WHICH VERIFIER ERRORS REACH THE GRADIENT? A MOMENT DECOMPOSITION FOR AUDITING GROUP-RELATIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-relative RL with verifiable rewards updates the policy with score-weighted *contrasts* of verifier rewards, yet verifiers are audited with the intraclass correlation (ICC) or effective sample size (ESS), statistics designed for a group *mean*. We characterize exactly which properties of verifier error reach the update. Casting the verifier-induced perturbation as a U-statistic, Hoeffding's variance formula decomposes its second moment into three channels: the within-group error variance, which a mean-square-within (MSW) audit estimates; a coupling between squared error and score norm; and a squared reward-hacking gradient, whose bias accumulates across a batch instead of averaging out. The ICC and ESS capture none of them. The decomposition covers length-normalized updates and shared verifier randomness, and yields an unbiased audit statistic. On two Qwen2.5 policies the MSW channel carries 0.90–1.00 of the per-group perturbation, so variance-based audits are a sound default; yet for a format-sensitive checker the systematic channel reaches 0.26 $[0.15, 0.37]$ of a 64-prompt batch update (post hoc). In a pre-specified three-seed audit with LLM judges, ESS ranked three verifiers in the reverse order of their gradient perturbation, while MSW ranked them correctly. These results give practitioners a principled, inexpensive audit target.

1 INTRODUCTION

Group-relative reinforcement learning with verifiable rewards (RLVR) is a standard recipe for post-training reasoning models (Shao et al., 2024; Guo et al., 2025). A prompt is answered by G sampled completions, a verifier scores them, and the scores become advantages by subtracting the group mean. Verifiers err, and not independently. Rule-based checkers reject equivalent answer formats (Huang et al., 2025a), model-based graders can be gamed (Helff et al., 2026), and different models make correlated mistakes (Kim et al., 2025). The natural summary of such dependence comes from survey statistics: the intraclass correlation (ICC) ρ of the within-group errors (Shrout & Fleiss, 1979) and the Kish effective sample size $\mathrm{ESS} = G/(1 + (G-1)\rho)$ (Kish, 1965). A concurrent audit of real GRPO groups reports $\hat{\rho} = 0.530$, i.e. $\mathrm{ESS} = 1.70$ at $G = 8$ (Xin, 2026).

The mismatch. An effective sample size describes the precision of a *mean*. The group-relative update discards the mean. It consumes the centered errors, weighted by the policy's score vectors (Figure 1a). The relevant question is therefore not how correlated the errors are, but *which moments of the error reach the second moment of the update*. Classical design-effect reasoning answers it only for isotropic scores.

This paper. The verifier-induced perturbation is a degree-two U-statistic, so Hoeffding's variance formula (Hoeffding, 1948) splits its second moment exactly into three channels: the within-group error variance that a mean-square-within (MSW) audit estimates, a coupling between squared error and score norm, and the squared reward-hacking gradient. An ICC or an ESS estimates none of them; an MSW audit estimates only the first. On two Qwen2.5 policies the first channel carries 0.90–1.00 of the per-group total (Figure 1b). The other two are small per group, but the hacking channel is a bias, which does not average out over the prompts of a batch; in one post hoc cell it becomes a substantial part of a batch update.

