \sigma_i w_i \in (-1, 1)$ and let $s(\theta)$ be the signed conditional block mass. Define the log odds along a unit direction $d$ by

$$\Lambda_d(r) = \log \frac{\sum_{i \in I_+} w_i e^{rd^\top \phi_i}}{\sum_{i \in I_-} w_i e^{rd^\top \phi_i}}. \tag{A.26}$$

Then $\Lambda_d(0) = \log[(1 + a)/(1 - a)] = 2 \operatorname{arctanh} a$ and $s(rd) = \tanh[\Lambda_d(r)/2]$. Differentiating the two log sums gives

$$\Lambda_d'(r) = 2d^\top v + \mathbb{E}_{P_{rd}(\cdot|I_+)} d^\top \xi_i - \mathbb{E}_{P_{rd}(\cdot|I_-)} d^\top \xi_i. \tag{A.27}$$

Each residual expectation lies in $[-\kappa, \kappa]$. Therefore $\Lambda_{d_M}'(r) \geq 2(a_* - \kappa)$ and $\Lambda_{d_S}'(r) \leq -2(a_* - \kappa)$. When $\kappa < a_*$, integration and monotonicity of $\tanh$ yield

$$s(r_M d_M) \geq \tanh[\operatorname{arctanh} a + (a_* - \kappa) r_M] > a, \tag{A.28}$$

$$s(r_S d_S) \leq \tanh[\operatorname{arctanh} a - (a_* - \kappa) r_S] < a. \tag{A.29}$$

The positive radii may differ, since the comparison matches KL rather than radius. Equation (A.3) implies that the mean-minus-standardized contrast has

$$A = \sum_I (P_M - P_S) = 0, \qquad B = \sum_I \sigma_i (P_M - P_S) = \frac{s(r_M d_M) - s(r_S d_S)}{6} > 0. \tag{A.30}$$

This is the actual probability contrast used in the next appendix. The factor $1/6$ includes both prompt probability and verifier acceptance, so it must not be confused with a change conditional on acceptance alone.

## B Finite-precision separation and disclosure lower bound

We now keep the candidates fixed and randomize only the locations of a fixed binary utility. The proof first produces a common report and opposite utility gaps with high prior probability. It then reduces successful selection to distinguishing the two original label priors. Keeping those two steps separate is essential: conditioning the prior on the first step's event would invalidate the simple query probabilities used in the second.

### B.1 Full quantitative statement

For a frozen pair, let $d_i = P_c(i) - P_b(i)$ include prompt mass. For any real vector $q$ on $I$, define

$$A = \sum_I d_i, \qquad B = \sum_I \sigma_i d_i, \tag{B.1}$$

$$\omega_\sigma(q) = \max_{i \in I_\sigma} q_i - \min_{i \in I_\sigma} q_i, \qquad r(q, \alpha) = \sqrt{\frac{R}{2} \log \frac{8}{\alpha}}\, [\omega_+(q) + \omega_-(q)], \tag{B.2}$$

$$r_\mu = r(w, \alpha), \qquad r_\Gamma = r(d, \alpha). \tag{B.3}$$

The report rounds $\mu(U) = \sum_I w_i U_i$ on a grid of width $2\tau$ with a bin centered at $1/2$. In world $z \in \{-1, 1\}$, independently in each block, choose uniformly a subset of $k_{z,\sigma} = R(1/2 + z\sigma\epsilon)$ positions and give exactly those positions utility one. Assume $0 < \epsilon \leq 1/8$ and integer counts. After this draw, $U$ is fixed. Repeating a query always returns the same label. The prior is a proof device for a worst-case statement about these fixed oracles.

The weights, features, verifier, candidate laws, $\epsilon$, and any label-free construction history are common and known. Let $\alpha, \delta > 0$ with $\alpha + \delta < 1/4$, and assume

$$\tau > \epsilon|a| + r_\mu, \qquad \eta := \epsilon B - |A|/2 - r_\Gamma > 0. \tag{B.4}$$

Define $\Gamma(U) = \sum_I d_i U_i = J_U(P_c) - J_U(P_b)$. Let $\mathcal{G}_z$ be the set of oracles with central-bin report and $z\Gamma(U) \geq \eta$.

**Theorem B.1** (Finite-precision separation)**.** *Under the setting above,* $\Pr_z(\mathcal{G}_z) \geq 1 - \alpha$ *for each* $z$*. A selector correct with probability at least* $1 - \delta$ *for every compatible fixed oracle with* $|\Gamma(U)| \geq \eta$*, using at most* $n$ *point-label queries, obeys Eq.* (8)*.*

The proof is in Appendix B.2. Both utility worlds can fall in the same reporting bin, while the locations of correct responses still reverse the preferred update. Individual labels must reveal this hidden block imbalance before the selector can reliably choose a direction.

The success probability is over the selector's randomization, whereas $\Pr_z$ averages over the auxiliary prior on fixed oracles. The argument below converts the former guarantee into a bound for the latter without assuming that the selector knows the world.

### B.2 Proof of Theorem B.1

*Proof. Concentration for fixed-count labels.* Let $T$ be the sum of a uniformly chosen size-$k$ subset of $R$ numbers $q_1, \ldots, q_R$, with mean $\overline{q}$ and range $\omega$. For $k \geq 1$ and $\omega > 0$,

$$\Pr(|T - k\overline{q}| \geq t) \leq 2\exp\left(-\frac{2t^2}{k\omega^2}\right), \qquad t > 0. \tag{B.5}$$

Here is a derivation that makes the sampling-without-replacement step explicit. For any real $\lambda$, the moment generating function is the normalized elementary symmetric polynomial in $a_i = e^{\lambda q_i} > 0$:

$$\mathbb{E}e^{\lambda T} = \binom{R}{k}^{-1} \sum_{|S|=k} \prod_{i \in S} a_i \leq \left(\frac{1}{R}\sum_i a_i\right)^k. \tag{B.6}$$

To see the inequality, hold the sum of the $a_i$ fixed and average a pair $a_i, a_j$. Terms involving neither or exactly one of them do not change, while terms involving both cannot decrease because their product increases. Repeated pair averaging, with continuity, bounds the symmetric polynomial by its value at equal coordinates. The right-hand side is therefore the moment generating function of $k$ independent draws with replacement.

The bounded-variable exponential inequality then gives $\mathbb{E}e^{\lambda(T - k\overline{q})} \leq e^{k\lambda^2\omega^2/8}$. Markov's inequality, optimized at $\lambda = 4t/(k\omega^2)$, proves the upper tail, and applying it to $-T$ proves the lower tail. These are standard concentration ingredients (Lattimore & Szepesvári, 2020). If $k = 0$ or $\omega = 0$, the sum is deterministic and no concentration failure is possible.

*Applying the bound to reports and gains.* In each block apply Eq. (B.5) to $q = w$ and $q = d$. At threshold $\omega_\sigma(q)\sqrt{R\log(8/\alpha)/2}$, the failure probability is at most $\alpha/4$, because $k_{z,\sigma} \leq R$. A union bound covers all four deviations. For any $q$, its prior mean is

$$\mathbb{E}_z \sum_I q_i U_i = \sum_\sigma \left(\frac{1}{2} + z\sigma\epsilon\right) \sum_{i \in I_\sigma} q_i = \frac{1}{2}\sum_I q_i + z\epsilon \sum_I \sigma_i q_i. \tag{B.7}$$

Taking $q = w$ uses $\sum_I w_i = 1$ and $\sum_I \sigma_i w_i = a$. Taking $q = d$ gives $A$ and $B$. Thus, on an event of probability at least $1 - \alpha$,

$$|\mu(U) - (1/2 + z\epsilon a)| \leq r_\mu, \qquad |\Gamma(U) - (A/2 + z\epsilon B)| \leq r_\Gamma. \tag{B.8}$$

The first condition in Eq. (B.4) implies $|\mu(U) - 1/2| \le \epsilon|a| + r_\mu < \tau$, strictly inside the common report bin. The second gives

$$z\Gamma(U) \ge zA/2 + \epsilon B - r_\Gamma \ge \epsilon B - |A|/2 - r_\Gamma = \eta. \tag{B.9}$$

Both desired properties therefore hold simultaneously. This proves $\Pr_z(\mathcal{G}_z) \ge 1 - \alpha$.

*Why the report is fixed instead of conditioning the prior.* Construct an auxiliary experiment that draws $U$ from the original world-$z$ prior but always sends the selector the central report. On every $U \in \mathcal{G}_z$, this is exactly the real report. Couple the two executions using the same oracle and selector randomness. They have identical queries, answers, and decisions on that event. Consequently, the auxiliary selector's probability of choosing the wrong world is at most

$$\delta \Pr_z(\mathcal{G}_z) + \Pr_z(\mathcal{G}_z^c) \le \delta + \alpha. \tag{B.10}$$

For this pair, choosing $c$ identifies world $+1$ and choosing $b$ identifies world $-1$. The same reduction for a larger menu is proved in Appendix C.

Crucially, the auxiliary experiment does not condition the label draw on $\mathcal{G}_z$. Such conditioning could favor particular label positions according to the weights and candidate contrasts. By forcing the report to be constant instead, we retain exchangeability among undisclosed identities and charge the exceptional oracles through the additive $\alpha$ in Eq. (B.10).

*Information in an adaptively selected new label.* Consider a common transcript with $k$ distinct queried identities and $s$ positive answers in block $I_\sigma$. Conditional on these answers, every remaining labeling with the required residual count is equally likely. Any particular fresh identity in that block therefore has success probability

$$p_z = \frac{R(1/2 + z\sigma\epsilon) - s}{R - k}. \tag{B.11}$$

Known unequal weights can influence which identity the selector requests, but not this conditional probability: the prior placed positives uniformly by identity. With at most $R/8$ total disclosures, $k \le R/8$, and the initial number of positives in either world lies between $3R/8$ and $5R/8$. Using $0 \le s \le k$ gives $p_z \ge 1/4$ and $p_z \le 5/7 < 3/4$. All these short label histories have positive probability in both worlds, since neither the positive nor the negative count can be exhausted. Moreover,

$$|p_+ - p_-| = \frac{2R\epsilon}{R - k} \le \frac{16\epsilon}{7}. \tag{B.12}$$

For Bernoulli distributions, the inequality $\log t \le t - 1$ implies

$$\mathrm{kl}(p\|q) \le \frac{(p-q)^2}{q(1-q)}. \tag{B.13}$$

Since $q \in [1/4, 3/4]$, each fresh answer has conditional divergence at most

$$\frac{(16\epsilon/7)^2}{3/16} = \frac{4096}{147}\epsilon^2 < 32\epsilon^2. \tag{B.14}$$

Repeated queries return already known labels, and queries outside $I$ return known zeros. They contribute no further divergence.

*From individual answers to an entire adaptive transcript.* Include the selector's independent random seed in the transcript. Given the same observed history, the rule for choosing the next identity is identical under both worlds, so query choices themselves add zero conditional divergence. If the selector stops before its deterministic cap $n$, pad the transcript with a common dummy symbol. The KL chain rule then gives, whenever $n \le R/8$,

$$\mathrm{KL}(\mathbb{P}_+^{\mathrm{tr}} \| \mathbb{P}_-^{\mathrm{tr}}) \le 32n\epsilon^2. \tag{B.15}$$

This explicitly permits adaptive identity choices and stopping before the cap. It is not a statement for an arbitrary expected stopping-time budget.

*Testing implies the disclosure bound.* For any two transcript laws $P, Q$, a test's two error probabilities sum to at least $1 - \mathrm{TV}(P, Q)$. One standard binary-testing inequality bounds this below by $\frac{1}{2}e^{-\mathrm{KL}(P\|Q)}$ (Lattimore & Szepesvári, 2020, Chapter 14). For clarity, let $\rho = \int \sqrt{dP\, dQ}$

be their affinity. Jensen's inequality gives $\rho \geq e^{-\mathrm{KL}(P\|Q)/2}$, and Cauchy–Schwarz gives $\mathrm{TV}(P,Q) \leq \sqrt{1-\rho^2}$. Since $1-\sqrt{1-x} \geq x/2$, the claimed testing inequality follows.

Combining this inequality with Eqs. (B.10) and (B.15) yields

$$2(\delta+\alpha) \geq \frac{1}{2}e^{-32n\epsilon^2}, \qquad n \geq \frac{\log(1/[4(\delta+\alpha)])}{32\epsilon^2}. \tag{B.16}$$

The logarithm is positive by assumption. If instead $n > R/8$, the finite-support branch of Eq. (8) already holds. These cases prove the stated minimum bound. □

### B.3 Support, precision, and perturbations

*A sufficient support size.* Suppose $\omega_\sigma(w) \leq C_w/R$, $\omega_\sigma(d) \leq C_g/R$, $|a| \leq a_0$, $B \geq b_0 > 0$, and $|A| \leq \epsilon b_0/2$. Then

$$r_\mu \leq C_w\sqrt{\frac{2\log(8/\alpha)}{R}}, \qquad r_\Gamma \leq C_g\sqrt{\frac{2\log(8/\alpha)}{R}}. \tag{B.17}$$

If $\tau > \epsilon a_0$, it suffices to require

$$R > \max\left\{\frac{2C_w^2\log(8/\alpha)}{(\tau-\epsilon a_0)^2}, \frac{32C_g^2\log(8/\alpha)}{\epsilon^2 b_0^2}\right\}. \tag{B.18}$$

The first inequality makes $r_\mu < \tau - \epsilon a_0$. The second gives $r_\Gamma < \epsilon b_0/4$, while $|A|/2 \leq \epsilon b_0/4$. Therefore $\eta \geq \epsilon b_0/2$. Integer label counts must still be respected. These are sufficient conditions, not necessary thresholds, and allow fixed relative within-block dispersion as support grows.

*Why a nonempty perturbation region exists.* Start at the balanced equal-weight point with $\kappa = 0$. Every $w_i = 1/(2R)$, and each candidate's probabilities are constant within each block. Hence $a = A = r_\mu = r_\Gamma = 0$, while Appendix A gives $B > 0$. Every fixed $\tau > 0$ therefore leaves a strict reporting margin and a positive decision margin.

At fixed support, probabilities and finite-group expected gradients are continuous in positive initial weights and bounded features. The constructed gradients are nonzero. Equation (A.25) makes every positive calibration root continuous as well. Maxima, minima, and the range radii are continuous functions of finitely many entries. Strict inequalities consequently persist in a sufficiently small neighborhood. Small perturbations may break accepted/rejected pairing, in which case $A$ need not stay zero. Its effect is already included in Eq. (B.4). This separation argument does not extend the paired reuse theorem to those non-paired policies.

*Why exact means require a different conclusion.* For distinct binary label vectors $u, u'$, equal exact means impose $w^\top(u-u') = 0$. This is a proper hyperplane of the weight simplex, or has no intersection with it. There are only finitely many binary pairs. The complement of their hyperplanes is dense and intersects the open separation region. At a weight vector in that complement, all binary means are distinct. Since there are finitely many, their minimum pairwise separation is positive. A grid finer than that separation cannot place two such means in the same bin. Thus exact or sufficiently precise real-valued reporting can encode all labels on fixed support, even though reporting at the specified coarser resolution cannot.

## C Complete menus, sampled reports, and finite rollouts

This appendix connects Theorem 2 to complete configuration menus, sampled reports, and empirical updates. Simultaneous gain intervals exclude a common good configuration, and coupling handles a sampled reliability report. The later sections control the probability of generating a separating menu before rollouts are drawn, then give symmetric and perturbed finite-budget examples.

### C.1 Full-menu statement

Let $K$ denote the number of nonbase candidate contrasts in the menu, counting configuration names even if some laws coincide. For each candidate set

$$d_{c,i} = P_c(i) - P_0(i), \quad A_c = \sum_I d_{c,i}, \quad B_c = \sum_I \sigma_i d_{c,i}. \tag{C.1}$$

Use the radius definition in Eq. (B.3), now with $\log(4(K+1)/\alpha)$ in place of $\log(8/\alpha)$, for both the report and all contrasts. Denote the candidate gain radius by $r_c$. If base is included, set its contrast and radius to zero. Define

$$q_{z,c} = A_c/2 + z\epsilon B_c, \quad L_z = \max_c (q_{z,c} - r_c), \quad \mathcal{C}_z(t) = \{c : q_{z,c} + r_c \geq L_z - t\}. \tag{C.2}$$

Here $L_z$ is a lower bound on the best gain, and $\mathcal{C}_z(t)$ is an outer set of potentially acceptable candidates. Membership alone does not certify low regret for a particular oracle.

**Corollary C.1** (Full-menu decision)**.** *Use the fixed-count families and error conditions of Appendix B. If the report precision condition holds with the simultaneous radius and*

$$\mathcal{C}_+(t) \cap \mathcal{C}_-(t) = \varnothing, \tag{C.3}$$

*every selector attaining regret at most $t$ with probability at least $1-\delta$ for every compatible good oracle obeys Eq.* (8)*.*

Adding configurations could remove a pairwise dilemma if one new candidate worked well in both worlds. The disjoint-set condition rules out exactly this escape: any sufficiently good menu choice reveals which world holds. The pairwise information lower bound therefore also constrains selection from the full menu.

*Proof. Simultaneous concentration.* There are $2(K+1)$ block/quantity deviations: two for the report and two for each nonbase contrast. The new logarithm makes each two-sided failure at most $\alpha/[2(K+1)]$. A union bound gives an event of probability at least $1-\alpha$ on which the report is common and

$$q_{z,c} - r_c \leq \Gamma_c(U) := J_U(P_c) - J_U(P_0) \leq q_{z,c} + r_c \quad \text{for every } c. \tag{C.4}$$

No independence between candidate contrasts is needed. For Theorem 2, define $\mathcal{G}_z$ to be the fixed-count oracles with the central report that satisfy all inequalities in Eq. (C.4). The event just constructed is contained in $\mathcal{G}_z$, so $\Pr_z(\mathcal{G}_z) \geq 1-\alpha$. This is the simultaneous good set for the full menu, rather than only the pairwise good set of Appendix B.

*Reduction from low regret to identification.* On this event, $\max_c \Gamma_c(U) \geq L_z$. If a returned candidate $\widehat{c}$ has regret at most $t$, then

$$q_{z,\widehat{c}} + r_{\widehat{c}} \geq \Gamma_{\widehat{c}}(U) \geq \max_c \Gamma_c(U) - t \geq L_z - t. \tag{C.5}$$

Thus $\widehat{c} \in \mathcal{C}_z(t)$. Under Eq. (C.3), membership identifies $z$ unambiguously whenever selection succeeds. Outputs outside both sets can be decoded arbitrarily, since they cannot occur on a successful good-oracle decision. The constant-report auxiliary experiment then has error at most $\delta + \alpha$, and the transcript proof in Appendix B applies unchanged. □

*Analytic nonemptiness and the role of the third candidate.* At the balanced equal-weight point, $A_c = r_c = 0$ for every candidate. Write $B_{\max} = \max_c B_c$ and $B_{\min} = \min_c B_c$. Then

$$\mathcal{C}_+(t) = \{c : B_c \geq B_{\max} - t/\epsilon\}, \qquad \mathcal{C}_-(t) = \{c : B_c \leq B_{\min} + t/\epsilon\}. \tag{C.6}$$

Intersection would require $B_{\max} - B_{\min} \leq 2t/\epsilon$. Appendix A supplies candidates with positive and negative redistribution, so the span is strictly positive. Any choice

$$0 < t < \frac{\epsilon}{2}(B_{\max} - B_{\min}) \tag{C.7}$$

therefore gives disjoint sets. This calculation includes all intermediate standardized candidates and any base candidate, rather than assuming a pairwise reversal excludes them.

To check robustness, for each candidate choose a world in whose acceptable set it does not lie. Nonmembership means a strict inequality $q_{z,c} + r_c < L_z - t$. Finitely many such strict inequalities persist under a sufficiently small common perturbation. Together with the report condition and continuous KL calibration, this gives the nonempty open full-menu separation region claimed in Theorem 2.

*A matching-rate audit within the fixed-count family.* Choose a block uniformly, then choose an identity uniformly within that block, and observe its label. Let $Z = \sigma(U - 1/2) \in \{-1/2, 1/2\}$. For every fixed oracle in world $z$, the known label counts imply

$$\mathbb{E}[Z \mid \sigma] = \sigma\left(\frac{1}{2} + z\sigma\epsilon - \frac{1}{2}\right) = z\epsilon. \tag{C.8}$$

This holds irrespective of the initial response weights. For $n$ independent draws with replacement, a sign test on their mean has error at most $2e^{-2n\epsilon^2}$, using a conservative two-sided bound. Thus $O(\epsilon^{-2}\log(1/\delta))$ draws identify the world. Cached labels on repeated identities do not alter this sampling calculation.

After identifying $z$, choose $\widetilde{c} \in \arg\max_c q_{z,c}$. On Eq. (C.4), if $c^*$ is actually optimal, then

$$\Gamma_{c^*}(U) - \Gamma_{\widetilde{c}}(U) \le q_{z,c^*} + r_{c^*} - q_{z,\widetilde{c}} + r_{\widetilde{c}} \tag{C.9}$$

$$\le 2\max_c r_c. \tag{C.10}$$

Attaining a specified tolerance also requires $2\max_c r_c \le t$. This upper bound uses the fixed-count family, not all binary utilities. In particular, its uniform-by-identity sampling differs from the initial-conditional-weight sampling in Theorem 3. Substituting one proposal for the other at unequal weights would estimate a different moment.

All arguments may be conditioned on a label-free finite-rollout history whose actual candidate laws meet the same certificates. This conditional statement does not assert success for every history. Appendix C.4 quantifies the construction probability under a stated sampling rule.