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

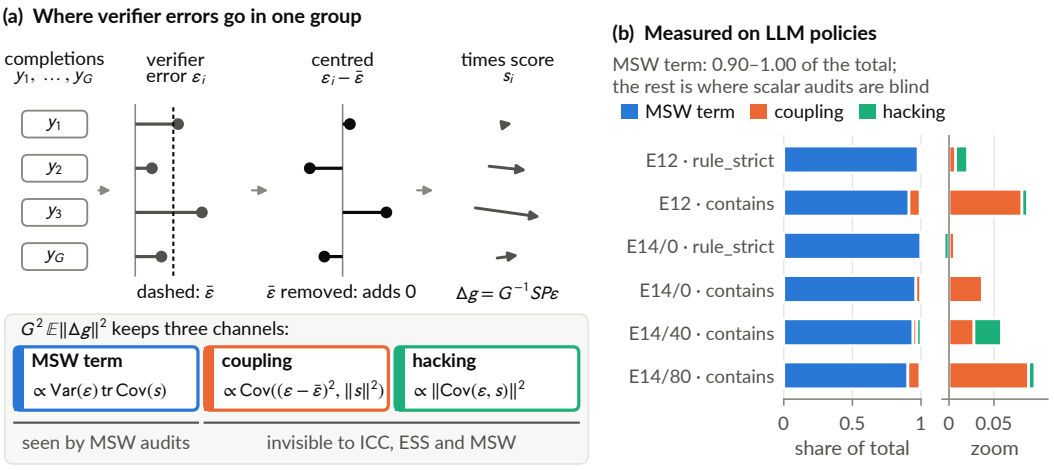

Figure 1: **The update consumes contrasts, so only three moments of the verifier error reach it, and a scalar audit sees at most one of them. (a)** Schematic (illustrative values, not data). Within a group the errors are centered by P, so their common part contributes nothing; the contrasts are then weighted by the completions' score vectors. Hoeffding's variance formula splits $G^2\mathbb{E}\|\Delta g\|^2$ into three channels (Theorem 2). **(b)** Per-group shares of the three channels on stored audits of two Qwen2.5 policies (post hoc, sequence-level operator, six audit · verifier cells, §5); right: the two channels an MSW audit cannot see, magnified. Segments left of 0 are negative point estimates of nonnegative quantities. Intervals are in Figure 3.

Contributions. **(1) Three audit channels** (§3). The decomposition is classical. We add its reading as audit channels, of which a scalar pipeline sees one; the centered-score treatment that makes it hold under length normalization; the functional for shared verifier randomness; and the batch-level share of the hacking bias. **(2) What classical audits capture** (§4). Common-mode error is annihilated, and the ESS reversal holds at fixed marginal variance but vanishes at fixed MSW. Both facts are classical, and we say where. **(3) Measurements** (§5–§6): the channels on stored audits of two policies under two operators, and a multi-seed audit with two LLM judges that tested four pre-specified predictions. Two held and two failed.

Scope and provenance. We establish no link between these quantities and training outcomes (§8). Empirical statements carry one of three labels, defined once: **pre-specified** = fixed before the run that tests it; **exploratory** = a study designed after the original simulation protocol was complete; **post hoc** = chosen after the relevant numbers were seen.

2 Setup: The Update, Three Verifier Regimes, and What an Audit Sees

Fix a prompt x and sample $y_1, \dots, y_G \sim \pi_\theta(\cdot \mid x)$. Let $s_i = \nabla_\theta \log \pi_\theta(y_i \mid x) \in \mathbb{R}^d$ be the score of completion i, $S = [s_1, \dots, s_G]$, $r \in \mathbb{R}^G$ the clean rewards, $\tilde{r} = r + \varepsilon$ the verifier's rewards, and $P = I_G - \mathbf{1}\mathbf{1}^\top/G$ the centering projector. The mean-centered update — REINFORCE with the group-mean baseline (Williams, 1992; Weaver & Tao, 2001), i.e. Dr. GRPO (Liu et al., 2025) — and its verifier-induced perturbation are

$$\tilde{g} = G^{-1}SP\tilde{r}, \qquad g = G^{-1}SPr, \qquad \Delta g = \tilde{g} - g = G^{-1}SP\varepsilon. \tag{1}$$

Standard GRPO (Shao et al., 2024) also divides by the group's reward standard deviation (§4) and normalises each completion's loss by its length $|o_i|$, which replaces s_i by $s_i/|o_i|$. We call the two operators `raw` (sequence-level) and `ln` (length-normalised).

Three verifier regimes. The error can depend on the completion and on the verifier's own randomness ω: a judge sample, a rubric, a reference solution. *(i)* A deterministic checker makes ε_i a function of y_i. *(ii)* A stochastic judge is called independently once per completion. *(iii)* Shared verifier randomness — one rubric or reference solution drawn per group — is the only source of error covariance across completions that remains once the completions are fixed.