### C.2 Proof of Theorem 2

*Proof.* We combine the preceding construction and information bounds, specifying where each part of the main statement enters.

First, Appendix A constructs a single reference policy and verifier with nonzero mean and standardized GRPO gradients. The normalized mean-16 and standardized-16 directions have opposite projections onto $v$, and every nonzero direction admits a unique positive radial calibration to a common KL target. The other prescribed configurations enter the same finite menu. At the balanced equal-weight paired point, the candidates are constant within each accepted block, so $A_c = r_c = 0$ for every candidate and the report radius is zero. Opposite redistribution implies $B_{\max} - B_{\min} > 0$. Choose any $t$ satisfying Eq. (C.7). The two full-menu acceptable sets are then disjoint, including all intermediate configurations and base if it is offered.

Second, assign the fixed-count utility priors of Appendix B. The initial policy, verifier, features, candidates, and label-free construction history are identical in the two worlds. At the balanced point the accepted correctness report is exactly $1/2$ in both worlds. More generally, the simultaneous concentration event in Eq. (C.4), together with the reporting precision condition, gives $\mathcal{G}_z$ with prior probability at least $1 - \alpha$ and the same report $\mathsf{R}_\tau(U)$ throughout both sets. On these sets, attaining regret at most $t$ identifies the utility world by Corollary C.1. The premise $\inf_{U \in \mathcal{G}_+ \cup \mathcal{G}_-} \Pr_U(\mathcal{R}_{\mathcal{C}}(\widehat{c}) \le t) \ge 1 - \delta$ therefore bounds the auxiliary test error by $\delta + \alpha$. Under the stipulated information interface and deterministic disclosure cap, the transcript argument in Appendix B yields Eq. (8).

Finally, the calibrated probabilities and all finite-menu margins are continuous in positive initial weights and bounded features. The report and menu inequalities at the constructed point are strict. They therefore persist on a nonempty neighborhood, with the concentration radii and any change in accepted mass included as in Appendices B and C. This establishes the existence claim and its stated robustness. When empirical gradients replace expected gradients, the information argument applies conditional on a label-free realized history whose actual candidates satisfy these same certificates. It does not require identifying an empirical candidate with its expectation. □

### C.3 An audit formed from actual sampled labels

Suppose the audit service draws $i_1, \ldots, i_N$ independently from $P_0(\cdot \mid I)$ and releases their unlabeled identities together with the rounded empirical mean $\widehat{\mu}_N = N^{-1}\sum_j U(i_j)$. Fix an oracle in the

simultaneous good set. Conditional on this fixed oracle, the sampled labels are iid bounded functions of independent identities, even though an identity may repeat. Hoeffding's inequality gives

$$\Pr(|\widehat{\mu}_N - \mu(U)| > \zeta \mid U) \le 2e^{-2N\zeta^2}. \tag{C.11}$$

It follows that the report remains in the central bin with probability at least $1-\beta$ if

$$N \ge \frac{\log(2/\beta)}{2\zeta^2}, \qquad \tau > \epsilon|a| + r_\mu + \zeta. \tag{C.12}$$

The second condition leaves enough reporting slack for both prior-oracle fluctuation and sampling fluctuation.

Couple real and auxiliary executions using the same oracle, sampled identities, and selector randomness. Give the auxiliary selector the same identities but always the central report. On the good-oracle, central-report event, the executions agree. The probability of a mismatch is at most $\alpha+\beta$, without requiring these two exceptional events to be independent.

In the auxiliary experiment, the identities are drawn from the common reference law independently of the label placement. They therefore add zero world information. Conditional on these identities, the label prior remains the original fixed-count prior. The subsequent adaptive query calculation remains Eq. (B.11), even if the selector chooses to disclose labels of previously sampled identities.

A real selector with error at most $\delta$ for each compatible oracle has auxiliary error at most $\delta+\alpha+\beta$. The disclosure bound follows with that sum replacing $\delta+\alpha$, provided it is below $1/4$. The protocol permits a free list of unlabeled identities, not free identity–label pairs. The $N$ initial draws already have an acquisition cost. Later disclosures count as new procurement only if the service actually requires new oracle calls rather than access to its cached labels.

### C.4 Finite rollout construction and KL calibration

We now control the probability of constructing a separating menu before observing the rollout history. The ideal utility classes remain fixed. This strengthens the conditional history statement above without identifying an empirical gradient with its expectation.

Let the prompt law $p_X$ be fixed. Each prompt has finite response support, positive initial probabilities, and features in $\mathbb{R}^d$. Write

$$\pi_\theta(i \mid x) = \frac{\pi_0(i \mid x)\exp(\theta^\top \phi_{x,i})}{Z_x(\theta)}, \qquad \|\phi_{x,i}\| \le F, \qquad P_\theta(x,i) = p_X(x)\pi_\theta(i \mid x). \tag{C.13}$$

Here $F > 0$, the verifier $V$ is fixed and binary, and the initial parameter is zero. The utility $U$ is fixed when rollouts are drawn. Candidate construction has no access to its unknown labels. Let $\mathsf{H}$ contain the entire construction history, including prompts, responses, verifier outcomes, algorithmic randomness, and the resulting candidates.

There are $K$ nonbase configurations. Configuration $c$ uses group size $m_c$ and either centered binary advantages or their division by the group standard deviation with denominator $m_c$. Silent groups have zero advantage. With all advantages detached, one group's gradient at zero is

$$Y_{c,g} = \frac{1}{m_c}\sum_{i=1}^{m_c} A_{c,g,i}\nabla \log \pi_0(Y_{g,i} \mid X_g), \qquad g_c = \mathbb{E}Y_{c,g}, \qquad \widehat{g}_c = \frac{1}{G}\sum_{g=1}^{G} Y_{c,g}. \tag{C.14}$$

For each configuration, the $G$ groups are iid. Configurations need not be independent of one another. In the six-name menu with $m_c \in \{4,8,16\}$, one may draw a common prompt and sixteen iid responses per outer group, and use the first four, eight, or sixteen responses. This supplies all six marginal sampling laws with $16G$ response draws. The result concerns one empirical update for each candidate, not $G$ sequential optimization steps.

Fix a true KL target $k_0 > 0$ and set

$$\mathsf{K}(\theta) = \mathrm{KL}(P_0 \| P_\theta), \qquad d_c = g_c/\|g_c\|, \qquad \theta_c = r(d_c)d_c, \qquad \mathsf{K}(r(d)d) = k_0. \tag{C.15}$$

The corresponding empirical candidate is $\widehat{\theta}_c = r(\widehat{d}_c)\widehat{d}_c$, with $\widehat{d}_c = \widehat{g}_c/\|\widehat{g}_c\|$ when nonzero. If an empirical gradient is exactly zero, use a fixed unit direction, for example $e_1$, and calibrate that

direction instead. This defines the procedure on every history without resampling until success. Base, if offered, remains $\theta = 0$ and is not assigned a positive KL target.

For an index $c$, distinguish the two regrets

$$\mathcal{R}_U^0(c) = \max_j J_U(P_{\theta_j}) - J_U(P_{\theta_c}), \qquad \widehat{\mathcal{R}}_U(c) = \max_j J_U(P_{\widehat{\theta}_j}) - J_U(P_{\widehat{\theta}_c}), \quad J_U(P) = \mathbb{E}_P U. \tag{C.16}$$

The selected empirical candidate is executed as evaluated. The construction count $G$, the later disclosure cap $n$, and the block support size $R$ are different resources.

*Group gradients.*

**Lemma C.1** (Group size does not inflate this norm bound)**.** *Under* (C.13)–(C.14)*, a centered group satisfies* $\|Y_{M,g}\| \le F/2$*, and a standardized group satisfies* $\|Y_{S,g}\| \le F$*. Consequently,*

$$\Pr\left(\max_{c \le K} \|\widehat{g}_c - g_c\| > \eta\right) \le 2Kd \exp\left(-\frac{G\eta^2}{2dF^2}\right). \tag{C.17}$$

*Proof.* For a fixed prompt, the score is $\phi_i - \mathbb{E}_{\pi_0(\cdot|x)}\phi$. Both advantage conventions satisfy $\sum_i A_i = 0$, including the silent case. Thus the common score offset cancels exactly and $Y = m^{-1}\sum_i A_i\phi_i$. If $p = k/m$ is the group's pass fraction, then

$$\frac{1}{m}\sum_i |A_i| = 2p(1-p) \le \frac{1}{2}$$

for centering. For standardization and $0 < p < 1$, it is $2\sqrt{p(1-p)} \le 1$. The norm claims follow. In particular, a single large standardized advantage does not require a bound proportional to $\sqrt{m}$.

Each coordinate of a group gradient is in $[-F, F]$. Hoeffding's inequality at coordinate tolerance $\eta/\sqrt{d}$ gives $2\exp[-G\eta^2/(2dF^2)]$. A union bound over the $Kd$ coordinates proves (C.17). Dependence across configurations is harmless for this union bound. Different group counts can instead be handled by summing the individual configuration bounds. □

The cancellation also removes the common feature mean from empirical probe gradients. Their expectation may vanish while their sampling variance remains positive.

*KL calibration from initial Fisher information.* Let

$$I_0 = \mathbb{E}_X \operatorname{Cov}_{\pi_0(\cdot|X)}(\phi_{X,Y}), \qquad I_0 \succeq \lambda_{\mathrm{F}} I_d, \qquad \lambda_{\mathrm{F}} > 0. \tag{C.18}$$

A certified positive lower bound on the smallest eigenvalue is sufficient. This condition holds with $\lambda_{\mathrm{F}} = 1/3$ in the model's paired family, because the two training prompts contribute $(e_1 e_1^\top + e_2 e_2^\top)/3$. At the balanced, homogeneous probe the exact matrix is $(I_2 + vv^\top)/3$.

Define the following known constants:

$$r_- = \frac{\sqrt{2k_0}}{F}, \qquad r_+ = \frac{2Fk_0 + \sqrt{4F^2k_0^2 + 2\lambda_{\mathrm{F}}k_0}}{\lambda_{\mathrm{F}}}, \tag{C.19}$$

$$\nu = \frac{\lambda_{\mathrm{F}}}{2F}(1 - e^{-2Fr_-}), \quad \Lambda = \min\{2F, F^2 r_+\}, \qquad L_{\mathrm{cal}} = r_+\left(1 + \frac{\Lambda}{\nu}\right). \tag{C.20}$$

They require neither a preselected angular neighborhood nor a minimum atom probability.

**Lemma C.2** (Global stability of radial calibration)**.** *Under* (C.13) *and* (C.18)*, every unit direction has a unique positive root in* (C.15)*. It satisfies* $r_- \le r(d) \le r_+$ *and, for all unit* $d, d'$*,*

$$|r(d) - r(d')| \le \frac{\Lambda r_+}{\nu}\|d - d'\|, \tag{C.21}$$

$$\|r(d)d - r(d')d'\| \le L_{\mathrm{cal}}\|d - d'\|. \tag{C.22}$$

*Proof.* The finite exponential family gives

$$\nabla^2 \mathsf{K}(\theta) = \mathbb{E}_X \operatorname{Cov}_{\pi_\theta(\cdot|X)}(\phi_{X,Y}), \qquad \mathsf{K}(0) = 0, \qquad \nabla \mathsf{K}(0) = 0.$$

For every prompt, the likelihood ratio of $\pi_\theta$ to $\pi_0$ lies in $[e^{-2F\|\theta\|}, e^{2F\|\theta\|}]$. Using $\operatorname{Var}_Q(T) = \inf_a \mathbb{E}_Q(T-a)^2$, for any vector $u$,

$$\operatorname{Var}_{\pi_\theta}(u^\top \phi) \geq e^{-2F\|\theta\|} \operatorname{Var}_{\pi_0}(u^\top \phi).$$

Also $u^\top \phi \in [-F\|u\|, F\|u\|]$, so its variance is at most $F^2\|u\|^2$. Therefore, globally,

$$\lambda_{\mathrm{F}} e^{-2F\|\theta\|} I_d \preceq \nabla^2 \mathsf{K}(\theta) \preceq F^2 I_d. \tag{C.23}$$

For a unit $d$, put $f_d(r) = \mathsf{K}(rd)$. Integrating (C.23) yields

$$f_d'(r) \geq \frac{\lambda_{\mathrm{F}}}{2F}(1 - e^{-2Fr}) > 0 \quad (r > 0), \qquad f_d(r) \leq \frac{F^2 r^2}{2}. \tag{C.24}$$

Since $1 - e^{-u} \geq u/(1+u)$ for $u \geq 0$, integration also gives

$$f_d(r) \geq \int_0^r \frac{\lambda_{\mathrm{F}} t}{1 + 2Ft}\, dt \geq \frac{\lambda_{\mathrm{F}} r^2}{2(1 + 2Fr)}. \tag{C.25}$$

The last lower bound diverges with $r$. Strict monotonicity and continuity prove existence and uniqueness. The upper bound in (C.24) implies $r(d) \geq r_-$. Solving $\lambda_{\mathrm{F}} r^2 = 2k_0(1 + 2Fr)$ proves $r(d) \leq r_+$. For every $r \in [r_-, r_+]$, (C.24) gives $f_d'(r) \geq \nu$. This derivative bound holds on the entire comparison interval, not just at an ideal root.

On the radius-$r_+$ ball, (C.23) and the gradient at zero give $\|\nabla \mathsf{K}(\theta)\| \leq F^2 r_+$. The gradient is also an average difference of two feature means and hence has norm at most $2F$. This proves the common bound $\Lambda$.

Suppose $r = r(d) \leq r' = r(d')$. Radial monotonicity and the derivative bound imply

$$\nu(r' - r) \leq \mathsf{K}(r'd') - \mathsf{K}(rd') = \mathsf{K}(rd) - \mathsf{K}(rd') \leq \Lambda r \|d - d'\|.$$

The parameter line segment used in the last step is inside the ball. Exchanging the directions gives (C.21). Finally,

$$\|rd - r'd'\| \leq r\|d - d'\| + |r - r'| \leq L_{\mathrm{cal}} \|d - d'\|.$$

No path along the unit sphere is required. □

**Lemma C.3** (Uniform utility continuity)**.** *For every two finite parameters and every $U$ with values in* $[0, 1]$,

$$|J_U(P_\theta) - J_U(P_\vartheta)| \leq \mathrm{TV}(P_\theta, P_\vartheta) \leq \tanh\left(\frac{F}{2}\|\theta - \vartheta\|\right) \leq \frac{F}{2}\|\theta - \vartheta\|. \tag{C.26}$$

*Proof.* For each prompt, $\pi_\theta$ is the normalized tilt of $\pi_\vartheta$ by $f_i = (\theta - \vartheta)^\top \phi_{x,i}$. Its oscillation is at most $2F\|\theta - \vartheta\|$. Lemma D.1 gives the conditional bound, and averaging over the common prompt law proves the joint bound. The final inequality is $\tanh u \leq u$ for $u \geq 0$. More precisely, one can average $\tanh(\mathrm{osc}_i(f_i)/4)$ over prompts instead of taking the common feature-norm bound. The first inequality is the variational property of total variation for bounded utility. □

### C.5 Realization from a fixed ideal separation

The ideal menu supplies sets of acceptable indices. In the manuscript they are

$$\mathcal{A}_z = \{c : q_{z,c} + r_c \geq L_z - t_0\}, \qquad L_z = \max_c (q_{z,c} - r_c), \qquad \mathcal{A}_+ \cap \mathcal{A}_- = \varnothing. \tag{C.27}$$

Here $r_c$ is the statistical gain radius, not the radial calibration root. The result below only needs the consequence of these intervals. Before drawing $\mathsf{H}$, fix good-oracle sets $\mathcal{G}_z^0$ of prior probability at least $1 - \alpha$ such that the reliability report is the common value and

$$U \in \mathcal{G}_z^0, \quad \mathcal{R}_U^0(c) \leq t_0 \quad \Longrightarrow \quad c \in \mathcal{A}_z. \tag{C.28}$$

The utility prior is not conditioned on these sets. For the realization statement itself only (C.28) is needed.

**Theorem C.1** (Unconditional realization with explicit constants)**.** *Assume* (C.13)*,* (C.18)*, and* (C.28)*, with* $\gamma = \min_c \|g_c\| > 0$ *and* $0 < t < t_0$*. Choose* $0 < \rho < 1$ *and*

$$0 < \eta \le \min\left\{\frac{\gamma}{2}, \frac{\gamma(t_0 - t)}{4FL_{\mathrm{cal}}}\right\}, \qquad G \ge \frac{2dF^2}{\eta^2}\log\frac{2Kd}{\rho}. \tag{C.29}$$

*With probability at least* $1 - \rho$ *over construction alone, simultaneously for every fixed* $U \in \mathcal{G}_+^0 \cup \mathcal{G}_-^0$ *and every index* $c$*,*

$$\widehat{\mathcal{R}}_U(c) \le t \quad \Longrightarrow \quad c \in \mathcal{A}_z \quad \text{when } U \in \mathcal{G}_z^0. \tag{C.30}$$

*All nonbase candidates have KL* $k_0$*, including on exceptional histories under the prescribed zero-gradient fallback.*

*Proof.* Let $\mathcal{E}_G = \{\max_c \|\widehat{g}_c - g_c\| \le \eta\}$. Lemma C.1 gives $\Pr(\mathcal{E}_G) \ge 1 - \rho$. On this event all empirical gradients are nonzero. Adding and subtracting $\widehat{g}_c/\|g_c\|$ and using the reverse triangle inequality gives

$$\|\widehat{d}_c - d_c\| \le \frac{2\eta}{\gamma}.$$

Lemmas C.2 and C.3 give, for every $U \in [0, 1]$ simultaneously,

$$\max_c |J_U(P_{\widehat{\theta}_c}) - J_U(P_{\theta_c})| \le q, \qquad q = \tanh\left(\frac{FL_{\mathrm{cal}}\eta}{\gamma}\right). \tag{C.31}$$

The base candidate has zero discrepancy. Taking the maximum over indices in the two menus shows

$$\mathcal{R}_U^0(c) \le \widehat{\mathcal{R}}_U(c) + 2q. \tag{C.32}$$

By (C.29) and $\tanh u \le u$, $2q \le (t_0 - t)/2$, so $t + 2q \le (t + t_0)/2 < t_0$. Equation (C.28) then proves (C.30). The event is defined using construction alone and the utility bound holds for all $U$, which proves the claimed order of quantifiers. Global root existence also proves that the fallback is defined and has the prescribed positive KL. □

The sharp total variation bound also permits $\eta \le \gamma \operatorname{arctanh}((t_0 - t)/4)/(FL_{\mathrm{cal}})$, in addition to $\eta \le \gamma/2$, when $0 < t_0 - t < 4$. This keeps the same half-margin slack because $2q \le (t_0 - t)/2$. The simpler condition in Eq. (C.29) remains sufficient.

For a utility independently drawn from either original prior, the report and selection property consequently hold jointly with probability at least $(1-\alpha)(1-\rho)$. The stronger formulation above specifies one construction event that works for every fixed good oracle at once.

The theorem preserves the actual decision separation. It does not assert that every entry of the original sample-dependent sufficient radius certificate must pass when recomputed. Avoiding that stronger, unnecessary demand allows the good-oracle sets to remain fixed before rollout sampling.

For the specified paired family the exact minimum over all six expected gradient norms is

$$\gamma = \|g_{M,4}\| = \frac{\sqrt{1 + \lambda^2}}{64} \approx 0.0253270108. \tag{C.33}$$

Indeed $g_{M,m} = (1 - 1/m)(\lambda, 1)/48$. Also the empirical standard deviation $\widehat{s}$ lies in $[0, 1/2]$, so $\widehat{s} \ge 2\widehat{s}^2$ and $\alpha_m(p) \ge 2(1 - 1/m)$. Both coordinates of the standardized expected gradient are at least twice those of the mean counterpart. Thus the minimum is attained at mean-4. One can use $F = 1 + \kappa$ and $\lambda_{\mathrm{F}} = 1/3$ while the two training prompts remain unchanged.

The theorem also applies to a fixed non-paired perturbation when its ideal certificate holds and its known $F, \lambda_{\mathrm{F}}, \gamma$ satisfy the assumptions. On a sufficiently small closed neighborhood of a strict ideal instance, these constants can be bounded uniformly by continuity. This does not extend the specified separate paired reuse theorem to non-paired models.

### C.6 Disclosure after random candidate construction

Take $0 < \delta < 1$ and $0 \le \alpha < 1$. Use the original fixed-count priors $\Pi_z$. In each signed block, the number of positive labels is $R(1/2 + z\sigma\epsilon)$, with $0 < \epsilon \le 1/8$ and integer counts. The selector receives the complete construction history and the specified reliability report, can disclose at most $n$ individual labels, and returns a candidate index. An independent random seed may be included in its transcript.

The success premise is now explicitly unconditional in the construction:

$$\sup_{z\in\{+,-\}} \sup_{U\in\mathcal{G}_z^0} \Pr_{\mathsf{H},\mathrm{selector}}(\widehat{\mathcal{R}}_U(\widehat{c}) > t) \le \delta. \tag{C.34}$$

It is not a guarantee postulated separately for every successful history.

**Theorem C.2** (Additional disclosures after random candidate construction)**.** *Under Theorem C.1 and* (C.34)*, if $\delta + \alpha + \rho < 1/4$, then*

$$n \ge \min\left\{\frac{R}{8}, \frac{\log(1/[4(\delta+\alpha+\rho)])}{32\epsilon^2}\right\}. \tag{C.35}$$

*A stronger bound is valid whenever $\alpha + \delta/(1-\rho) < 1/4$:*

$$n \ge \min\left\{\frac{R}{8}, \frac{\log(1/[4(\alpha+\delta/(1-\rho))])}{32\epsilon^2}\right\}. \tag{C.36}$$

*Proof.* Generate $U$ from the original, unconditioned prior $\Pi_z$. In an auxiliary experiment, provide the same $\mathsf{H}$ but always the common central report. For a good oracle the auxiliary and real executions coincide. On $\mathcal{E}_G$, a successful empirical choice belongs to $\mathcal{A}_z$. Decode by these disjoint sets, and choose arbitrarily outside their union. The auxiliary test's error in either world is at most $\alpha + \rho + \delta$. Importantly, its prior has not been conditioned on $\mathcal{G}_z^0$ or on a label-dependent empirical certificate.

The law of $\mathsf{H}$ is the same in both worlds and is independent of the label placement. After $k$ distinct queried labels in a block, of which $s$ are positive, a new identity in block $\sigma$ has conditional probability

$$p_z = \frac{R(1/2 + z\sigma\epsilon) - s}{R - k}. \tag{C.37}$$

For histories shorter than $R/8$, both probabilities are in $[1/4, 3/4]$ and $|p_+ - p_-| \le 16\epsilon/7$. Thus

$$\mathrm{kl}(p_+\|p_-) \le \frac{(p_+ - p_-)^2}{p_-(1-p_-)} \le \frac{4096}{147}\epsilon^2 < 32\epsilon^2. \tag{C.38}$$

Queries outside the unknown blocks and repeated labels add no information. Query choices have the same kernel given a common transcript. Padding early stopping with a common dummy symbol and applying the chain rule gives $\mathrm{KL}(\mathbb{P}_+^{\mathrm{tr}}\|\mathbb{P}_-^{\mathrm{tr}}) \le 32n\epsilon^2$ when $n \le R/8$.

The binary-testing inequality proved in Appendix B bounds the sum of errors below by $\frac{1}{2}e^{-\mathrm{KL}}$. Combining it with the auxiliary error bound proves Eq. (C.35). If $n > R/8$, the first branch of the displayed minimum is immediate.

To obtain the stronger result, let $p_E = \Pr(\mathcal{E}_G) \ge 1 - \rho$ and condition the auxiliary construction history on $\mathcal{E}_G$ only. This is legitimate because the event depends on no utility labels. Its history law remains common to both worlds and the utility prior remains unconditioned. For each fixed good $U$, the unconditional premise implies

$$\Pr(\widehat{\mathcal{R}}_U > t \mid \mathcal{E}_G, U) \le \delta/p_E.$$

Hence each conditional auxiliary test error is at most $\alpha + \delta/p_E \le \alpha + \delta/(1-\rho)$. Conditional on any permitted history, (C.37) and the same transcript KL calculation still hold. Applying the testing inequality proves (C.36). Conditioning here is a proof operation, not a resampling algorithm and not an assumption of conditional success with probability $1 - \delta$. □

Equivalently, for $n \le R/8$, the worst fixed-good-oracle unconditional failure probability of any selector obeys

$$\sup_z \sup_{U \in \mathcal{G}_z^0} \Pr(\widehat{\mathcal{R}}_U > t) \ge (1-\rho)\left[\frac{1}{4}e^{-32n\epsilon^2} - \alpha\right]_+. \tag{C.39}$$

This explicitly separates a construction guarantee from the information needed to choose after construction.

*Sampled reports and numerical calibration.* If the initial report is formed from independent initial-law draws, and the manuscript's reporting-slack condition makes its coupling failure at most $\beta$ uniformly on fixed good oracles, replace $\alpha$ by $\alpha+\beta$ in (C.36). Independence of these draws from $\mathsf{H}$ keeps the bound $\beta$ valid after conditioning on $\mathcal{E}_G$. Without that independence, the conservative additive bound with $\delta+\alpha+\rho+\beta$ follows from the original coupling, provided its marginal failure bound is available. The original audit draws already have an acquisition cost. Unlabeled identities are permitted, not free identity–label pairs.

The theorems calibrate the finite model's true KL exactly. A numerical root returned in the certified interval $[r_-, r_+]$ with $|\mathsf{K}(\widetilde{r}d) - k_0| \le \zeta$ differs from the exact root by at most $\zeta/\nu$. Lemma C.3 adds at most $F\zeta/(2\nu)$ to $q$ in (C.31). It suffices to require $t + 2q + F\zeta/\nu < t_0$. Numerical tolerance must not be silently equated with an exact KL equality.

### C.7 An explicit symmetric realization

The general transfer theorem deliberately controls all bounded utilities. In the balanced, homogeneous construction the fixed-count structure allows a substantially sharper route that avoids bounding each candidate's utility separately. This is an additional specialization, not a replacement for the perturbed-family theorem.

*Model and sampling.* Set $\kappa = 0$, $w_i = 1/(2R)$, and

$$p_* = 1/16, \quad \lambda = \sqrt{\frac{\alpha_{16}(p_*)}{\alpha_{16}(1/2)}}, \quad \alpha_m(p) = \frac{\sum_{k=1}^{m-1}\binom{m}{k}p^k(1-p)^{m-k}\sqrt{k(m-k)}/m}{p(1-p)}. \tag{C.40}$$

The probe features are $\sigma v$ for accepted atoms and their rejected partners. The training prompts have exactly the specified probabilities:

| $x$ | Probabilities | Features | Accepted |
|---|---|---|---|
| 1 | $((1/2+k_1)/2, (1/2-k_1)/2,$ $(1/2-k_1)/2, (1/2+k_1)/2)$ | $(e_1, -e_1, e_1, -e_1)$ | First two |
| 2 | $(p_*, 1/2-p_*, 1/2)$ | $(e_2, e_2, -e_2)$ | First one |

Here $k_1 = p_*\lambda$. All probabilities are positive. The finite coefficient calculation gives $\lambda > 1$, with

$$a_* = \frac{\lambda - 1}{\sqrt{2(\lambda^2+1)}} > 0, \quad d_M = \frac{(\lambda, 1)}{\sqrt{1+\lambda^2}}, \quad d_S = \frac{(1, \lambda)}{\sqrt{1+\lambda^2}}. \tag{C.41}$$

The full empirical menu uses mean and standardized configurations at $m = 4, 8, 16$, with the shared sixteen-response prefix sampling described above and optional base. Fallback directions are fixed before sampling.

For any finite parameter, all atoms within a signed block have equal probability. Consequently, every fixed-count oracle in world $z$ has

$$J_z(P_\theta) = \frac{1}{12} + \frac{z\epsilon}{6}s(\theta), \qquad s(\theta) = \tanh(v^\top\theta). \tag{C.42}$$

The exact initial accepted correctness is $1/2$ for every oracle, so every fixed positive reporting resolution returns the same report. There is no bad-oracle event in this special case, hence $\alpha = 0$.

The KL is

$$\mathsf{K}(\theta) = \frac{1}{3}\left[\log\cosh(v^\top\theta) + \log\cosh\theta_1 + \log\cosh\theta_2\right]. \tag{C.43}$$

Its Hessian is bounded above by $(I_2 + vv^\top)/3 \preceq (2/3)I_2$, which improves the lower radius bound to $r(d) \geq \sqrt{3k_0}$.

*Only two angular witnesses are needed.* Choose $0 < a_0 < a_*$ and set

$$w = (1,1)/\sqrt{2}, \qquad q_0 = \frac{a_0}{\sqrt{1-a_0^2}},$$
$$u_{c,s} = \varsigma_c v + s q_0 w, \qquad \varsigma_M = 1, \quad \varsigma_S = -1, \quad s \in \{-1,1\}. \tag{C.44}$$

Only the two $m = 16$ witness configurations occur in this definition. For their one-group laws, define known scalars

$$\mu_{c,s} = \mathbb{E}[u_{c,s}^\top Y_{c,g}], \quad v_{c,s} = \mathrm{Var}(u_{c,s}^\top Y_{c,g}), \quad B_c = 2M_c\sqrt{1+q_0^2}, \quad M_M = 1/2, \quad M_S = 1, \tag{C.45}$$

and

$$C_{c,s} = \frac{2v_{c,s} + (2/3)B_c\mu_{c,s}}{\mu_{c,s}^2}, \qquad C_* = \max_{c,s} C_{c,s}. \tag{C.46}$$

All four means are positive. Indeed $w^\top d_M = w^\top d_S = \sqrt{1-a_*^2}$, so $\mu_{c,s} = \|g_c\|[a_* + sq_0\sqrt{1-a_*^2}] > 0$.

**Theorem C.3** (Direct finite-rollout separation in the symmetric model)**.** *If $G \geq C_* \log(4/\rho)$, then with construction probability at least $1-\rho$ the empirical mean-16 and std-16 directions satisfy*

$$v^\top \widehat{d}_M > a_0, \qquad v^\top \widehat{d}_S < -a_0. \tag{C.47}$$

*On this event, for every fixed-count oracle and every tolerance*

$$0 < t < \frac{\epsilon}{6}\tanh(a_0\sqrt{3k_0}), \tag{C.48}$$

*the complete empirical menu has disjoint sets of $t$-acceptable indices in the two utility worlds. This statement includes all intermediate configurations, their possible empirical duplicates, and base.*

*Proof.* The centered scalar $u_{c,s}^\top Y_{c,g} - \mu_{c,s}$ has absolute value at most $B_c$, by Lemma C.1. The one-sided Bernstein inequality gives

$$\Pr(u_{c,s}^\top \widehat{g}_c \leq 0) \leq \exp\left[-\frac{G\mu_{c,s}^2}{2v_{c,s} + (2/3)B_c\mu_{c,s}}\right].$$

A union bound over the four scalar inequalities is at most $4e^{-G/C_*} \leq \rho$. On the complementary event, $\varsigma_c v^\top \widehat{g}_c > q_0 |w^\top \widehat{g}_c|$. In the orthonormal basis $(v, w)$ this implies $\varsigma_c v^\top \widehat{d}_c > q_0/\sqrt{1+q_0^2} = a_0$. It also rules out a zero gradient for the two witnesses.

For clarity, the Bernstein inequality follows by bounding the moment series of a mean-zero scalar $X$ with $|X| \leq B$ and variance $v$. Since $\mathbb{E}|X|^k \leq vB^{k-2}$ and $k! \geq 2\,3^{k-2}$ for $k \geq 2$,

$$\mathbb{E}e^{hX} \leq \exp\left(\frac{vh^2}{2(1-hB/3)}\right), \qquad 0 < hB < 3.$$

Apply this to $-X$, multiply mgfs over groups, and use $h = \mu/(v + B\mu/3)$, with the zero-variance case understood directly.

By the radius bound, the two witness block coordinates exceed $s_0 = \tanh(a_0\sqrt{3k_0})$ and are below $-s_0$, respectively. Let $s_{\max}, s_{\min}$ be the extreme coordinates of the entire actual menu. They obey $s_{\max} - s_{\min} > 2s_0$. From (C.42), an index $c$ has world regrets

$$\widehat{\mathcal{R}}_+(c) = \frac{\epsilon}{6}(s_{\max} - s_c), \qquad \widehat{\mathcal{R}}_-(c) = \frac{\epsilon}{6}(s_c - s_{\min}).$$

Their sum is $\epsilon(s_{\max} - s_{\min})/6 > 2t$, so no index can have both regrets at most $t$. A successful index can be decoded by whether $s_c$ is above or below $(s_{\max} + s_{\min})/2$, a threshold known from $\mathsf{H}$. No closeness assumption is needed for the other candidates. Adding an intermediate or more extreme candidate cannot remove this particular full-menu separation. □

*Exact variances and a numerical certificate.* The constants in (C.45) do not require simulation. For a fixed prompt, put $p = \mathbb{E}V$, let $h_{x,1}, h_{x,0}$ be conditional feature means given $V = 1, 0$, and let $\Sigma_1, \Sigma_0$ be the corresponding covariances. Write $D_x = h_{x,1} - h_{x,0}$. Given $k$ accepted responses in a group,

$$\mathbb{E}[Y \mid x, k] = f_c(k) D_x, \tag{C.49}$$

$$\mathbb{E}[YY^\top \mid x, k] = f_c(k)^2 \left[ D_x D_x^\top + \frac{\Sigma_1}{k} + \frac{\Sigma_0}{m-k} \right], \tag{C.50}$$

$$f_M(k) = \frac{k(m-k)}{m^2}, \qquad f_S(k) = \frac{\sqrt{k(m-k)}}{m}, \qquad 1 \le k \le m-1. \tag{C.51}$$

Both moments are zero for $k = 0, m$. These formulas follow from $Y = f_c(k)(\overline{\phi}_+ - \overline{\phi}_-)$ and conditional independence of the two samples. Average over $\mathrm{Bin}(m, p)$ and the three prompts, then subtract $g_c g_c^\top$ to obtain the group covariance. Thus $C_*$ is a finite expression involving rational numbers and square roots.

For $a_0 = 3/50$, the supplied rational interval computation gives $\lambda \in (1.27569, 1.27571)$, $a_* \in (0.12026, 0.12028)$ and

| Witness | Cone sign | Mean, approximate | $C_{c,s}$, enclosing interval |
|---|---|---|---|
| mean-16 | $-$ | 0.00191846 | $[5270.90, 5270.92]$ |
| mean-16 | $+$ | 0.00569674 | $[706.45, 706.48]$ |
| std-16 | $-$ | 0.00504910 | $[3909.04, 3909.09]$ |
| std-16 | $+$ | 0.01499297 | $[494.08, 494.12]$ |

The intervals use integer square roots and exact fraction operations, not rounded floating-point evaluations as inequalities. In particular $C_* < 5300$.

**Corollary C.2** (An explicit finite budget)**.** *Take $k_0 = 1/100$, $\epsilon = 1/16$, $a_0 = 3/50$, and any positive block size $R$ divisible by 16. With $G = 24000$ shared sixteen-response groups, the actual six-name menu, with optional base, has disjoint $t$-acceptable sets at $t = 1/10000$ with probability greater than $0.95$ over rollout construction. Every nonbase candidate is calibrated to KL $0.01$. The raw construction count is $384000$ responses.*

*Proof.* Since $C_* < 5300$ and $24000/5300 > 22/5 > \log 80$, the four-event Bernstein bound is below $1/20$. The strict bound $e^{22/5} > 80$ follows already from the first 26 terms of its positive exponential series, and is checked using exact fractions.

For the tolerance, $\sqrt{3k_0} > 17/100$, so $a_0\sqrt{3k_0} > 51/5000$. For $x \ge 0$, $\tanh x \ge x - x^3/3$, because the derivative of the difference is $x^2 - \tanh^2 x \ge 0$ and both sides agree at zero. Hence $\tanh(a_0\sqrt{3k_0}) > 1/100$. It follows that

$$\frac{\epsilon}{6} \tanh(a_0\sqrt{3k_0}) > \frac{1}{9600} > \frac{1}{10000}.$$

Integer label counts are $9R/16$ and $7R/16$. The construction and all inequalities are independent of which identities receive these labels. □

Here $t = 10^{-4}$ is a probability difference, or $0.01$ percentage points. It is not a one-percent training gain. The budget is a sufficient finite-model certificate, not an empirically necessary threshold or a recommended LLM rollout count. The $G$ condition in Theorem C.3 is independent of $R$, $\epsilon$, and $k_0$ when $a_0$ and $\rho$ are fixed. These parameters still control integrality, the permitted regret tolerance, and the label complexity. The general perturbation-transfer budget need not share this special cancellation.

*Implications for decisions using only reliability metrics.* In the symmetric model, $J_z$ depends only on $z$ and $s_c$, not on label locations within a block. The history law and correctness report remain the same in the two worlds. Conditional on the angular event, the successful index sets are disjoint. Thus a selector reading no trusted labels has the same output distribution in the two worlds, and the sum of its two failure probabilities on that event is at least one. Unconditionally,

$$\sup_z \sup_{U \in \mathrm{supp}(\Pi_z)} \Pr(\widehat{\mathcal{R}}_U > t) \ge \frac{1-\rho}{2}. \tag{C.52}$$

For Corollary C.2, the lower bound is $0.475$. This is a statement about decisions restricted to the stated metric and common history. Physical deletion of the original audit records is not required for the metric-insufficiency interpretation. If raw labels are used instead, the information channel has changed.

For an $n$-disclosure selector, repeat the proof of Theorem C.2 with $\alpha = 0$ and the midpoint decoder from Theorem C.3. This decoder may depend on $\mathsf{H}$, because it remains a test based only on the permitted transcript. Conditioning on the label-free angular event leaves the same fixed-count prior and incurs $\delta/\Pr(\mathcal{E})$ in place of the unconditional selection error. The resulting lower bound is

$$n \geq \min\left\{\frac{R}{8}, \frac{\log((1-\rho)/(4\delta))}{32\epsilon^2}\right\}, \qquad \delta < (1-\rho)/4. \tag{C.53}$$

The symbolic expression is primary. Its constants do not establish a tight minimum audit budget.

A constructive comparison is also possible. Draw a block uniformly and an identity uniformly within it, with replacement. Every fixed oracle gives $\mathbb{E}[\sigma(U - 1/2)] = z\epsilon$. A sign test has error at most $e^{-2n\epsilon^2}$. On a correct sign, selecting the actual menu's largest $s_c$ for $z = +$ or smallest $s_c$ for $z = -$ is exactly optimal, for every construction history, including histories outside the angular event. Therefore $n \geq \log(1/\delta)/(2\epsilon^2)$ draws suffice for zero-regret selection with probability $1 - \delta$. This upper bound applies to this fixed-count family and is also available to a direct method retaining the same raw labels. It gives no exclusive cost advantage to a compressed summary.

### C.8 Variance bounds for perturbed constructions

The preceding symmetric certificate used the exact variances of four projected group gradients. Those variances are also computable for an unequal or non-paired finite model. This section first replaces the range-only concentration step in Theorem C.1, then combines the variance calculation with Lemma D.1 to obtain an explicit nonzero perturbation example. No additional trusted labels enter these construction bounds.

For prompt $x$, put $p_x = \Pr_0(V = 1 \mid x)$, let $h_{x,1}, h_{x,0}$ and $\Sigma_{x,1}, \Sigma_{x,0}$ be the feature means and covariances conditional on acceptance and rejection, and set $D_x = h_{x,1} - h_{x,0}$. Prompts with $p_x \in \{0, 1\}$ contribute zero group gradient. For other prompts, let $p_{x,k} = \binom{m}{k} p_x^k (1-p_x)^{m-k}$. The conditional calculation in Eqs. (C.49)–(C.50) does not require pairing or equal response weights. It gives

$$\begin{aligned} g_c &= \sum_x p_X(x) \sum_{k=1}^{m_c-1} p_{x,k} f_c(k) D_x, \\ \Sigma_c &= \sum_x p_X(x) \sum_{k=1}^{m_c-1} p_{x,k} f_c(k)^2 \left[D_x D_x^\top + \frac{\Sigma_{x,1}}{k} + \frac{\Sigma_{x,0}}{m_c - k}\right] - g_c g_c^\top. \end{aligned} \tag{C.54}$$

The final subtraction includes variability between prompts, not just conditional response variance. In particular, a probe with zero expected gradient can have a nonzero contribution to $\Sigma_c$.

Write $\sigma_{c,j}^2 = (\Sigma_c)_{jj}$ and let $M_c = F/2$ for mean centering and $M_c = F$ for standardization. A valid centered coordinate bound is $B_{c,j} = M_c + |g_{c,j}|$. All of these quantities depend only on the known initial policy, verifier, and features.

**Corollary C.3** (Variance-sensitive construction budget)**.** *For the sampling in Eq.* (C.14)*,*

$$\Pr\Big(\max_c \|\widehat{g}_c - g_c\| > \eta\Big) \leq 2\sum_{c=1}^{K}\sum_{j=1}^{d} \exp\left[-\frac{G\eta^2}{2d\sigma_{c,j}^2 + (2/3)B_{c,j}\eta\sqrt{d}}\right]. \tag{C.55}$$

*Consequently, the conclusion of Theorem C.1 holds when its condition on $\eta$ is kept and its group budget is replaced by*

$$G \geq \max_{c,j}\left[\frac{2d\sigma_{c,j}^2}{\eta^2} + \frac{2B_{c,j}\sqrt{d}}{3\eta}\right]\log\frac{2Kd}{\rho}. \tag{C.56}$$

*Proof.* Conditional on $x, k$, the accepted and rejected feature averages are independent and have covariances $\Sigma_{x,1}/k$ and $\Sigma_{x,0}/(m_c - k)$. Since $Y = f_c(k)(\overline{\phi}_1 - \overline{\phi}_0)$, averaging their first and second moments proves Eq. (C.54). For coordinate $j$, the centered variable has variance $\sigma_{c,j}^2$ and absolute value at most $B_{c,j}$. Apply the two-sided Bernstein inequality at threshold $\eta/\sqrt{d}$ and take a union bound. A zero variance coordinate is deterministic and can be omitted. Taking every summand at most $\rho/(Kd)$ proves the displayed budget. The remainder of Theorem C.1's proof uses only this common gradient event, so no change to its utility class or calibration is needed. □

The Hoeffding and Bernstein bounds can both be computed in advance, and the smaller sufficient budget used. Estimated variances cannot simply replace $\sigma_{c,j}^2$ without a separate valid upper confidence bound. For illustration, in Eq. (C.57) with $a = 0.005$, $\omega = 0.1$, and $\kappa = 0.001$, the budgets for $\eta = 0.001$, $\rho = 0.05$, and all six two-dimensional gradients are 24,695,170 groups from the common Hoeffding bound and 2,772,728 from Eq. (C.56). These deterministic sufficient counts control full gradient norms. They are not measured sample requirements and are not comparable to the smaller, decision-specific cone budget below.

*A paired perturbation with unequal weights.* Let $w = (1, 1)/\sqrt{2}$, divide each block into two classes of $R/2$ atoms indexed by $\zeta = \pm 1$, and assign every atom in class $(\sigma, \zeta)$

$$w_i = \frac{(1 + \sigma a)(1 + \zeta\omega)}{2R}, \qquad \phi_i = \sigma v + \kappa\zeta w. \tag{C.57}$$

Copy each atom to its rejected partner, and keep the two training prompts unchanged. The probabilities are unequal, the initial block masses are unbalanced, and each block has nonconstant features. The construction has $F = \sqrt{1 + \kappa^2}$ and $I_0 \succeq I_2/3$. The utility prior still places the prescribed number of positive labels uniformly across all $R$ identities in each block, not separately within $\zeta$ classes.

Fix the following parameters before drawing labels or rollouts:

$$R = 4096, \quad a = \frac{1}{200}, \quad \omega = \frac{1}{10}, \quad \kappa = \frac{1}{1000}, \quad \epsilon = \frac{1}{16}, \quad k_0 = \tau = \alpha = \frac{1}{100}. \tag{C.58}$$

Here $\tau$ is the reliability bin half-width and $\alpha$ is the prior exception probability. Define $r_b = \omega\sqrt{2\log(4/\alpha)/R}$ and fix, before rollouts,

$$\mathcal{G}_z^0 = \{U : |\mu_\sigma(U) - (1/2 + z\sigma\epsilon)| \le r_b \text{ for both } \sigma\}. \tag{C.59}$$

**Corollary C.4** (Explicit realization away from equal weights)**.** *For Eqs.* (C.57)–(C.59)*,* $\Pi_z(\mathcal{G}_z^0) \ge 0.99$*, and both good sets share the central reliability report. With* $G = 24000$ *shared sixteen-response groups, a single construction event of probability greater than* $0.95$ *makes the complete empirical menu have disjoint acceptable index sets across these good worlds at* $t = 1/25000$*, or* $0.004$ *pp. The statement is simultaneous over all oracles in the fixed good sets.*

*Proof. The good utility sets.* Within block $\sigma$, the conditional weights are $(1 + \omega\zeta_i)/R$, with range $2\omega/R$. Applying the without-replacement concentration argument of Appendix B to the fixed positive count gives a two-sided failure probability at most $\alpha/2$ at radius $r_b$ for each block. The union has probability at most $\alpha$. Since $\log 400 < 6$, $r_b < 0.0055$, so every good utility has

$$zb(U) \ge \epsilon - r_b > 0.057, \qquad |\mu(U) - 1/2| \le \epsilon|a| + r_b < \tau. \tag{C.60}$$

Thus the oracle class and common report are fixed independently of sampling.

*The four empirical directions.* Accepted and rejected feature laws at the probe remain identical, so its expected group gradient is zero. The other gradients are unchanged. The probe feature covariance is

$$(1 - a^2)vv^\top + \kappa^2(1 - \omega^2)ww^\top. \tag{C.61}$$

For each of the four cone functionals $u_{c,s} = \varsigma_c v + sq_0 w$ with $a_0 = 3/50$, the change in probe projection variance is proportional to

$$-a^2 + \kappa^2(1 - \omega^2)q_0^2 < 0. \tag{C.62}$$

The positive proportionality factor, already including prompt mass, is $(m - 1)/(12m^2)$ for centering and $(1 - 2^{1-m})/(3m)$ for standardization. Hence all four projected variances are no larger than in

the symmetric example, and their means are unchanged. The range constants increase by a factor $F < 1 + 10^{-6}$. The certified symmetric values have $C_{c,s} < 5271$, $\mu_{c,s} > 0.0019$, and the old centered range bound below $2.01$. Therefore the increase of any Bernstein coefficient is less than

$$\frac{2}{3}\frac{2.01 \times 10^{-6}}{0.0019} = \frac{67}{95000} < 0.001, \tag{C.63}$$

so $C_* < 5300$ still holds. The same $G = 24000$ gives the angular event in Eq. (C.47) with probability greater than $0.95$.

*Uniform radii and conditional change.* The two signs in the probe factorize under every tilt. Its covariance is bounded above by $vv^\top + \kappa^2 ww^\top$, while the training prompts contribute at most $I_2$. Thus $\nabla^2 \mathsf{K} \preceq (2/3)I_2$ and every calibrated radius exceeds $\sqrt{3k_0} > 0.173$. Conversely, the coercive bound in Eq. (C.25) at $H = 1/3$ exceeds $k_0$, using $\lambda_{\mathrm{F}} = 1/3$ and $F < 1 + 10^{-6}$. Every menu candidate therefore lies in the ball of radius $H$. Lemma D.1 gives the utility remainder bound $|e(\theta)| \le E_H = \tanh(H\kappa/2)/6$ for all these candidates.

The block log-odds argument in Appendix A gives witness coordinates at least $\tanh(\operatorname{arctanh} a + u)$ and at most $\tanh(\operatorname{arctanh} a - u)$, where $u = (a_0 - \kappa)\sqrt{3k_0}$. In particular the whole menu has span

$$S > \frac{2(1-a^2)\tanh u}{1 - a^2 \tanh^2 u} > 0.02041. \tag{C.64}$$

The last inequality follows from $\sqrt{0.03} > 0.173$ and $\tanh x \ge x - x^3/3$. For a candidate below the midpoint of the menu's extreme $s$ values, its regret in every good positive world is at least $0.057S/12 - 2E_H$. For a candidate above that midpoint, the same lower bound holds in every good negative world. At equality both bounds apply. Finally,

$$\frac{0.057S}{12} - 2E_H > \frac{0.057 \times 0.02041}{12} - \frac{H\kappa}{6} = \frac{149011}{3600000000} > \frac{1}{25000}. \tag{C.65}$$

No candidate can consequently be $t$-acceptable in both good worlds. The construction event does not involve $U$, so this argument applies to all fixed good oracles on the same event. □

The joint probability when an oracle is also drawn from its original prior is greater than $0.99 \times 0.95 = 0.9405$. The $0.95$ guarantee is the construction probability for the fixed good classes, not a joint $0.95$ claim over arbitrary label draws. The disclosure argument of Theorem C.2 applies with $\alpha = 0.01$ and the history-dependent midpoint decoder. General non-paired perturbations still use the transfer theorem and their actual margins, not this numerical budget.

This example makes the constant refinement operational for a sufficient certificate. At the same $H, S$, and $b$ lower bounds, the earlier remainder $\tanh(H\kappa)/6$ gives no positive tolerance through the corresponding radius-based calculation. The sharper normalized-tilt bound makes Eq. (C.65) positive. That comparison is between sufficient calculations, not a claim that the original model lacked any other successful certificate. The attached arithmetic check verifies the displayed rational inequalities. It does not sample 24,000 groups or change any historical perturbation experiment.

# D Proofs for directional reuse and supporting identities

This appendix proves the reuse statements in Section 5, followed by the general characterization in Section 3. The paired-model proofs and their refinements appear first. Sections D.7–D.10 give the broader moment criterion, its limits for binary utilities, and conditions for reuse without pairing. Each extension states its own utility class and structural assumptions.

**Lemma D.1** (Total variation of a normalized exponential tilt)**.** *Let $Q$ be a probability law, let $f$ have finite oscillation $L = \operatorname{ess\,sup}_Q f - \operatorname{ess\,inf}_Q f$, and set $dP = e^f\, dQ/\mathbb{E}_Q e^f$. With $\mathrm{TV}(P,Q) = \frac{1}{2}\int |dP - dQ|$,*

$$\mathrm{TV}(P,Q) \le \tanh(L/4). \tag{D.1}$$

*The bound is attainable by two-point laws.*

This is the sharp comparison with the Hilbert distance, in the half-$L^1$ convention (Cohen & Fausti, 2024, Theorem 5.1). We give an elementary proof for the normalized tilt used here, and use it to refine the earlier range-based estimate rather than as a separate novelty claim.

*Proof.* Write $a = \operatorname{ess\,inf} e^f$, $b = \operatorname{ess\,sup} e^f$, and $z = \mathbb{E}_Q e^f$. When $a = b$ the laws coincide. Otherwise convexity of $x \mapsto |x - z|$ and the chord between $a$ and $b$ give

$$\mathrm{TV}(P, Q) = \frac{\mathbb{E}_Q|e^f - z|}{2z} \le \frac{(b - z)(z - a)}{z(b - a)} = \frac{a + b - z - ab/z}{b - a}. \tag{D.2}$$

Since $z + ab/z \ge 2\sqrt{ab}$, the last expression is at most $(\sqrt{b} - \sqrt{a})/(\sqrt{b} + \sqrt{a}) = \tanh(L/4)$. Equality holds when $e^f$ takes only $a, b$ and $Q(e^f = b) = \sqrt{a}/(\sqrt{a} + \sqrt{b})$, which makes $z = \sqrt{ab}$. Thus sharpness concerns a bound based only on oscillation, not the regret bound or every fixed policy instance. □

### D.1 Proof of Theorem 3

*Proof. Step 1: the audit estimates the initial block contrast.* Let $Q_\sigma$ be the initial conditional law inside $I_\sigma$, and let $\mu_\sigma = \mathbb{E}_{Q_\sigma} U$. Select $\sigma$ uniformly, then draw under $Q_\sigma$. For $Z = \sigma(U - 1/2)$,

$$\mathbb{E}Z = \frac{1}{2}(\mu_+ - 1/2) - \frac{1}{2}(\mu_- - 1/2) = \frac{\mu_+ - \mu_-}{2} = b. \tag{D.3}$$

Because $Z \in [-1/2, 1/2]$, $n$ iid draws give $\Pr(|\widehat{b} - b| > e_n) \le 2e^{-2ne_n^2} = \delta$ for the $e_n$ in Eq. (10). This event concerns one scalar under a fixed sampling law. No candidate is involved in defining it, and caching repeated labels changes distinct acquisitions rather than the iid draw count.

*Step 2: control each block's conditional change uniformly.* Let $Q_{\sigma,\theta}$ denote the conditional law at $\theta$. The factor $e^{\sigma\theta^\top v}$ cancels inside a block, leaving

$$\frac{Q_{\sigma,\theta}(i)}{Q_\sigma(i)} = \frac{e^{\theta^\top \xi_i}}{\mathbb{E}_{Q_\sigma} e^{\theta^\top \xi}}. \tag{D.4}$$

Apply Lemma D.1 to $f_i = \theta^\top \xi_i$. Its oscillation is at most $2H\kappa$, hence, simultaneously in the ball,

$$\mathrm{TV}(Q_{\sigma,\theta}, Q_\sigma) \le \tanh(H\kappa/2). \tag{D.5}$$

The earlier density-ratio range argument gives the valid but weaker $\tanh(H\kappa)$. Keeping the normalization structure gives the factor $1/2$ in the argument above. Because $0 \le U \le 1$, $|\mu_\sigma(\theta) - \mu_\sigma| \le \mathrm{TV}(Q_{\sigma,\theta}, Q_\sigma)$.

*Step 3: expose the one-moment utility formula.* Pairing preserves Eq. (A.3), and $U$ vanishes outside the accepted probe. Hence

$$J_U(P_\theta) = \frac{1}{6}\left[\frac{1 + s(\theta)}{2}\mu_+(\theta) + \frac{1 - s(\theta)}{2}\mu_-(\theta)\right]. \tag{D.6}$$

Replacing the two current means by their initial means is a convex combination of errors bounded by Eq. (D.5). The resulting error $e(\theta)$ obeys $|e(\theta)| \le \tanh(H\kappa/2)/6$. By the definitions of $\mu$ and $b$,

$$\frac{1 + s}{2}\mu_+ + \frac{1 - s}{2}\mu_- = \frac{1 + a}{2}\mu_+ + \frac{1 - a}{2}\mu_- + \frac{s - a}{2}(\mu_+ - \mu_-) \tag{D.7}$$

$$= \mu + (s - a)b. \tag{D.8}$$

This proves Eq. (11). The offset $\mu/6 - ab/6$ is common to all candidates, so it need not be estimated for ranking them.

*Step 4: translate scalar estimation error into regret.* Fix any nonempty finite candidate set in the ball, let $\theta^*$ be a true utility maximizer, and choose $\widehat{\theta}$ to maximize $\widehat{b}\, s(\theta)$. Write $D = s(\theta^*) - s(\widehat{\theta})$. The selection rule, including any tie rule, implies $\widehat{b}D \le 0$. Therefore, on $|\widehat{b} - b| \le e_n$,

$$J_U(P_{\theta^*}) - J_U(P_{\widehat{\theta}}) = \frac{bD}{6} + e(\theta^*) - e(\widehat{\theta}) \tag{D.9}$$

$$\le \frac{(b - \widehat{b})D}{6} + \frac{1}{3}\tanh(H\kappa/2) \tag{D.10}$$

$$\le \frac{e_n}{6}\left[\max_{\theta \in \mathcal{C}} s(\theta) - \min_{\theta \in \mathcal{C}} s(\theta)\right] + \frac{1}{3}\tanh(H\kappa/2). \tag{D.11}$$

The span coefficient is obtained directly from the difference of two candidates, rather than paying twice a separate worst-case error for each one.

*Step 5: justify uniformity and state its limit.* Steps 2–3 are deterministic and hold for every $\theta \in \Theta_H$. Step 1 supplies the event $E_n = \{|\widehat{b} - b| \le e_n\}$ with probability at least $1 - \delta$. On $E_n$, Step 4 holds for every finite nonempty $\mathcal{C} \subseteq \Theta_H$ and every $\widehat{\theta}$ in the prescribed argmax, including all ties. This proves the simultaneous quantifiers in Theorem 3 and permits a menu constructed after observing $\widehat{b}$. No candidate union bound or independence from the chosen menu is needed.

At $\kappa = 0$, conditional block laws are unchanged and the deterministic error vanishes. If the oracle, features, pairing, or block definitions change, the approximation must be established again. The event is uniform over a fixed parameter region, not arbitrary future training trajectories. In the symmetric pair of norm $h$, the span is $2\tanh(ha_*)$ and the estimation term is $e_n \tanh(ha_*)/3$. The numerical interpretation of this term depends on the decision gap, as quantified in Appendix D.6. □

### D.2 Proof of Proposition 1

*Proof. The old observations are exactly the same.* Split each accepted signed block into two equal subblocks indexed by $\zeta = \pm 1$, with all four initial masses equal. Copy the feature of every accepted atom to its rejected partner and use Eq. (13). For either $z$, half of each initial block has utility one, giving $\mu_+ = \mu_- = 1/2$ and $b = 0$. The two worlds also share all weights, features, verifier outputs, and label-free construction information.

*The GRPO candidates and their KL remain explicit.* The probe still has pass probability $1/2$ at every parameter, hence zero expected verifier gradient. The two training prompts and their gradients are unchanged, so the directions remain $d_M, d_S$. Put $t = v^\top \theta$. Summing over the four equally weighted signs gives the probe partition

$$\frac{1}{4} \sum_{\sigma,\zeta \in \{-1,1\}} e^{(\sigma + \kappa\zeta)t} = \cosh t \cosh(\kappa t). \tag{D.12}$$

Its initial feature mean is zero. Together with the training prompts this gives

$$F_\kappa(\theta) = \frac{1}{3}\left[\log\cosh(v^\top\theta) + \log\cosh(\kappa v^\top \theta) + \log\cosh\theta_1 + \log\cosh\theta_2\right]. \tag{D.13}$$

Swapping coordinates reverses $v^\top\theta$ and exchanges the last two terms. Since $\cosh$ is even, $F_\kappa$ is invariant under this swap. The directions $d_M, d_S$ are themselves exchanged, and uniqueness of positive radial calibration therefore gives the same norm $r$ for both candidates at a common KL target.

*Compute the actual utility gap.* Within the accepted probe the tilted signs factorize. In particular,

$$\mathbb{E}_\theta \zeta = \frac{e^{\kappa t} - e^{-\kappa t}}{e^{\kappa t} + e^{-\kappa t}} = \tanh(\kappa t). \tag{D.14}$$

The accepted probe has joint mass $1/6$, and $U_z = (1 + z\zeta)/2$, so

$$J_z(P_\theta) = \frac{1}{12} + \frac{z}{12}\tanh(\kappa v^\top \theta), \qquad J_z(P_{rd_M}) - J_z(P_{rd_S}) = \frac{z}{6}\tanh(\kappa r a_*). \tag{D.15}$$

This is an exact finite-policy calculation, not a first-order expansion.

*Convert indistinguishability into expected regret.* Let $p$ be the probability that a selector of this pair chooses the mean candidate. Because its permitted observations have the same law in both worlds, the same $p$ applies to both. The wrong-choice probabilities are $1 - p$ in world $+1$ and $p$ in world $-1$. Consequently

$$\max_z \mathbb{E}_z \mathcal{R} = \frac{1}{6}\tanh(\kappa r a_*)\max\{p, 1-p\} \ge \frac{1}{12}\tanh(\kappa r a_*). \tag{D.16}$$

Every randomized rule $\chi(O_z)$ induces such a common $p$. Taking the infimum over these rules proves the minimax statement in Eq. (14). For $\kappa = 0$ the lower bound is zero, as it should be. For fixed $a_* > 0$ and small $\kappa r$, $\tanh(\kappa r a_*) = \kappa r a_* + O((\kappa r)^3)$. The floor is therefore linear in $\kappa r$, as is the exact-moment heterogeneity term of Theorem 3 at $H = r$. This matches that term's order, not its optimal constant or every term in a minimax rate. A selector that receives raw labels and computes a new $\zeta$-aligned statistic has a different information channel and is not covered by this impossibility statement. □

### D.3 SHARED-LABEL SUMMARIES AND THE GROUP-SURROGATE IDENTITY

*Two summaries of the same sampled labels.* In the equal-within-block illustration, initial block probabilities are $(1 \pm a)/2$, while the fixed positive fraction in block $\sigma$ is $1/2 + z\sigma\epsilon$. Thus a draw from the initial law has

$$\Pr_z(U = 1) = \frac{1+a}{2}\left(\frac{1}{2} + z\epsilon\right) + \frac{1-a}{2}\left(\frac{1}{2} - z\epsilon\right) = \frac{1}{2} + z\epsilon a. \tag{D.17}$$

For the signed observation $Z = \sigma(U - 1/2)$, the shifted variable $Z + 1/2$ is binary. Its success probability is

$$\Pr_z(Z + 1/2 = 1) = \Pr_z(\sigma = 1, U = 1) + \Pr_z(\sigma = -1, U = 0) = \frac{1}{2} + z\epsilon. \tag{D.18}$$

The coarse and signed sample means therefore distinguish Bernoulli parameters separated by $2\epsilon a$ and $2\epsilon$, respectively. Since these probabilities stay away from zero and one, their per-draw testing information has order $\epsilon^2 a^2$ and $\epsilon^2$. For nonzero $a$, the resulting sample requirements differ by order $a^{-2}$. At $a = 0$, the coarse scalar has exactly the same law in both worlds.

This comparison releases only the respective final scalar and the draw count. In particular, the coarse channel does not also receive realized block counts or raw response–label pairs. A method retaining those pairs can compute the signed statistic at the same acquisition cost. The experiment compares retained information, not an advantage over such a direct method.

*Deriving the fixed-sample surrogate identity.* For a group of size $m_g$, let $\ell_i(\theta)$ be the implemented per-response log-probability score, possibly divided by answer length. Hold the rewards and computed advantages fixed during differentiation. The surrogate and gradient are

$$L_g(\theta \mid \mathbf{r}) = \frac{1}{m_g}\sum_i (C_g\mathbf{r})_i \ell_i(\theta), \qquad g_g(\mathbf{r}) = \frac{1}{m_g}\Psi_g^\top C_g\mathbf{r}, \tag{D.19}$$

where row $i$ of $\Psi_g$ is $\psi_i^\top = \nabla\ell_i(\theta)^\top$. Subtracting the gradients for $\widehat{\mathbf{r}}$ and $\mathbf{u}$ gives Eq. (3) exactly. To obtain the norm bound, set $e = C_g(\widehat{\mathbf{r}} - \mathbf{u})$. Its coordinates sum to zero, so

$$\widehat{g}_g - g_{U,g} = \frac{1}{m_g}\sum_i e_i(\psi_i - \overline{\psi}), \qquad \overline{\psi} = \frac{1}{m_g}\sum_i \psi_i. \tag{D.20}$$

Vector Cauchy–Schwarz now yields

$$\|\widehat{g}_g - g_{U,g}\| \le \sqrt{\frac{1}{m_g}\|C_g(\widehat{\mathbf{r}} - \mathbf{u})\|^2}\sqrt{\frac{1}{m_g}\sum_i \|\psi_i - \overline{\psi}\|^2}. \tag{D.21}$$

The first factor measures centered reward error, and the second records score geometry. If a predictor is constant within a prompt, centering removes it regardless of its cross-prompt accuracy. Conversely, lowering the first factor alone need not improve the direction of a particular finite update. Answer-length normalization must be included in $\psi_i$. Reward standardization, clipping, optimizer transformations, and later response sampling require their own analysis and are not silently included in this mean-centered fixed-sample identity.

### D.4 COMPARING THE REMAINDERS OF TWO CANDIDATES

This refinement of Theorem 3 bounds a difference of utilities directly. It uses the same fixed oracle, blocks, prompt masses, and accepted/rejected pairing. Let $Q_{\sigma,\theta}$ be the conditional law inside block $I_\sigma$, put $q_\sigma(\theta) = (1 + \sigma s(\theta))/2$, and define

$$T_\theta = \max_{\sigma\in\{-1,1\}} \mathrm{TV}(Q_{\sigma,\theta}, Q_{\sigma,0}), \qquad D_{\theta,\vartheta} = \max_{\sigma\in\{-1,1\}} \mathrm{TV}(Q_{\sigma,\theta}, Q_{\sigma,\vartheta}). \tag{D.22}$$

These quantities depend on the policies and the specified partition, not on the unknown utility labels. For parameters in the prescribed ball, set

$$\mathcal{E}(\theta,\vartheta) = \frac{1}{6}\min\{T_\theta + T_\vartheta,\ D_{\theta,\vartheta} + |s(\theta) - s(\vartheta)|\min(T_\theta, T_\vartheta)\}. \tag{D.23}$$

The two alternatives respectively bound separate remainders and exploit cancellation between them.

**Theorem D.1** (Comparison of candidate utilities)**.** *In the paired setting of Theorem 3, every pair in the parameter ball satisfies*

$$\left| J_U(P_\theta) - J_U(P_\vartheta) - \frac{b}{6}[s(\theta) - s(\vartheta)] \right| \le \mathcal{E}(\theta, \vartheta). \tag{D.24}$$

*On $|\widehat{b} - b| \le e_n$, simultaneously for all nonempty finite $\mathcal{C}$ in that ball, its specified selection rule obeys*

$$\mathcal{R}_{\mathcal{C}} \le \frac{e_n}{6}|s(\theta^*) - s(\widehat{\theta})| + \mathcal{E}(\theta^*, \widehat{\theta}), \qquad \theta^* \in \arg\max_{\theta \in \mathcal{C}} J_U(P_\theta). \tag{D.25}$$

*Proof.* Write $\delta_\sigma(\theta) = \mathbb{E}_{Q_{\sigma,\theta}} U - \mathbb{E}_{Q_{\sigma,0}} U$. The exact decomposition in Eq. (11) has remainder

$$e(\theta) = \frac{1}{6} \sum_{\sigma \in \{-1,1\}} q_\sigma(\theta)\delta_\sigma(\theta). \tag{D.26}$$

Because $0 \le U \le 1$, total variation controls a difference of conditional expectations. In particular, $|\delta_\sigma(\theta)| \le T_\theta$ and $|\delta_\sigma(\theta) - \delta_\sigma(\vartheta)| \le D_{\theta,\vartheta}$. Subtract the two remainders by adding and subtracting the same conditional means:

$$\begin{aligned} 6[e(\theta) - e(\vartheta)] &= \sum_\sigma q_\sigma(\theta)[\delta_\sigma(\theta) - \delta_\sigma(\vartheta)] \\ &\quad + \sum_\sigma [q_\sigma(\theta) - q_\sigma(\vartheta)]\delta_\sigma(\vartheta). \end{aligned} \tag{D.27}$$

The first sum has absolute value at most $D_{\theta,\vartheta}$ since the $q_\sigma(\theta)$ form a probability vector. Writing $\Delta s = s(\theta) - s(\vartheta)$, the second sum is $\Delta s[\delta_+(\vartheta) - \delta_-(\vartheta)]/2$, of absolute value at most $|\Delta s| T_\vartheta$. Interchanging the two parameters gives the alternative $|\Delta s| T_\theta$. Separately, Eq. (D.26) gives $6|e(\theta) - e(\vartheta)| \le T_\theta + T_\vartheta$. Taking the smaller of these valid bounds proves Eq. (D.24).

For selection, $\widehat{b}[s(\theta^*) - s(\widehat{\theta})] \le 0$. Hence, on the stated event,

$$\begin{aligned} \mathcal{R}_{\mathcal{C}} &= \frac{b}{6}[s(\theta^*) - s(\widehat{\theta})] + e(\theta^*) - e(\widehat{\theta}) \\ &\le \frac{b - \widehat{b}}{6}[s(\theta^*) - s(\widehat{\theta})] + \mathcal{E}(\theta^*, \widehat{\theta}) \\ &\le \frac{e_n}{6}|s(\theta^*) - s(\widehat{\theta})| + \mathcal{E}(\theta^*, \widehat{\theta}). \end{aligned} \tag{D.28}$$

The pair bound is deterministic over the whole ball, and the estimation event involves only $b$. Therefore this argument also covers finite menus chosen after observing the moment. It does not need a union bound over menus or candidates. □

The conditional exponential family makes the comparison quantitative. Relative to $Q_{\sigma,\vartheta}$, the law $Q_{\sigma,\theta}$ is its normalized tilt by $e^{(\theta-\vartheta)^\top \xi_i}$. Applying Lemma D.1 with this parameter difference gives

$$T_\theta \le t_\theta := \tanh(\kappa\|\theta\|/2), \qquad D_{\theta,\vartheta} \le d_{\theta,\vartheta} := \tanh(\kappa\|\theta - \vartheta\|/2). \tag{D.29}$$

Thus $\mathcal{E}$ can be replaced by the explicit envelope

$$\overline{\mathcal{E}}(\theta, \vartheta) = \frac{1}{6}\min\{t_\theta + t_\vartheta,\, d_{\theta,\vartheta} + |s(\theta) - s(\vartheta)| \min(t_\theta, t_\vartheta)\} \le \frac{1}{3}\tanh(H\kappa/2). \tag{D.30}$$

The envelope is zero when the candidates coincide and never weakens the strengthened uniform bound. It can be smaller when the candidates are close to each other even if both have moved from the reference. Both their conditional change and their change in block weights matter.

An observable comparison follows on the same event. For any candidate pair, with $\Delta s = s(\theta) - s(\vartheta)$,

$$J_U(P_\theta) - J_U(P_\vartheta) \ge \frac{\widehat{b}\,\Delta s - e_n|\Delta s|}{6} - \overline{\mathcal{E}}(\theta, \vartheta). \tag{D.31}$$

A positive right-hand side certifies that comparison in this model. Its being nonpositive means that this sufficient certificate is inconclusive, not that the first candidate is worse. Taking the maximum of the right-hand side of Eq. (D.25) over $\theta \in \mathcal{C}$, with $\widehat{\theta}$ fixed, also removes its reference to the unknown optimizer $\theta^*$.

### D.5 SELECTION WITH A NONZERO BLOCK CONTRAST

The uniform bound remains valid when $b$ is zero or arbitrarily close to zero. A different conclusion is available when the block contrast is separated from zero. For a nonempty finite menu define

$$S_{\mathcal{C}} = \max_{\theta \in \mathcal{C}} s(\theta) - \min_{\theta \in \mathcal{C}} s(\theta), \qquad \Delta_{\mathcal{C}} = \frac{|b| S_{\mathcal{C}}}{6}, \qquad \mathcal{E}_{\mathcal{C}} = \max_{\theta, \vartheta \in \mathcal{C}} \mathcal{E}(\theta, \vartheta). \tag{D.32}$$

Here $\Delta_{\mathcal{C}}$ is the range of the affine part of the utility, not the minimum gap between distinct candidates. It equals the total utility range when $\kappa = 0$. All audit draws use the fixed proposal of Eq. (10), with replacement. A repeated identity returns its fixed label. Fix a measurable tie rule.

**Theorem D.2** (Selection with a nonzero contrast)**.** *Under the paired setting, suppose $b \neq 0$. For every realized audit and every eligible finite menu,*

$$\mathcal{R}_{\mathcal{C}} \leq \mathcal{E}_{\mathcal{C}} + \Delta_{\mathcal{C}} \mathbf{1}\{b\widehat{b} \leq 0\}. \tag{D.33}$$

*If the menu is fixed independently of the audit draws, then*

$$\Pr(\mathcal{R}_{\mathcal{C}} > \mathcal{E}_{\mathcal{C}}) \leq e^{-2nb^2}, \qquad \mathbb{E}\mathcal{R}_{\mathcal{C}} \leq \mathcal{E}_{\mathcal{C}} + \Delta_{\mathcal{C}} e^{-2nb^2}. \tag{D.34}$$

*When $\kappa = 0$, $\mathcal{E}_{\mathcal{C}} = 0$ and these bounds control error probability and expected regret directly. In particular, the fixed-menu expected-regret bound becomes*

$$\mathbb{E}\mathcal{R}_{\mathcal{C}} \leq \frac{|b| S_{\mathcal{C}}}{6} e^{-2nb^2}, \qquad S_{\mathcal{C}} = \max_{\theta \in \mathcal{C}} s(\theta) - \min_{\theta \in \mathcal{C}} s(\theta). \tag{D.35}$$

*Proof.* On $b\widehat{b} > 0$, multiplying $s(\theta)$ by $b$ or by $\widehat{b}$ gives the same maximizers. Consequently $b[s(\theta^*) - s(\widehat{\theta})] \leq 0$, and Theorem D.1's deterministic comparison gives $\mathcal{R}_{\mathcal{C}} \leq \mathcal{E}_{\mathcal{C}}$. On $b\widehat{b} \leq 0$, including a zero estimated contrast, the affine difference is at most $|b| S_{\mathcal{C}}/6$. Its remainder difference is at most $\mathcal{E}_{\mathcal{C}}$. This proves Eq. (D.33) for every audit outcome and tie rule.

The independent observations $Z_j \in [-1/2, 1/2]$ have mean $b$. For $b > 0$, the event $b\widehat{b} \leq 0$ is $\widehat{b} - b \leq -b$. The one-sided Hoeffding bound gives probability at most $e^{-2nb^2}$. For $b < 0$, use the other tail with displacement $|b|$. Thus

$$\Pr(b\widehat{b} \leq 0) \leq e^{-2nb^2}. \tag{D.36}$$

For a fixed menu, the quantities in Eq. (D.32) are deterministic conditional on the fixed model and oracle. Taking probabilities and expectations in Eq. (D.33) proves Eq. (D.34). When $\kappa = 0$, the conditional laws do not change and the remainders vanish. □

Two uniformity distinctions are useful. The pointwise inequality holds even for a menu depending on $\widehat{b}$. In that case the event $\mathcal{R}_{\mathcal{C}} > \mathcal{E}_{\mathcal{C}}$ is still contained in the same sign-error event, but the expectation cannot pull the now-random $\Delta_{\mathcal{C}}$ and $\mathcal{E}_{\mathcal{C}}$ outside. Deterministic envelopes, such as $S_{\mathcal{C}} \leq 2$ and $\mathcal{E}_{\mathcal{C}} \leq \tanh(H\kappa/2)/3$, remain available. Alternatively, one may condition on an independent history that fixes the menu before the audit.

On the common confidence event $|\widehat{b} - b| \leq e_n$, either $|b| > e_n$ or the observable condition $|\widehat{b}| > e_n$ is sufficient for a correct, nonzero sign. The latter gives a certificate for every finite menu on this event, not a conditional coverage claim after selecting only audits that pass the test. The former is ensured by

$$n > \frac{\log(2/\delta)}{2b^2}. \tag{D.37}$$

On that event $\mathcal{R}_{\mathcal{C}} \leq \mathcal{E}_{\mathcal{C}}$, and it is zero at $\kappa = 0$. This is a sufficient draw count, not a necessary label budget. The bounds hold at a specified $n$, not after arbitrary optional stopping. When $b = 0$, Eq. (D.33) reduces directly to $\mathcal{R}_{\mathcal{C}} \leq \mathcal{E}_{\mathcal{C}}$ with no sign decision, while Theorem 3 continues to apply without a contrast assumption.

### D.6 When the finite-budget guarantee is informative

The numerical illustration in Appendix D.1 has $b = 1/16$, $h = 0.25$, $a_* \simeq 0.1202700373$, and $\kappa = 0$. The actual two-candidate gap is

$$\Delta_{\text{pair}} = \frac{|b|}{3} \tanh(ha_*) \simeq 0.062622 \text{ pp}, \tag{D.38}$$

whereas Eq. (12)'s estimation term at $n = 128$ and $\delta = 0.05$ is about $0.120274$ pp. The inequality is valid but its numerical value does not improve on the pair's entire possible loss at this budget. It should not be interpreted as a tight 128-draw certificate.

Table D.1: Deterministic calculations for the fixed symmetric pair at $\delta = 0.05$. The uniform term and gap are in pp. Exact errors use iid signed draws and a fair tie between the two endpoint candidates, not new policy experiments.

| Draws $n$ | $e_n$ | Uniform estimation term | Exact sign-error probability |
|---|---|---|---|
| 128 | 0.120040 | 0.120274 | 0.078239 |
| 256 | 0.084881 | 0.085047 | 0.022432 |
| 473 | 0.062446 | 0.062567 | 0.003160 |
| 512 | 0.060020 | 0.060137 | 0.002269 |
| 2,048 | 0.030010 | 0.030069 | $6.80 \times 10^{-9}$ |

There is nevertheless an effective budget region. In this example, the right-hand side of Eq. (D.37) is about $472.177$, so $n \geq 473$ makes the original two-sided confidence event certify the correct sign at level $0.95$. A sign decision alone has the separate one-sided sufficient count $n \geq 384$ from Eq. (D.36). Neither is an optimized necessary threshold. At $n = 128$, the latter bound gives expected regret at most $0.062622e^{-1} \simeq 0.023037$ pp, already below the entire pair gap. The exact binomial calculation gives a still smaller expected regret of about $0.004899$ pp.

For completeness, $Z + 1/2 \sim \text{Bernoulli}(1/2 + b)$. With $b > 0$ and $K \sim \text{Bin}(n, 1/2 + b)$, a fair tie between the endpoint candidates gives error

$$\Pr(K < n/2) + \tfrac{1}{2} \Pr(K = n/2). \tag{D.39}$$

The second term is zero for odd $n$. This is the calculation in Table D.1, distinct from the historical 20,000-repeat estimates in Table 1.

When $\kappa > 0$, a sufficient sample count for the strengthened uniform bound to be at most a prescribed $\rho$ is

$$n \geq \frac{S_{\mathcal{C}}^2 \log(2/\delta)}{72[\rho - \tanh(H\kappa/2)/3]^2}, \qquad \rho > \tanh(H\kappa/2)/3. \tag{D.40}$$

For a fixed menu with $S_{\mathcal{C}} = 0$, only the remainder term needs control. For a menu chosen after the audit, its span must instead be replaced by a fixed upper envelope when prescribing $n$. If the displayed condition on $\rho$ fails, this bound supplies no certificate, not an impossibility for the particular task. Equation (D.31) and the contrast condition can be sharper. The uniform theorem remains useful because it requires no nonzero margin and covers menus selected from the whole parameter region on a single event.

### D.7 Exact information for a general candidate menu

This subsection proves Theorem 1, independently of the paired construction. Let $\Omega$ be finite, let $p_0(\omega) > 0$, and let $\{p_c\}_{c \in \mathcal{C}}$ be probability vectors on $\Omega$. All candidates and summary functions are fixed before utility is chosen. Write $L_c = p_c/p_0$, $f_{c,b} = L_c - L_b$, and use the inner product $\langle f, g \rangle_0 = \sum_\omega p_0(\omega) f(\omega) g(\omega)$. The utility class in this subsection is the entire cube $[0,1]^\Omega$. The report contains $M_j(U) = \langle U, h_j \rangle_0$. Let $\mathcal{H}$ be the span of the $h_j$. Redundant functions are allowed.

*Proof of Theorem 1.* We prove the cycle (i)⇒(ii)⇒(iii)⇒(i), using the same maps $M, D$ and optimizer sets $\mathcal{A}$ as in the statement. The final implication is by contraposition; its construction also proves the theorem's last assertion.

*Step 1: (i)⇒(ii), reconstruction of every contrast.* Assume $\mathcal{F} \subseteq \mathcal{H}$. For each pair $c, b$, choose coefficients $\beta_{c,b,j}$ with $f_{c,b} = \sum_j \beta_{c,b,j} h_j$. The coefficients depend only on the fixed policies and summary functions, not on $U$. Hence

$$D_{c,b}(U) = J_U(P_c) - J_U(P_b) = \langle U, f_{c,b} \rangle_0 = \sum_j \beta_{c,b,j} M_j(U). \tag{D.41}$$

Equal reports give equal right-hand sides for every pair, proving $M(U) = M(U') \Rightarrow D(U) = D(U')$.

*Step 2: (ii)⇒(iii), a selector defined by the report.* Assume (ii), and fix one baseline $b_0 \in \mathcal{C}$ and an ordering of the finite menu. For each attainable report $m \in M(\mathcal{U})$, define $\widetilde{D}_{c,b_0}(m) = D_{c,b_0}(U)$ using any $U$ with $M(U) = m$. Condition (ii) makes this definition independent of which compatible utility is used. Define

$$\chi(m) = \text{first} \arg\max_{c \in \mathcal{C}} \widetilde{D}_{c,b_0}(m), \tag{D.42}$$

where "first" uses the fixed ordering. The argmax is nonempty because the menu is finite and nonempty. For every $U$, subtracting the common value $J_U(P_{b_0})$ leaves the maximizers unchanged, so $\chi(M(U)) \in \mathcal{A}(U)$. This proves (iii) without assuming that the selector observes $U$.

*Step 3: ¬(i)⇒ ¬(iii), indistinguishable reports with disjoint optima.* If $\mathcal{F} \nsubseteq \mathcal{H}$, at least one generating contrast $f = f_{c,b}$ lies outside $\mathcal{H}$. Project it using $\langle \cdot, \cdot \rangle_0$ and set $r = f - \Pi_{\mathcal{H}} f$. Then $r \perp \mathcal{H}$, $\langle r, f \rangle_0 = \|r\|^2_{L^2(P_0)} > 0$, and finite positive support gives $0 < \|r\|_\infty < \infty$. Define

$$U_\pm = \frac{1}{2} \pm \frac{r}{2\|r\|_\infty}. \tag{D.43}$$

These are admissible utilities in $\mathcal{U}$. For every reported function,

$$M_j(U_+) - M_j(U_-) = \frac{\langle r, h_j \rangle_0}{\|r\|_\infty} = 0, \tag{D.44}$$

which proves the equal-report requirement in the theorem's last assertion. The selected pair already has different contrasts, because $\langle 1, f \rangle_0 = 0$:

$$D_{c,b}(U_\pm) = \pm \frac{\|r\|^2_{L^2(P_0)}}{2\|r\|_\infty}. \tag{D.45}$$

A pair reversal is not enough to rule out a common optimal third candidate. To verify the full-menu requirement, set $z_d = \langle L_d, r \rangle_0$ for every $d \in \mathcal{C}$. Since each $p_d$ sums to one,

$$J_{U_\pm}(P_d) = \frac{1}{2} \pm \frac{z_d}{2\|r\|_\infty}, \qquad \mathcal{A}(U_+) = \arg\max_d z_d, \quad \mathcal{A}(U_-) = \arg\min_d z_d. \tag{D.46}$$

The values $z_d$ are not all equal: $z_c - z_b = \langle r, f \rangle_0 > 0$. Thus their maximizing and minimizing sets are disjoint. Any function $\chi$ of the common report must return the same index in both worlds, which cannot belong to both optimal sets. Condition (iii) fails. This proves (iii)⇒(i) by contraposition and completes the equivalence. Equations (D.44) and (D.46) establish the final assertion explicitly. □

Step 3 explains failure of exact selection. The same two utilities also quantify the resulting regret, including for randomized selectors that might distribute their choices across candidates. Any rule receiving only the common report and candidate laws satisfies

$$\max_{z \in \{+,-\}} \mathbb{E}_z \mathcal{R}_{\mathcal{C}} \geq \frac{\max_d z_d - \min_d z_d}{4\|r\|_\infty} \geq \frac{\|r\|^2_{L^2(P_0)}}{4\|r\|_\infty}. \tag{D.47}$$

For each output index, the two regrets sum to $(\max z_d - \min z_d)/(2\|r\|_\infty)$. The same sum holds in expectation under a common randomized output law. The binary-oracle and rollout results in the preceding appendices use their own restricted utility classes.

*How many exact moments?* Fixing one baseline $b$ gives $\mathcal{F} = \text{span}\{L_c - L_b : c \neq b\}$ and $r_{\mathcal{C}} := \dim \mathcal{F} \leq |\mathcal{C}| - 1$. A basis of $\mathcal{F}$ yields $r_{\mathcal{C}}$ sufficient moments. Conversely, $q$ linear moments span a

space of dimension at most $q$, so Theorem 1 requires $q \geq r_{\mathcal{C}}$. Given an existing function space $\mathcal{H}_0$, the minimum additional number is

$$\dim(\mathcal{F} + \mathcal{H}_0) - \dim \mathcal{H}_0 = \dim \mathcal{F} - \dim(\mathcal{F} \cap \mathcal{H}_0). \tag{D.48}$$

To attain this value, extend a basis of $\mathcal{H}_0$ to a basis of $\mathcal{H}_0 + \mathcal{F}$. These statements count unrestricted exact linear functionals. They say nothing by themselves about estimation variance, response-scoring cost, reporting precision, or sample complexity. For the paired model at $\kappa = 0$, the relevant contrasts restricted to $I$, where utility is unknown, have rank one. Contrasts outside $I$ contribute zero because utility there is known to vanish. The full policy contrast space need not have rank one.

*Utility class and precision.* The necessity argument uses the full bounded-utility cube. It must not be applied unchanged to a fixed binary label class. For example, on two atoms with $P_0 = (1/3, 2/3)$, the single exact mean $\mathbb{E}_0 U$ takes four distinct values on the four binary utilities, identifying all of them, even though nonconstant contrasts are not in the span of the constant function. The bounded utilities in Eq. (D.43) need not be binary. Likewise, a pair reversal alone does not generally exclude a common good third candidate, it does so here because all candidate values are symmetric around the same constant $1/2$. The binary full-menu construction in Section 4 supplies its own argument under a narrower oracle class and finite reporting precision.

### D.8 Approximate information from a partition

Let $S$ be any finite partition of $\Omega$, with zero-reference-mass cells removed. Exact conditional utility means correspond to the moment functions $h_s(\omega) = \mathbf{1}\{S(\omega) = s\}/P_0(S = s)$. Their span consists exactly of functions constant within cells. Theorem 1 therefore shows that they are universally sufficient for the bounded utility class precisely when every $f_{c,b}$ is constant within cells. If $P_0$ is itself in the candidate menu, this is equivalent to each $L_c$ being constant within cells. In that case $P_c(\cdot \mid S = s) = P_0(\cdot \mid S = s)$ whenever $P_c(S = s) > 0$. Without base in the menu, only the contrasts, not the individual likelihood ratios, must be cellwise constant.

**Theorem D.3** (Comparison error from cell summaries)**.** *Set $\mu_s = \mathbb{E}_0[U \mid S = s]$, $a_s = P_c(S = s) - P_b(S = s)$, and $r_{c,b} = f_{c,b} - \mathbb{E}_0[f_{c,b} \mid S]$. Whenever $|\widehat{\mu}_s - \mu_s| \leq e_s$ in all cells, every candidate pair and every $U \in [0, 1]$ satisfy*

$$\left| J_U(P_c) - J_U(P_b) - \sum_s \widehat{\mu}_s a_s \right| \leq \sum_s e_s |a_s| + \tfrac{1}{2}\mathbb{E}_0|r_{c,b}|, \quad a_s = P_c(S = s) - P_b(S = s). \tag{D.49}$$

*Proof.* The exact decomposition is

$$\mathbb{E}_0[U f_{c,b}] = \sum_s \mu_s a_s + \mathbb{E}_0[U r_{c,b}]. \tag{D.50}$$

Conditional centering gives $\mathbb{E}_0 r_{c,b} = 0$. The positive and negative parts of $r_{c,b}$ consequently have equal expectations, each $\mathbb{E}_0|r_{c,b}|/2$. Since $0 \leq U \leq 1$, the remainder is between the negative of this quantity and the quantity itself. Replacing $\mu_s$ by $\widehat{\mu}_s$ adds at most $\sum_s e_s|a_s|$. This proves the claim. □

The factor $1/2$ in the residual bound is sharp uniformly over unrestricted bounded utilities, as witnessed by $U = \mathbf{1}\{r_{c,b} > 0\}$. This need not be a tight bound after particular cell means are fixed. A positive estimated gain larger than the right-hand side certifies the pair ordering on the cell-estimation event. A large residual bound, however, is not evidence that a given fixed utility causes an error.

For a selection rule maximizing $\widehat{J}_c = \sum_s P_c(S = s)\widehat{\mu}_s$, let $c^*$ maximize true utility. The same event gives the direct pairwise regret bound

$$\mathcal{R}_{\mathcal{C}}(\widehat{c}) \leq \sum_s e_s |P_{c^*}(S = s) - P_{\widehat{c}}(S = s)| + \frac{1}{2}\mathbb{E}_0|r_{c^*,\widehat{c}}|. \tag{D.51}$$

This follows by applying the theorem to $(c^*, \widehat{c})$ and using $\widehat{J}_{c^*} - \widehat{J}_{\widehat{c}} \leq 0$. The cell-error event must be simultaneous. If its construction or the partition uses the same audit data adaptively, that dependence must be controlled rather than ignored. In a known finite model the residual is computable without utility labels. On sampled neural responses, estimating likelihood ratios and conditional expectations introduces additional coverage and estimation requirements.

### D.9 THE ANALOGOUS CRITERION FOR A FIXED GROUP UPDATE

Equation (3) has a related elementary linear interpretation. Fix all response scores and grouping before considering a reward vector $u \in [0,1]^N$. Write its implemented centered reward update as $Bu$, and suppose the retained linear measurements are $Au$. Then $Au$ determines $Bu$ for every such $u$ if and only if

$$\ker A \subseteq \ker B, \quad \text{equivalently} \quad \operatorname{row}(B) \subseteq \operatorname{row}(A). \tag{D.52}$$

If the row inclusion holds, $B = CA$ for some matrix $C$. If the kernel inclusion fails, choose $v \in \ker A$ with $Bv \neq 0$ and $u_\pm = \mathbf{1}/2 \pm v/(2\|v\|_\infty)$. Their measurements coincide but their updates differ. This proves necessity as well as sufficiency. Here $B$ contains the actual implementation's score vectors and group weights, including length normalization if used. It describes that surrogate gradient, not an unbiased sequence-utility gradient or a finite-step optimization guarantee. The matched-feedback experiment uses binary allocations with fixed counts, not the entire bounded cube, and establishes its equal statistics by direct counting rather than by this converse.

### D.10 WHEN THE REUSE ASSUMPTIONS HOLD

The positive paired result in Section 5 supplies a tractable sufficient structure. Its general content is fixed conditional response composition, not pairing itself. The following bound makes that distinction explicit for an arbitrary finite partition, without restricting utility to the probe or requiring a two-dimensional policy.

For each candidate $\theta$, let $p_\theta(s) = P_\theta(S = s)$ and $Q_{\theta,s} = P_\theta(\cdot \mid S = s)$. Require $P_\theta \ll P_0$, define $\mu_s = \mathbb{E}_{Q_{0,s}} U$, and set

$$A_\theta = \sum_s p_\theta(s)\, \mathrm{TV}(Q_{\theta,s}, Q_{0,s}). \tag{D.53}$$

Terms with $p_\theta(s) = 0$ are defined as zero and require no conditional law there. The probabilities and bounds below refer to the true laws, not unverified sample estimates.

**Theorem D.4** (Reuse under conditional stability)**.** *For any fixed $U \in [0,1]$, $|J_U(P_\theta) - \sum_s p_\theta(s)\mu_s| \leq A_\theta$. On a simultaneous event $|\widehat{\mu}_s - \mu_s| \leq e_s$, choosing $\widehat{\theta} \in \arg\max_{\theta \in \mathcal{C}} \sum_s p_\theta(s)\widehat{\mu}_s$ gives*

$$\mathcal{R}_{\mathcal{C}} \leq \sum_s e_s |p_{\theta^*}(s) - p_{\widehat{\theta}}(s)| + A_{\theta^*} + A_{\widehat{\theta}}. \tag{D.54}$$

*If the conditional bounds hold throughout a fixed policy region, the same event covers all finite menus in that region.*

*Proof.* Condition on $S = s$ in $J_U(P_\theta)$. Each difference $\mathbb{E}_{Q_{\theta,s}} U - \mu_s$ is at most the corresponding TV in absolute value, giving the first statement after weighting. For the regret, subtract the two approximate values, insert the estimated means, and use optimality of $\widehat{\theta}$ for their estimated values. The moment error is weighted by $p_{\theta^*} - p_{\widehat{\theta}}$ and the two deterministic residuals give $A_{\theta^*} + A_{\widehat{\theta}}$. Only the same cell-error event is random, proving the stated uniformity. □

This theorem gives sufficient conditions, not an algorithm for discovering a partition or guaranteeing a favorable decision margin. It also shows that a worst-case bound in every cell is not always necessary, because drift is weighted by the candidate's cell probability. Applying it to the paired probe recovers its fixed-mass factors, while the contrast $b$ removes a common offset and makes the estimation term sharper.

*Exact stability by policy design.* Consider a fixed response-level partition and candidates of the form

$$P_\theta(\omega) = P_0(\omega) \frac{\exp(\beta_{\theta,S(\omega)})}{Z_\theta}. \tag{D.55}$$

Conditioning on a cell cancels the common exponential factor and normalization, so $Q_{\theta,s} = Q_{0,s}$ and $A_\theta = 0$. For a fixed prompt distribution, the analogous formula is normalized within each prompt and uses $S = (X, s(Y))$. It describes, for example, changing category weights in a frozen candidate pool or adding a cell-specific response-level score to a frozen generator distribution. It is not a claim that arbitrary token-level logit changes preserve response-level conditionals.

Similarly, in a system that first chooses an observed branch $Z$ and then samples from a fixed generator $Q_z(\cdot \mid X)$, updating only the branch probabilities preserves the conditional law given $(X, Z)$. This includes routing among frozen response generators when routing is the only trained component. If routing depends on the prompt, means indexed only by $Z$ need not remain valid because the prompt distribution within that branch changes. One must condition on $(X, Z)$ or justify additional pooling. If branch identity is unobserved and component supports overlap, a partition of responses alone need not recover this structure. A token-level mixture-of-experts network is not automatically a fixed mixture of complete-response generators.

*Approximate stability.* For policies of common positive support, write a log density change as

$$\log \frac{dP_\theta}{dP_0}(\omega) = \beta_{\theta,S(\omega)} + \varepsilon_\theta(\omega). \tag{D.56}$$

If $L_{\theta,s}$ bounds the oscillation of $\varepsilon_\theta$ within cell $s$, Lemma D.1 gives

$$\mathrm{TV}(Q_{\theta,s}, Q_{0,s}) \le \tanh(L_{\theta,s}/4). \tag{D.57}$$

The paired family uses $\varepsilon_\theta(i) = \theta^\top \xi_i$, hence $L_{\theta,s} \le 2H\kappa$. For a general shared neural update, small parameter norm or low adapter rank alone supplies no such oscillation bound. Sparse-prefix KL controls an even more limited set of histories. The v2 trajectory diagnostic is consequently not evidence that the reuse assumptions hold.

*Severity and practical precision.* Exact conditional invariance is strong for an unconstrained generator, but automatic in the two restricted designs above. Approximate reuse is useful when its weighted residual and estimation error are small relative to the candidate margin or allowed regret. For a concrete pair, a sufficient ordering test compares the estimated gain with the cell-estimation term and $A_\theta + A_\vartheta$, or with the more direct residual in Eq. (D.49). A bound that fails this test is inconclusive, not proof that the true ranking is wrong. A finer partition may reduce approximation error while increasing the number of means, scoring cost, and sample requirements. Stable conditional composition makes reuse possible, it does not establish a label or computation advantage over direct use of the same records.

# E Finite experiments and numerical checks

The shared-label comparison in Section 6.1 isolates the retained statistic. The checks here establish the scope of the model behind it: whether the GRPO menu remains separating away from symmetry, how large its utility gap is, and how reporting precision differs from sample size. We then examine the reuse error and budget comparison. Settings that fail the sufficient certificate are retained, so the numerical study does not select only favorable instances.

## E.1 Settings and actual candidate construction

To check that the information gap belongs to actual constructed policies, rather than an abstract pair of value vectors, the finite experiments use exponential-family policies, fixed binary labels, exact finite-group binomial coefficients, float64 probability calculations, and bisection for KL calibration. The common KL target is the original mean-16 candidate at norm $0.25$. The principal experiment uses seed 906101, and the sampled-report study uses seed 906102. Maximum reported calibration error in the finite-history checks is below $1.4 \times 10^{-15}$, which quantifies numerical calibration rather than symbolic KL equality.

Robustness is evaluated on a fixed grid, rather than by repeatedly shrinking perturbations until the certificate passes. The 24 settings cross $\kappa \in \{0, 0.02, 0.05\}$, weight dispersion $\{0.1, 0.3\}$, initial block imbalance $\{0.005, 0.05\}$, and pairing or a non-paired perturbation of amplitude $0.0005$. Each setting has one fixed generated feature/weight table. At $R = 4096$ and $\tau = 0.01$, all 12 dispersion-0.3 settings fail the sufficient precision condition, whereas the other 12 pass. The support curve fixes dispersion $0.3$, $\kappa = 0.02$, $a = 0.005$, and $\tau = 0.005$, passing only at $R = 65536$ among $\{256, 1024, 4096, 16384, 65536\}$. This is a specified grid, not a distributional prevalence estimate.

The representative paired setting uses $\kappa = 0.02$, dispersion $0.1$, $a = 0.005$, $R = 4096$, $\epsilon = 1/16$, $\alpha = 0.01$, and $\tau = 0.01$. It has $B = 0.01001698$, $r_\mu = 0.00570780$, $r_\Gamma = 0.0000672745$, and $\eta = 0.000558787$. Across 2,000 fixed oracles per world, actual pair gaps average approximately

$\pm 0.0626$ pp. All sampled oracles satisfy the good event and sign, and zero failures give a Wilson 95% upper endpoint of about $0.00192$, not zero population failure.

Table E.1: Full-menu outer acceptable sets in the representative finite model. Both worlds report the central $1/2$ bin. At $t = 0.000208687$, the signed-audit upper bound is $2 \max_c r_c = 0.000161780 < t$.

| World | Potentially $t$-acceptable configurations |
|---|---|
| $z = +1$ | mean-4, mean-8, mean-16, std-4 |
| $z = -1$ | std-8, std-16 |

The remaining construction check replaces expected gradients with sampled ones. For each $G \in \{1024, 4096, 16384\}$, 100 label-free rollout histories use shared groups for the pair. All generated pairs pass the actual contrast precision/margin certificate, with Wilson lower endpoint about $0.963$. This event differs from requiring each candidate to remain on its prescribed side of base. In an earlier 300-repeat study, increasing $G$ from 16 to 16,384 increases that stricter opposite-side frequency from $14.3\%$ to $100\%$, and the average actual-to-ideal gap ratio from about $0.485$ to $0.998$.

### E.2 Gap magnitude across a fixed KL grid

To quantify the scale of the construction, a deterministic sensitivity analysis evaluates all six configurations and base at KL targets $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1$. It keeps the model in Appendix A and $\epsilon = 1/16$ fixed, rather than optimizing its parameters for a large gap. At the balanced point, the two worlds have exact opposite gaps for every fixed-count oracle. A second, fixed paired instance uses $R = 4096$, $a = 0.005$, dispersion $0.1$, and residual norms at most $0.02$. Its table is generated once with NumPy seed 946240 and reused at every KL target. It is a separate specified sensitivity instance from the seed-906101 experiment.

For the full menu, define the tolerance threshold

$$t_* = \min_c \max_{z \in \{-1,1\}} \{L_z - q_{z,c} - r_c\}. \tag{E.1}$$

A candidate belongs to both outer acceptable sets precisely when both displayed deficits are at most $t$. Hence $t < t_*$ suffices for their disjointness. All candidates, including base and the coinciding mean configurations, enter this calculation. Table E.2 reports every target. The half-width grid is

Table E.2: Actual KL and decision-gap scale. Gaps and tolerances are in pp. The balanced gap is exact. The perturbed margin is the sufficient pair bound $\eta$, and $t_*$ uses simultaneous full-menu radii at $\alpha = 0.01$.

| $\mathrm{KL}(P_0 \| P_c)$ | Balanced pair gap | Balanced $t_*$ | Perturbed $\eta$ | Perturbed $t_*$ |
|---|---|---|---|---|
| $10^{-4}$ | 0.006094 | 0.003047 | 0.005449 | 0.002100 |
| $10^{-3}$ | 0.019274 | 0.009637 | 0.017234 | 0.006640 |
| $10^{-2}$ | 0.061075 | 0.030537 | 0.054559 | 0.021035 |
| $10^{-1}$ | 0.197066 | 0.098533 | 0.174890 | 0.067763 |
| 1 | 0.732828 | 0.366414 | 0.625887 | 0.247745 |

$\tau \in \{0.001, 0.005, 0.01, 0.02\}$. At the balanced point, the exact report is $1/2$, so all 20 KL/precision combinations pass. In the perturbed instance, the simultaneous report condition requires $\tau > 0.00653602$, independently of the KL target. The ten combinations with $\tau = 0.01$ or $0.02$ pass, and the other ten fail that sufficient condition. Failure of a certificate does not prove absence of ambiguity. These are probability-scale half-widths, whereas the table multiplies utility differences by 100.

The grid makes the quantitative limitation visible: at KL $0.01$, the ideal gap is about $0.061$ pp, and even KL $1$ gives less than one pp in this particular family. The construction establishes realizability and robustness of the information gap. It does not establish that its magnitude explains the larger neural training gains. The supplied script recomputes the candidate probabilities, calibrations, intervals, and all precision checks without a model or new labels.

### E.3 Reporting resolution, exact encoding, and shared-label summaries

To distinguish additional labels from additional reporting precision, we keep the report bin fixed while increasing the sample count. For the representative model, each of the 2,000 fixed oracles per world is audited 20 times at each $N$. The central-bin frequencies are shown in Table E.3. The largest $N$ follows a conservative sufficient bound using $\zeta = 0.00198985$ and $\beta = 0.01$, not a recommended audit budget. Repeated audits share the same oracle and cannot be counted as independent oracle replications. This count-law simulation checks marginal report frequencies, while Appendix C handles retained identities.

Table E.3: Observed central-bin frequency at fixed reporting half-width $\tau = 0.01$.

| Raw draws $N$ | World $+1$ | World $-1$ |
|---|---|---|
| 128 | 0.2081 | 0.2110 |
| 8,192 | 0.9278 | 0.9306 |
| 669,063 | 1.0000 | 1.0000 |

The opposite limit, exact real-valued reporting, explains why the separation theorem states a precision restriction. A 16-atom example uses accepted weights proportional to $2^{24} + 2^i$, $i = 0, \ldots, 15$, with denominator 268500991. All 65,536 binary subset sums differ. At $R = 8$ and $\epsilon = 1/8$, each world contains 3,136 fixed-count assignments, all inside the predetermined $1/2 \pm 0.001$ bin. Positive actual pair gaps lie in $[0.0012512046, 0.0012532852]$, with the corresponding negative interval. Integer arithmetic checks the encoding, whereas floating-point candidate evaluation checks the gaps. This is not an empirical test of asymptotic lower-bound constants.

The summary comparison in Table 1 uses 20,000 iid multinomial repetitions, shared raw samples, and random tie breaking. At $n = 128/a^2$, coarse errors become approximately $0.0799, 0.0790, 0.0768$. A finite oracle with $2^{40}$ atoms per block makes the maximum 320,000 draws correspond to about 319,999.9767 expected unique labels. Both summaries incur identical unique-query costs on the shared draws.

### E.4 Reuse and finite verification scope

The reuse checks separate drift from sampling error, as in Section 5. With the old means known exactly, Proposition 1's worst-world regret evaluates to approximately $0.001253$, $0.005011$, and $0.012528$ pp at $\kappa = 0.005, 0.02, 0.05$. Its increase reflects within-block changes that additional precision on those means cannot resolve; no estimation error is included.

An independent finite implementation enumerates all 256 binary oracles of an unequal-weight, heterogeneous eight-atom probe over 119 parameter points and checks utility approximation and selection with perturbed moments. It also checks 876 conditional-label histories and exact subset sums. These numerical checks do not replace the written proofs of the separation and reuse results.

A separate budget comparison tests a different requirement: whether acquiring a useful profile is preferable to spending the same labels directly. At 128 distinct labels, direct-estimation MSE is about $0.001936$, versus $0.00001405$ when a useful moment is already available and $0.007971$ when the budget must purchase a split profile. The distinction matters for interpreting Theorem 3: useful side information need not be worth acquiring under a fixed total budget. This comparison does not establish a profiling advantage over direct estimation; the underlying budget records are retained in the bundle.

## F Response prediction under full and limited supervision

Section 6.3 evaluates feedback through the response differences used by group-relative training. The prediction studies here explain why cross-task correctness prediction is not a substitute for that evaluation. We first hold the handcrafted features fixed and change the fitting objective, then replace them with frozen response representations. The full-label and sparse-label stages distinguish whether

information is present from whether it can be learned under a limited budget. Neither unsuccessful sparse predictor is used for policy training.

The group diagnostic uses $C_g = I - \mathbf{1}\mathbf{1}^\top/m_g$ and compares centered predicted rewards with centered complete labels. Its skill is

$$100\left(1 - \frac{\sum_g \|C_g(\widehat{\mathbf{r}}_g - \mathbf{u}_g)\|^2}{\sum_g \|C_g\mathbf{u}_g\|^2}\right). \tag{F.1}$$

The sums use the stated diagnostic subset, with nonzero denominator. Positive skill means lower error than zero centered feedback, not a policy-improvement guarantee.

### F.1 Data and evaluation targets

The recorded public-reward and complete-reward histories each contain 1,024 generated code responses, in groups of eight, from the controlled record-transformation domain. They are denoted C-V and C-U in the experiment records. Each history spans 61 structure signatures and contains 128 consecutive eight-response groups. There are only 127 distinct $(\mathrm{step}, \mathrm{task_id})$ pairs, so repeated pairs do not merge groups. Sorted signatures are shuffled using NumPy seed 943300 and allocated round-robin to five folds, with no task–program identity shared between folds. Program identities are task–extracted-program pairs, with repeated programs sharing fixed labels. The v4 complete checker is the finite utility reference.

Five-fold evaluation holds out entire structure signatures. The primary diagnostic subsets contain 29 C-V and 30 C-U groups in which public accepts all eight responses while complete labels are mixed. The predictors do not receive private tests, private failure logs, or reference programs. These subsets are selected for *analysis* using trusted labels, not made available to a sparse acquisition policy in advance.

Centered skill is defined by Eq. (F.1), relative to zero centered advantage. The concordance metric compares predicted scores of trusted-positive and trusted-negative responses within the same group. Global Brier skill uses a training-fold constant label-mean comparator on its broader evaluation population. Thus global Brier skill and diagnostic-group skill are not scores for the same population and baseline.

### F.2 Handcrafted features and changing the fitting objective

The first comparison asks whether a fitting objective focused on absolute correctness misses information useful within a group. The fixed 28 public features describe operation categories, response length and termination, AST size, loops, conditionals, comprehensions, and call patterns. A global logistic fit uses ridge coefficient $0.1$. Global-linear and group-centered-linear controls then change the linear objective without adding features or labels; the logistic fit remains a separate historical reference. Table F.1 reports both response histories so that a result on complete-reward trajectories is not mistaken for one learned without that training information. The global logistic fits predict cross-task correctness

Table F.1: Full-label prediction diagnostics. Skill is measured on public-silent, trusted-mixed groups; histories and predictors remain separate.

| History | Predictor | Skill (%) | Concordance |
|---|---|---|---|
| C-V | Global logistic, 28 features | 0.0911 | 0.5372 |
| C-V | Group linear, 28 features | −2.3608 | 0.5243 |
| C-V | Length + average log probability | 1.8835 | 0.5890 |
| C-V | Frozen representation, PCA64 | 7.5256 | 0.6117 |
| C-U | Global logistic, 28 features | 3.3882 | 0.6283 |
| C-U | Group linear, 28 features | −0.5323 | 0.5664 |
| C-U | Length + average log probability | 2.3564 | 0.6136 |
| C-U | Frozen representation, PCA64 | 4.9290 | 0.6165 |

reasonably well: AUROC/Brier skill are $0.7932/27.0349\%$ on C-V and $0.7828/25.8036\%$ on C-U. Their weak group scores therefore expose a mismatch between the two evaluation targets. Sixteen

features are constant within the diagnostic groups and vanish under centering. Fitting the linear model directly to group differences, including training-fold scale calibration, does not resolve the shortfall. An alternative solver reproduces the predictions, but feature-collision checks do not establish an information-theoretic impossibility. The next comparison consequently changes the representation rather than treating this particular failure as a limit on reward prediction generally.

**Exact linear design.** The 28-dimensional vector consists of 12 response features and 16 operation-category indicators. Feature means and standard deviations are fitted only on unique public-positive programs in the training fold, with a standard-deviation floor of $10^{-8}$. Let $X_g$ append an intercept to the standardized features and write $w = (b, \beta)$. The prediction is public-gated, $\mathrm{diag}(V_g)X_g w$, and the group objective is

$$\frac{1}{G}\sum_{g}\frac{1}{8}\left\|C_g\left[\mathrm{diag}(V_g)X_g w - u_g\right]\right\|^2 + 0.1\|\beta\|^2. \tag{F.2}$$

The intercept is unpenalized. The global-linear control removes $C_g$. An augmented least-squares system is solved with `numpy.linalg.lstsq`, using `rcond=1e-12`. Scores are not clipped to $[0, 1]$. Each history is analyzed separately, particularly because C-U was itself generated using complete rewards.

### F.3 Frozen representations and the 128-label check

To test whether the handcrafted features omit response-level information, a frozen Qwen3-4B reads the original specification, public examples, and actual program. Its final normalized hidden layer at the last nonpadding program token has 2,560 dimensions. Extraction uses no adapter, evaluation mode, BF16, SDPA, batch size four, right padding with an attention mask, and contexts of at most 4,096 tokens. The fixed base revision is the one specified in Appendix G. We extract it once for 1,864 distinct task–program pairs, then fit whitened PCA to 64 dimensions on unique public-positive programs in each training fold. Randomized SVD uses seed 944170, ten oversampling dimensions, two power iterations, and a scale floor of $10^{-8}$. The downstream model uses Eq. (F.2). Code-correctness information in response representations has also been studied for candidate ranking (Ribeiro et al., 2026), whereas here the target is within-group feedback under structural holdout.

The frozen-representation predictor has positive centered skill in four of five structural folds in each history; folds are not independent model-training replications. C-V is the primary decision history, while C-U is secondary because its generation trajectory already used trusted reward. The full-label C-V result meets the development criteria of skill at least $5\%$ and concordance at least $0.60$, permitting one prespecified sparse check. C-U's $4.9290\%$ is not rounded up to a passing $5\%$ result.

The full-label result establishes a signal, but not its availability at low label cost. The sparse stage therefore samples complete groups within each outer training fold, subject to a cap of 128 distinct task–program labels. Both PCA and regression are fitted only to the selected programs. Acquisition seeds 944200–944215, offset by $100$ times the fold index, randomize the training-group order. Acquisition stops before the next whole group would exceed 128 distinct task–programs, without skipping that group or inspecting its labels to find a more favorable group. The realized counts are 121–128 labels per fold. On C-V, only 2/16 satisfy both criteria, against a required 12/16, with median skill $-5.2573\%$ and concordance $0.6003$. C-U also passes only 2/16, with medians $-5.1699\%$ and $0.5870$. Neither predictor is used in a new policy-training run.

This sparse procedure reduces both supervised labels and the unlabeled representation sample used for PCA. It therefore tests the joint sparse pipeline, not an impossibility at 128 labels with an independently well-estimated representation. Folds share training data, and acquisition seeds are not fresh language-model training seeds. A per-fold budget of 128 is not the total label cost of the multi-fold study.

### F.4 Costs and checks

The simple reference features are $\log(1 + \text{program token count})$ and the frozen model's average program-token log probability. Representation extraction processes 958,032 context tokens in 177.45 seconds, including 3.57 seconds loading, with peak CUDA memory about 8.38 GB. A separate

small check uses 65.56 seconds of program time. Extraction acquires no new labels, but consumes additional forwards, and the 1,864 historical v4 labels remain part of the information cost. Another normal-equation implementation agrees on 20 full-label fits with maximum prediction difference $1.94 \times 10^{-14}$. Budget and isolation checks cover 160 sparse acquisitions. These are implementation checks, not independent-task replications.

# G Diagnostic execution feedback for actual policy training

The training comparison in Section 6.3 requires three distinct checks: the test suites must supply valid information, that information must affect the generated-program endpoint, and any benefit must be interpreted with its acquisition and execution costs. This appendix gives the shared implementation first, then the offline and on-trajectory diagnostics supporting the endpoint, and finally the resource accounting. These analyses distinguish feedback quality from a claim of inexpensive moment reuse.

## G.1 Task domain, information access, and training comparison

The controlled domain asks for Python programs that transform records using specified operations, including filtering, duplicate-key handling, ordering, and aggregation. A deterministic domain interpreter supplies expected outputs. The training task collection contains 320 specifications. Public inputs and the complete-v4 checker are fixed, and generated program success refers to this finite checker, not formal correctness over all legal inputs.

Diagnostic inputs are generated from specifications using frozen rules for empty cases, duplicate-key conflicts, ties/order, and boundaries. They are not tailored after observing a particular evaluated program. Four extra inputs per task are used by both diagnostic and random conditions. Public tests precede the additional checks, and repeated task–program executions may be cached. The diagnostic rules access domain semantics through the interpreter. This information source and its construction cost are distinct from a learned reward predictor or a four-label audit.

Each of four arms starts afresh from the fixed Qwen3-4B base and performs actual LoRA training for 64 steps using 1,024 generated responses. The arms are public, random-four, diagnostic-four, and complete-v4 rewards. Seeds 944190 and 944192 form two paired runs, with the second keeping the trainer, task collection, testing rules, length, and endpoints fixed. The first result activates the single prespecified replication, and no third seed or parameter search is used to improve the reported endpoint. The model never uses execution feedback to rerank final greedy outputs.

## G.2 Model, sampling, and exact optimization objective

All arms use `Qwen/Qwen3-4B` at the fixed revision

```
1cfa9a7208912126459214e8b04321603b3df60c
```

The base is loaded in bfloat16 with SDPA attention and no quantization. Rank-8 LoRA adapts the query and value projections in all applicable layers, with scaling parameter 16, zero dropout, no bias adaptation, and default initialization. Each saved adapter contains 2,949,120 parameters stored in float32. Gradient checkpointing uses `use_reentrant=False`, with input gradients enabled. Gradient forwards disable the cache and generation enables it.

AdamW parameters not explicitly overridden in the script use the recorded PyTorch version's defaults. The prompt is one user message requesting only a Python implementation named `solve`, retaining needed imports but excluding example calls and tests. It includes the specification and one public example, uses the model's chat template with `add_generation_prompt=True` and `enable_thinking=False`, and does not reveal reference code, private tests, or failure logs. The complete-reward arm exposes full training tests only to its scorer.

The tokenizer uses left padding with the EOS token as padding. Generation honors all configured stop IDs from the pinned model assets. The saved answer ends at the first stop token, including that token but excluding later padding. Observed training stops are 151645, without inferring that the configuration allows only this ID. A separately saved complete generation configuration is not present

Table G.1: Actual controlled policy-training configuration. These values are reconciled with the shared trainer, per-run protocols, and saved adapter configurations.

| Item | Setting |
|---|---|
| Optimizer | AdamW, learning rate $2 \times 10^{-5}$, weight decay zero, no scheduler or warmup |
| Gradient clipping | Global norm 1.0 after accumulating all 16 response losses |
| Sampling | Two prompts per step, uniform with replacement from 320 specifications, eight fresh responses per prompt |
| Prompt RNG | NumPy `default_rng(seed+801)`, shared schedules within a paired seed |
| Training | 64 optimizer steps, 1,024 responses per arm, one update per fresh batch |
| Reference penalty | Coefficient 0.01, sampled $k_3$ form with token log-ratio clamped to $[-10, 10]$ |
| Generation | Temperature 1.0, top-$p$ 1.0, top-$k$ 0, maximum 384 new tokens |
| Evaluation | One greedy response per specification, batch size eight, no test-based reranking |
| Seeds | 944190 and 944192, fixed endpoint at step 64 |
| Hardware | One NVIDIA RTX A5000 per main run |

in the original result directories, so exact reconstruction requires the pinned tokenizer and model assets.

For group $g \in \{1, 2\}$ and response $j \in \{1, \ldots, 8\}$, let

$$A_{gj} = r_{gj} - \frac{1}{8} \sum_{k=1}^{8} r_{gk}$$

be detached. Write $L_{gj}$ for saved answer length, $h_{gjt}$ for the prompt and answer prefix, and

$$\delta_{gjt} = \mathrm{clip}_{[-10,10]} \left( \log \pi_0(y_{gjt} \mid h_{gjt}) - \log \pi_\theta(y_{gjt} \mid h_{gjt}) \right). \tag{G.1}$$

The actual loss minimized each step is

$$\mathcal{L}(\theta) = \frac{1}{16} \sum_{g=1}^{2} \sum_{j=1}^{8} \left[ -\frac{A_{gj}}{L_{gj}} \sum_{t=1}^{L_{gj}} \log \pi_\theta(y_{gjt} \mid h_{gjt}) + \frac{0.01}{L_{gj}} \sum_{t=1}^{L_{gj}} \left( e^{\delta_{gjt}} - \delta_{gjt} - 1 \right) \right]. \tag{G.2}$$

Logits and log-softmax are evaluated in float32. Sixteen response losses are backpropagated sequentially before clipping and a single optimizer step. Each response also receives a reference forward with the adapter disabled, giving 1,024 reference forwards per arm. There is no standard-deviation division, probability-ratio clipping, old-policy ratio replay, repeated optimization epoch, or terminal radial KL matching. The reward-surrogate score in Eq. (3) is therefore

$$\psi_{gj} = L_{gj}^{-1} \sum_{t} \nabla_\theta \log \pi_\theta(y_{gjt} \mid h_{gjt}).$$

It is length-normalized, not the unnormalized sequence policy gradient. The sampled and clamped reference term is not an exact KL between sequence distributions constraint. This specification makes the implementation distinct from a default PPO-clipped GRPO training recipe.

### G.3 Evaluation, test definitions, and execution limits

The complete data contain 320 training, 200 in-domain, and 200 compositional specifications. Python `random.Random(944191)` selects 16 training specifications for auxiliary endpoint generation and 64 from each development set. Each of the eight runs saves 144 endpoint responses, while the principal table uses the shared 128 development responses. The 16 auxiliary responses are not additional held-out evidence.

The fixed v4 utility requires one public test and a suite of 41 private tests to pass. Random-four and diagnostic-four first check public, then execute up to four additional inputs in one sandbox process with early exit. Random test construction uses seed $944171 + \mathrm{task_id}$, list lengths 0–15, keys $-3$ to 3, values $-12$ to 16, and binary flags. Diagnostic rules cover empty input, duplicate-key/filter conflicts, equal retained values with different-key order, and threshold/modulo boundaries. All

expected outputs come from the domain interpreter. Test validity therefore uses domain knowledge rather than an oracle-free generic generator.

Program execution uses a Linux bubblewrap sandbox with filesystem and network namespace isolation, a two-second CPU limit, one-GiB address-space limit, five-second outer wall timeout, and `PYTHONHASHSEED=0`. These limits are part of the operational utility. The model and scoring environments are Python 3.12.13, PyTorch 2.8.0+cu128, Transformers 4.57.3, PEFT 0.19.1, and NumPy 1.26.4. Historical cached-regression analysis uses Python 3.8 and NumPy 1.24.4.

Evaluation uses the same 128 development specifications in both seeds: 64 in-domain specifications spanning 49 structure signatures and 64 compositional specifications spanning 40 signatures absent from training. Related specifications are not 128 independent semantic tasks. No new-task statistical superiority or noninferiority claim is inferred from the two training seeds.

### G.4 Offline feedback diagnostics

Before interpreting a training effect, we check that the diagnostic inputs reveal errors in actual model responses beyond an equal-count random suite. The existing C-V and C-U histories allow this comparison on identical programs with fixed complete-v4 labels (Table G.2). This controls response variation in the offline analysis; the later training trajectories need not remain the same. These are

Table G.2: Offline behavior information on recorded histories. Zero observed rejection of $U$ positives does not guarantee zero rejection outside these programs.

| History | Tests | Hidden-error recall | $U$-positive rejection | Skill | Concordance |
|---|---|---|---|---|---|
| C-V | Random four | 39.23% | 0% | 29.13% | 0.6942 |
| C-V | Diagnostic four | 95.00% | 0% | 89.00% | 0.9482 |
| C-U | Random four | 41.23% | 0% | 25.07% | 0.6652 |
| C-U | Diagnostic four | 95.18% | 0% | 91.15% | 0.9558 |

development diagnostics on existing responses, not independent confirmations before training. The random tests contain useful signal and remain the primary comparator with the same test count. Reference suites are checked across the training collection, including 640 reference-suite runs for 320 specifications and both four-test methods.

### G.5 Both training seeds and paired changes

Table 3 gives all split-level endpoints. Paired repair/break counts below distinguish a net gain from merely preserving different subsets of the base model's successes.

Relative to public reward, diagnostic-four repairs/breaks 17/7 programs in the first seed and 12/3 in the second. Relative to random-four, these counts are 13/8 and 11/2. Net gains over random are $5/128 = 3.90625$ pp and $9/128 = 7.03125$ pp. Against public they are $10/128 = 7.8125$ pp and $9/128 = 7.03125$ pp. Both splits improve over public in both seeds. Complete-v4 improves over diagnostic-four by 2/128 and 6/128, so the experiment does not establish equivalence to complete feedback.

As a separate strong-supervision reference, expert-program SFT reaches 200/200 on each of the larger related in-domain/composition collections, with three training seeds for the 80-program condition. Later 20/40-program conditions also saturate these original domains at the same answer-token budget. These models use reference program supervision, not the same resources as the four reward arms. Their success remains a relevant alternative, without treating SFT and RL as a comparison matching all resources.

### G.6 Sensitivity to related specifications

A sensitivity analysis uses the saved complete-v4 endpoint labels and the original operation-structure signatures. The 128 specifications span 89 signatures: 49 in-domain and 40 held-out composition signatures. We retain the original specification-weighted endpoint, analyze the two seeds separately, and

report two descriptive sensitivities. First, each signature receives equal weight after averaging its own specification differences. This changes the evaluation weighting and is not a replacement for the main endpoint. Second, we remove each complete signature in turn and recompute the gain on all remaining specifications. Thus no single signature accounts for the entire overall advantage in either seed. This

Table G.3: Diagnostic-four minus the stated baseline on saved endpoints. The signs count signatures with positive, zero, or negative net contributions. Gain columns are pp. Removal ranges are descriptive sensitivity ranges, not confidence intervals.

| Seed | Baseline | Specification gain | Signature gain | Signs $(+, 0, -)$ | Removal range |
|---|---|---|---|---|---|
| 944190 | Public | 7.8125 | 7.3034 | $14, 69, 6$ | $[5.6000, 9.5238]$ |
| 944190 | Random | 3.9063 | 3.5581 | $10, 72, 7$ | $[1.6000, 5.5556]$ |
| 944192 | Public | 7.0313 | 5.9925 | $8, 79, 2$ | $[5.5556, 7.9365]$ |
| 944192 | Random | 7.0313 | 6.5543 | $8, 79, 2$ | $[5.5556, 7.9365]$ |

does not make all subsets equally strong. In the first seed, the in-domain diagnostic-minus-random gain is only $1/64$, and its removal range is $[0, 4.8387]$ pp. The corresponding composition range is $[1.6393, 8.0645]$ pp. In the second seed these ranges are $[4.7619, 8.0645]$ and $[4.8387, 9.5238]$ pp. The bundled minimal input retains all five methods' endpoint labels, and the analysis output lists every signature, including zero and negative contributions. It adds no model runs or independent tasks and does not turn the shared development collection into a population-level confirmation.

### G.7 Signal along the new trajectories

The offline histories cannot establish whether diagnostic information remains useful on programs produced during the new training runs. We therefore examine those trajectories after optimization, separately querying complete labels for the public, random, and diagnostic arms. These labels do not enter their optimizers, while the complete-reward arm reuses labels it already accessed. On diagnostic-four's own new responses, hidden-error recall is $95.1807\%$ and $93.75\%$ in the two seeds, with zero observed rejection of finite-$U$ positives. Its diagnostic-group skill is $90.8197\%$ and $80.7107\%$, compared with random-four's $33.7580\%$ and $28.5285\%$ on the respective random trajectories.

Among the 26 and 33 groups where public feedback is constant but trusted labels differ, only 2 and 5 remain silent under diagnostic feedback. This is the response distinction needed for a nonzero centered reward signal. Hidden-error recall also stays high in both halves of each run: $94.2029/96.3964\%$ and $92.6966/94.9367\%$. Because both tasks and policies change, these observations do not isolate policy drift or prove that the restored signal causes the full endpoint gain. They use fresh complete labels, not a reused old moment. The additional analyses re-evaluate 2,566 and 2,593 distinct task–programs within each analysis batch and execute 4,896 and 4,917 suites; their costs are separate from training.

### G.8 Information and computation cost

The utility gain over random tests is an information-quality result at equal test count. To assess the separate efficiency question, Table G.4 compares diagnostic and complete verification using the same recorded timing definitions. Generation and optimization remain in both workflows, so reducing executed inputs need not reduce total time proportionally. Fewer inputs do not imply proportional end-to-end savings: policies produce different programs and lengths, and the second diagnostic trainer run is slower. Initial test construction saves 2,560 expected outputs for two four-test sets across 320 tasks, although the implementation makes 5,120 computations of interpreter outputs because of repeated construction work. Both semantic knowledge and computation belong in the accounting. Test assets are reused, but every newly executed program produces a fresh behavior observation.

## H Matched-feedback updates and limits of transfer

Section 6.2 asks whether matched feedback summaries determine the outcome of a neural update procedure. Its interpretation depends on which aspects of that procedure were actually controlled. We

Table G.4: Recorded diagnostic/complete costs. Input counts are upper bounds including public checks and caching, not exact executed assertions. Training-loop time includes generation, reward execution, reference/policy forwards, backpropagation, optimization, and some record writing. Trainer wall time adds loading, adapter export, and endpoint generation, but excludes separate endpoint scoring, post-hoc trajectory scoring, test construction, and queueing. Each value is one recorded run, not a repeated timing estimate.

| | Seed 944190 | | Seed 944192 | |
|---|---|---|---|---|
| Measure | Diagnostic | Complete | Diagnostic | Complete |
| Test-input upper bound | 4,493 | 37,138 | 4,480 | 37,199 |
| Training-loop time (s) | 1,982.13 | 2,035.12 | 2,094.70 | 1,976.98 |
| Trainer wall time (s) | 2,295.10 | 2,353.04 | 2,423.12 | 2,281.15 |

therefore document the matched identities and errors, the failed common-norm calibration, and the completed prefix-calibrated comparison before giving the exact implementation. The final subsection examines a different boundary: whether the controlled diagnostic-test benefit transfers to natural code when public information and evaluator validity are accounted for.

### H.1 Matched feedback, calibration, and independent utility

The diagnostic-test training study changes both the information in feedback and its error rate. The matched-feedback intervention removes that ambiguity by fixing a response batch and its group confusion matrices. Its two versions differ only in step calibration: the first stops at its guardrail, while the second reaches utility evaluation. Keeping both records is necessary to separate a calibration failure from evidence about post-update utility.

*Version 1, a stopped common-norm update.* This follow-up to Section 6.3 fixes a response batch and matches the two feedback vectors' confusion matrix in every group, rather than only their overall accuracy. A calibration batch contains 32 prompts with eight responses each. It fixes a direction $v$ and parameter step $h = 0.26$, whose full-vocabulary forward KL on those prefixes is $0.001033417$ nat/token, within the specified tolerance of the $0.001$ target. Three further batches each contain 64 groups of eight responses. Each exports updates for $r^+$, $r^-$, and complete $U$, all with the same Euclidean displacement norm. The signs name larger and smaller projections onto the calibration direction, not a known ordering of final utilities.

A separate counting check verifies exact equality of TP/FP/TN/FN, reward means, variances, and centered MSE in all 192 groups. Repeated identities have consistent labels and feedback. The two vectors differ in all groups. Their accuracies across batches are $77.1484\%$, $76.9531\%$, and $77.3438\%$, with mean centered MSE $0.18774414$, $0.18603516$, and $0.17968750$. Reading the nine exported adapters against their intended updates gives zero per-parameter discrepancy and displacement norms near $0.2599999904$, with maximum between-arm difference about $1.7 \times 10^{-10}$.

Table H.1: Stopped matched-feedback experiment. Full-vocabulary forward KL on fixed calibration prefixes, in nat/token. The prespecified cap is $0.01$. Candidate utility evaluation was not launched.

| Batch | $r^+$ | $r^-$ | Complete $U$ |
|---|---|---|---|
| 1 | 0.00293880 | 0.03592896 | 0.00121843 |
| 2 | 0.00635341 | 0.07935648 | 0.00516508 |
| 3 | 0.00179070 | 0.02209027 | 0.09449751 |

All three batches violate the guardrail, including one complete-reward update (Table H.1). The candidate generation and utility-scoring jobs were cancelled by the dependency rule. Only base was evaluated, with 317/512 sampled programs and 77/128 greedy programs passing. These are not updated-policy outcomes. Version 1 therefore supplies no positive or zero-effect finding for its primary endpoint.

The stop exposes a calibration limitation. An equal Euclidean step calibrated along one direction need not control policy change along other directions. The recorded KLs concern fixed prefixes, not the full generated sequence law. No Hessian or finite-difference curvature measurement was made, so the data do not separate local anisotropy from finite-step nonlinear effects. The complete-reward gradient's cosine with the calibration direction is $0.3304$, $0.2175$, and $0.0163$ across the batches. The calibration projection is therefore not a guarantee of agreement with each new batch's trusted gradient, nor is either quantity an independent finite-policy utility test.

The records account for 1,792 construction responses, 1,715 distinct construction task-program labels, and 1.973611 allocated GPU-hours including smoke, cancelled jobs, and the stops. With base evaluation the unique recorded task-program count is 2,322. This is a controlled full-information intervention, not a low-label method. The original feedback, gradients, adapters, guardrail failures, and stop decision are retained. Version 2 below uses those directions rather than replacing failed batches.

*Version 2, completed calibration and utility evaluation.* The follow-up retains all original feedback assignments and all nine directions. Each generated candidate passes the prescribed four-position prefix forward-KL interval $[0.00095, 0.00105]$ nat/token. The previous $h = 0.26$ KL endpoints replay exactly. Saved parameter displacements and generation adapters agree with the globally frozen candidate hashes, repeated evaluation programs have consistent utility, and the full fixed-batch gradient identity is checked. These checks concern the exported candidates and saved analysis, rather than a second set of policy updates.

After calibration, independent generation supplies the endpoint that the constructed gradient projections cannot establish. Table 2 gives the 512-response outcomes on 128 common tasks, using the unchanged base cache. The primary contrast is $131/1536 = 0.085286\ldots$, an equal-batch average over shared tasks rather than 1,536 independent tasks. The paired-task bootstrap interval $[4.9479, 12.3047]$ pp and three positive batch contrasts meet the prespecified directional criterion. This approximate interval does not quantify resampling of new training batches; individual and secondary comparisons are not multiplicity-corrected.

Base and full-$U$ comparisons distinguish separation between assignments from harm or superiority to trusted feedback. The $r^+$ changes from the cached base are $+7.421875$, $+4.882813$, and $+9.960938$ pp. The $r^-$ changes are $-1.171875$, $-0.976563$, and $-1.171875$ pp, each with a reported interval containing zero. Full $U$ has a positive sampled-utility point change in each batch, with an equal-batch increase of $8.723958$ pp. These results do not establish that $r^-$ causes harm or that $r^+$ exceeds complete utility feedback.

Table H.2: Secondary greedy endpoint for v2, each entry is a success count among 128 common tasks. The base cache is 77/128. The pooled $r^+ - r^-$ contrast is $10.417$ pp, with paired approximate 95% bootstrap interval $[4.427, 16.406]$.

| Batch | $r^+$ | $r^-$ | Full $U$ | Difference (pp) |
|---|---|---|---|---|
| 1 | 89 | 75 | 95 | 10.9375 |
| 2 | 91 | 78 | 92 | 10.1563 |
| 3 | 92 | 79 | 81 | 10.1563 |

*Calibration and trajectory diagnostic.* The utility contrast alone cannot show that policy-change magnitude has been controlled. Calibration samples four prescribed positions; to inspect what this omits, a separate diagnostic sums full-vocabulary conditional KL along 16 frozen base-policy trajectories. Their mean summed KLs range from $0.182461$ to $4.275528$ nat across the candidates, and $r^+$ is larger than $r^-$ in every batch. These histories are independent of the utility evaluation but overlap the calibration sample. The diagnostic is a finite-horizon Monte Carlo estimate, not exact population sequence KL or an independent calibration validation.

The observed outcome difference is conditional on the prescribed feedback-to-update procedure. Equal empirical group statistics and compliance with its prefix-KL tolerance coexist with different utility, while the trajectory diagnostic leaves the extent and location of policy change coupled to the intervention. The theoretical comparison fixes the verifier and candidates and varies hidden utility.

This experiment instead fixes utility, changes feedback, and constructs new candidates, without matching their sequence KL.

*Sources, precision, and costs.* The archived execution records document calibration, global freezing, trajectory diagnostics, nine candidate generations and scores, analysis, and independent arithmetic checks. No evaluation samples or training seeds were added after observing the results. Exact integer recomputation gives 54 positive, 26 negative, and 48 zero pooled per-task differences. A machine-precision residual had instead assigned one zero to the negative category in the frozen descriptive JSON. The principal effect, intervals, and prespecified decision are unchanged, and the historical JSON is not overwritten.

Version 2 uses 4.258056 allocated GPU-hours and 17.225000 allocated CPU-core-hours. Its ledger includes 1,440 calibration response forwards, 1,536 gradient-verification backwards, 160 transformer forwards for the trajectory diagnostic, 1,170,433 generated candidate tokens, and 10,468 evaluator executions. The v1 allocation of 1.973611 GPU-hours is separate. One six-second CPU analysis test failed by comparing a $2.8 \times 10^{-17}$ residual with literal zero, and only that test assertion was corrected. The frozen scientific analysis was unchanged. All 191 checked v1 files remain byte-identical.

**Executed policy and update.** The model and tokenizer are Qwen3-4B at revision `1cfa9a7208912126459214e8b04321603b3df60c`. LoRA trains only the query and value projections (rank 8, scaling parameter 16, dropout 0, no bias), with 2,949,120 trainable parameters. Initialization uses seed 947250. Base weights, adapters, and logits use FP32; attention uses PyTorch SDPA, TF32 is disabled, and gradient checkpointing is non-reentrant. The executed environment is Python 3.12.13, PyTorch 2.8.0+cu128, Transformers 4.57.3, PEFT 0.19.1, and NumPy 1.26.4, on an NVIDIA RTX A6000. The model, tokenizer, generation-config, and initial-adapter hashes are recorded in the source manifest. Weight hashes were checked against the model cache during v2; v1 had recorded the revision rather than independent historical hashes of all base weights.

For response $y_{gi}$, let $\ell_{gi}(\theta) = \sum_t \log \pi_\theta(y_{gi,t} \mid x_g, y_{gi,<t})$ include every saved answer token through the first EOS, or through the length limit if no EOS occurs. Prompt tokens do not contribute to this sum, and there is no answer-length normalization. With advantages held fixed, the ascent direction is the gradient of

$$L_r(\theta) = \frac{1}{64 \cdot 8} \sum_{g=1}^{64} \sum_{i=1}^{8} (r_{gi} - \bar{r}_g)\ell_{gi}(\theta), \qquad g(r) = \nabla L_r(\theta_0), \qquad d_r = \frac{g(r)}{\|g(r)\|_2}. \tag{H.1}$$

Per-response score gradients are computed in FP32 and accumulated in FP64; normalized directions are stored in FP32. Each export assigns $\theta_0 + hd_r$ directly from the same initial trainable tensors. There is no Adam step, clipping, iterative optimization, or added KL penalty in this update. Thus the experiment changes the location of reward errors in one fixed batch, rather than comparing diverging on-policy training trajectories.

**Data, feedback construction, and identity constraints.** The calibration set $A$ contains 32 tasks and 256 responses from the frozen 320-task record-transformation domain. Its complete-$U$ mean-centered gradient, normalized to unit length, defines $v$. Each $B_b$ contains 64 tasks and 512 responses; no task repeats within a batch, and each $B_b$ is disjoint from $A$. The three $B$ sets share 18, 9, and 19 tasks for pairs $(1, 2)$, $(1, 3)$, and $(2, 3)$ respectively. Generation seeds are 947252 for $A$ and 947270–947272 for $B_1$–$B_3$. The exact ordered task lists, rather than seeds alone, determine the joins in the supplied `v1/inputs/splits.json`.

For a response score $s_{gi} = \nabla \ell_{gi}(\theta_0)$, its construction priority is $v^\top s_{gi}$. Within each group the script enumerates binary assignments to distinct task–program identities; repeated identities receive one common label. The nominal false-positive and false-negative quotas are one quarter of the negative and positive response counts, respectively. Each quota is rounded to the nearest attainable integer under those identity constraints, choosing the lower count on a tie. Among assignments with these same two quotas, $r^+$ maximizes and $r^-$ minimizes $\sum_i r_{gi} v^\top s_{gi}$; equal objectives are resolved by lexicographic order of the response-level binary vector. No group is removed because it lacks an attractive assignment. Identity uses the task ID and SHA-256 of the extracted program. True labels are consistent across saved construction responses; feedback consistency is enforced within each batch, not asserted to define a single verifier across different batches.

The per-batch aggregate $(\mathrm{TP}, \mathrm{FP}, \mathrm{TN}, \mathrm{FN})$ counts are $(216, 51, 179, 66)$, $(229, 47, 165, 71)$, and $(259, 37, 137, 79)$ for either feedback. Equality also holds separately in every one of the 192 groups. The supplied CPU replay rechecks the confusion counts, means, variances, centered MSE, duplicate constraints, and extrema from the saved priorities. These controlled assignments consume all construction labels; their names describe the prescribed projection, not an independently learned ordering of utilities.

The frozen evaluation set $E$ has IDs 950000–950127 and one task for each of 128 distinct complete operation signatures, 59 of which also occur in the training domain. Its specifications and finite tests were amended before the retained base evaluation to remove duplicate specifications, as documented in `EVALUATION_AMENDMENT.md`; v2 uses exactly this amended set. Distinct signatures are not a claim of independent natural-code tasks. The original and amended records are preserved, and no task or evaluation sample was changed in response to v2 candidate utility.

**Four-position forward KL and deterministic calibration.** The probe comprises 16 saved $P_0$ responses from $A$, with zero-based response indices $\mathrm{linspace}(0, 248, 16, \mathrm{dtype} = \mathrm{int})$ in the ordered 256-response array. For a probe with $n$ prompt tokens and $L$ retained answer tokens, the executed hidden-state positions are

$$S = \mathrm{unique}\big(\mathrm{linspace}(n-1, n+L-2, 4, \mathrm{dtype} = \mathrm{int})\big). \tag{H.2}$$

Position $n-1$ predicts the first answer token; position $n+L-2$ predicts the last retained answer token. The manifest includes each probe's token IDs, lengths, and actual indices. Teacher forcing uses an unpadded prompt–answer sequence. Full-vocabulary FP32 log-softmax gives $p_0(\cdot \mid H_s)$ and $p_\theta(\cdot \mid H_s)$, and the statistic is

$$K_{\mathrm{prefix}}(\theta_0, \theta) = \frac{1}{16} \sum_{j=1}^{16} \frac{1}{|S_j|} \sum_{s \in S_j} \sum_{z} p_0(z \mid H_{js}) \log \frac{p_0(z \mid H_{js})}{p_\theta(z \mid H_{js})}. \tag{H.3}$$

The vocabulary and position reductions follow the original FP32 implementation, followed by the NumPy mean over responses. Sampling temperature is 1 with no top-$k$ or top-$p$ truncation. The quantity is a mean over these prescribed positions, in nat/token, rather than a mean over all generated tokens or a sequence-distribution KL.

For each retained direction, the grid $h \in \{0, 0.0325, 0.065, 0.13, 0.195, 0.26\}$ is evaluated from $\theta_0$. The first adjacent crossing of $0.001$ brackets a deterministic midpoint search, with at most 24 bisections and acceptance interval $[0.00095, 0.00105]$. The upper bracket endpoint is accepted immediately if it lies in this interval. The baseline must have absolute KL at most $10^{-6}$; replay at $0.26$ must agree with v1 within $\max(10^{-6}, 10^{-4}|K_{\mathrm{v1}}|)$; a local decrease larger than $10^{-7}$, missing crossing, nonfinite result, or exhausted search stops the entire primary comparison. Every accepted candidate must also satisfy the hard cap $0.01$. No candidate utility enters this search. All nine saved states were reloaded, their parameter deltas checked, and their prefix KL recomputed before the global candidate freeze.

Table H.3: Executed v2 candidates. Prefix KL is the statistic in (H.3), in nat/token. Diagnostic summed KL averages all answer-history conditional KLs over the same 16 base trajectories, in nat/sequence. Different $h$ values are required for prefix matching; the final column shows that this does not match sequence KL.

| Batch | Arm | Step $h$ | Achieved prefix KL | Diagnostic summed KL |
|---|---|---|---|---|
| 1 | $r^+$ | 0.13406250 | 0.0010098635 | 2.043123 |
| 1 | $r^-$ | 0.07515625 | 0.0009982308 | 0.889451 |
| 1 | Full $U$ | 0.23562500 | 0.0009952055 | 2.244657 |
| 2 | $r^+$ | 0.13000000 | 0.0010488921 | 1.920124 |
| 2 | $r^-$ | 0.05484375 | 0.0010428602 | 0.252927 |
| 2 | Full $U$ | 0.13812500 | 0.0009767594 | 0.735630 |
| 3 | $r^+$ | 0.18687500 | 0.0010237916 | 4.275528 |
| 3 | $r^-$ | 0.08531250 | 0.0009591594 | 1.043787 |
| 3 | Full $U$ | 0.04875000 | 0.0010264919 | 0.182461 |

**Saved directions and full-history diagnostic.** The normalized directions are taken from the original saved gradients, not inferred by dividing a recalibrated adapter by $0.26$. The largest reconstruction discrepancy between a v1 delta divided by $0.26$ and its stored direction is below $4 \times 10^{-9}$ per coordinate. Every v2 export reload has zero recorded per-parameter discrepancy from its intended update. Exact displacement norms, direction hashes, adapter hashes, all grid values, brackets, and forward counts appear in `calibrated_steps.json` and `CANDIDATES_FROZEN.json`. The protocol freeze precedes the global candidate freeze, which precedes updated-candidate $E$ generation.

Writing $\Psi_g$ for the matrix of response score rows, v2 also replays

$$g(r^+) - g(r^-) = \frac{1}{512} \sum_{g=1}^{64} \Psi_g^\top C_g (r_g^+ - r_g^-). \tag{H.4}$$

The full-vector relative errors are below $7 \times 10^{-7}$. The three gradient-difference norms are $7.9873$, $7.9739$, and $8.3753$; the cosines between the $r^+$ and $r^-$ gradients are $-0.6727$, $-0.5715$, and $-0.6713$. Cosines with $g(U)$ are $(0.4706, 0.0970)$, $(0.4465, 0.1539)$, and $(0.1180, 0.3597)$ for $(r^+, r^-)$ in the three batches. This confirms the saved construction geometry, not independent evidence of efficacy: batch 3 even has the larger trusted-gradient cosine for $r^-$. The independent endpoint remains the generated-program utility.

The optional trajectory calculation uses exactly the same 16 base responses as calibration, comprising 3,139 retained answer tokens. It evaluates every answer-prediction position, including the first EOS when present and retaining length-truncated responses through the fixed horizon 384. For each candidate it computes full-vocabulary $\mathrm{KL}(p_0 \| p_\theta)$ at those histories, sums conditional KL within each trajectory, then averages the 16 sums with equal trajectory weight. Logits and vocabulary reductions are FP32; the resulting per-token scalars are converted to FP64 before summation. Sixteen-position head chunks preserve every position and vocabulary entry. Tiny negative conditional values from rounding are retained; the implementation rejects values below $-10^{-5}$. Token-weighted averages and all per-token and per-trajectory values are supplied separately. Recovering the four-position statistic from this calculation agrees within $2.90 \times 10^{-8}$. No step was retuned from the diagnostic. Its overlap with calibration and limited trajectory count preclude treating it as an independent sequence-KL matching test.

**Generation and fixed finite utility.** The prompt requests only a Python implementation of `solve`, appends the frozen specification and public example, and uses the model's chat template with `enable_thinking=False`. Stochastic generation uses temperature 1, top-$p = 1$, top-$k = 0$, four returned sequences per task, and two tasks per generation call. Greedy generation uses eight tasks per call and one sequence each. The maximum answer length is 384 tokens. Tokenizer padding is left-sided with ID 151645; generation stops at the first ID in $\{151645, 151643\}$, includes that token in the saved response, and discards later padding. No-EOS responses remain valid length-truncated samples. No test outcome reranks or filters generation.

The stochastic base seed is 947400; batch $b \in \{1, 2, 3\}$ uses $947400 + 10b + (0, 1, 2)$ for $(r^+, r^-, U)$. Greedy seeds add 10000. These are policy-specific streams, not token-level common-random-number coupling; comparisons are paired by task. The base generations and labels are reused byte-for-byte from v1. Saved response ordering is the task-list order, with four contiguous stochastic responses per task. The source bundle retains raw text, token IDs, termination flags, program hashes, and every binary label, including duplicates.

The evaluator extracts the first Python/`py` fenced block, or the full stripped response if none exists. It checks the single public assertion and, only if that passes, the concatenated 41 private assertions from the fixed v4 task definition. Code runs under a filesystem/network-isolated `bubblewrap` process, with CPU limit 2 seconds, address-space limit 1 GiB, file-size limit 1 MiB, 32 file descriptors, and an outer 5-second timeout. The environment is cleared and `PYTHONHASHSEED=0`; only the system runtime is exposed. There is no host-execution fallback. Exceptions and ordinary timeouts yield failure; sandbox-process failures abort the job. Repeated task–program labels are cached, while repeated sampled responses remain in the utility denominator. This is correctness under the frozen finite evaluator, not a claim that all possible program inputs have been verified.

**Paired-task uncertainty and source replay.** For policy $c$ and task $i$, let $\bar{U}_{ic}$ be the mean of its four saved stochastic labels. Each batch contrast is the mean of $\bar{U}_{i,b,+} - \bar{U}_{i,b,-}$ over the 128 tasks; the

primary contrast averages the three signed batch contrasts equally. The executed NumPy generator `default_rng(947299)` draws 20,000 arrays of 128 task indices with replacement. Each array is shared by every arm and all three batches. The 2.5th and 97.5th percentiles of these paired resampled contrasts give the reported intervals. This preserves the common task effects and shared base control; it does not count the four responses as independent semantic tasks. One complete structure signature per task makes the recorded structure-cluster sensitivity identical to this task bootstrap. Batch and secondary intervals are descriptive and unadjusted for multiple comparisons. A separate within-task Monte Carlo standard error is supplied in the analysis JSON; it is not the primary bootstrap interval.

The included `scripts/replay_v2_evidence.py` reads the original per-response label records, reconstructs the signed contrasts, and reruns the original resampling configuration without generating or executing new programs. It reproduces the principal interval as well as the secondary comparisons and uses integer counts to distinguish exact zero task differences. The accompanying registry maps implementation settings and results to byte-hashed inputs. Actual protocol and candidate-freeze timestamps, environment pins, commands, and included-versus-archive-only boundaries are documented in `data/v2_sources/README.md`. Base weights, saved gradient tensors, and adapter tensors are archive-only; their hashes and verification results are included. Thus the distributed CPU replay verifies the empirical controls and reported arithmetic, while reproducing GPU forwards requires those separately identified assets. CPU replay verifies these saved records without adding model runs or changing the historical outcomes.

### H.2 Limits of transfer to natural-code tasks

Bounded feasibility checks on HumanEval+/EvalPlus (Liu et al., 2023) and TACO (Li et al., 2023) did not establish a natural-code training benefit. In HumanEval, a selector improved coverage of reference-code mutations, but its four chosen tests and four random tests assigned identical rewards to all 320 model responses. The proxy errors covered by the selector did not coincide with the model's residual errors.

The initial TACO comparison exposed an information-interface problem: six of eight additional detected errors could already be identified from public examples omitted by the baseline verifier. Incorporating those examples reduced the residual advantage to two responses from one task and revealed five programs accepted by the original finite evaluator despite violating a public example. A subsequent full-visible comparison on new tasks still failed its cross-task criterion. A targeted validity audit then found illegal test inputs and legal boundary errors missed by the finite evaluator. The original scores and failure decisions were retained; none of these studies launched further policy training.

These results limit the interpretation of Section 6.3. Diagnostic tests must be compared with a baseline that uses all available public information, and both test legality and evaluation coverage must be established before a reward difference is treated as an improvement. The controlled interpreter supplies that domain-specific information, which is part of its supervision cost. The two-seed controlled result therefore supports the value of diagnostic feedback in its stated domain, without establishing general natural-code transfer or end-to-end savings.