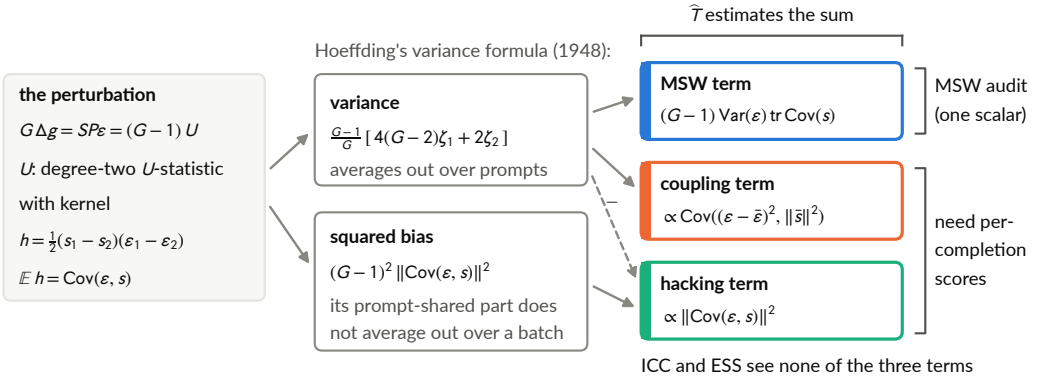

Figure 2: **The perturbation is a U-statistic, so its second moment is a variance plus a squared bias, which regroup into three audit channels.** Arrows show which part of Hoeffding's decomposition feeds which channel of (3). The dashed arrow is a negative variance correction, $-\frac{(G-1)(G-2)}{G}\|\mathrm{Cov}(\varepsilon, s)\|^2$: the hacking term is the squared bias minus a variance part. Brackets: an MSW audit (equivalently the ICC-implied baseline) estimates only the MSW term; $\widehat{T}$ estimates the sum; the split needs per-completion scores. ICC and ESS are scale-free and estimate none of the terms. Diagram only; no data.

Assumption 1 (Symmetric verifier, i.i.d. completions)**.** Given x there are a group-level randomness ω_0 and per-completion randomness $\omega_1, \ldots, \omega_G$ such that, conditionally on (x, ω_0), the pairs (y_i, ω_i) are i.i.d., and $\varepsilon_i = f(y_i, x, \omega_0, \omega_i)$, $s_i = s(y_i, x)$ for fixed measurable f, s, with finite second moments. Regimes (i)–(ii) have ω_0 trivial.

The assumption excludes verifiers that see the slot index or that build their reference from the sampled completions (majority vote). It does not require unbiased or independent errors.

What an audit sees. Scalar audits report one-way ANOVA summaries of the error matrix: the within-group mean square MSW, the ICC and the Kish ESS (Xin, 2026; Kohli, 2026; Miller, 2024). With gold labels one can instead measure the target directly. For m audited groups with observed errors $e_j = \tilde{r}_j - r_j^{\mathrm{gold}}$, let $z_j = S_j P e_j = \nabla_\theta \sum_i (e_{j,i} - \bar{e}_j) \log \pi_\theta(y_{j,i} \mid x_j)$, the usual surrogate gradient with advantages Pe_j, and

$$\widehat{T} = \frac{1}{m}\sum_{j=1}^{m} G^{-2}\|z_j\|^2. \tag{2}$$

$\widehat{T}$ costs one extra backward pass per group. It is the cluster-robust (CR0) sandwich meat (Liang & Zeger, 1986) with design $S_j P$, and it is unbiased for $\mathbb{E}\|\Delta g\|^2$ when the gold labels are error-free and the groups are drawn from deployment — immediately, since the errors are observed (Theorem 11, App. B.4). None of this is new. The question is what $\widehat{T}$ contains.

3 Which error moments reach the update

Write $\mu_s = \mathbb{E}[s \mid x]$ for the score mean and $\bar{s} = s - \mu_s$ for the centered score. For the sequence-level score $\mu_s = 0$ (the score identity). Under `ln` it is not, which is why centering matters below.

Theorem 2 (Three channels; Hoeffding's variance formula (Hoeffding, 1948))**.** *Under Assumption 1 with ω_0 trivial, and with no assumption on μ_s, writing $\bar{\varepsilon} = \mathbb{E}[\varepsilon \mid x]$,*

$$G^2\,\mathbb{E}\big[\|\Delta g\|^2 \mid x\big] = \underbrace{(G-1)\,\mathrm{Var}(\varepsilon)\,\mathrm{trCov}(s)}_{\textit{MSW term}} + \underbrace{\tfrac{(G-1)^2}{G}\mathrm{Cov}\big((\varepsilon-\bar{\varepsilon})^2, \|\bar{s}\|^2\big)}_{\textit{coupling term}} + \underbrace{\tfrac{(G-1)(G^2-2G+2)}{G}\,\|\mathrm{Cov}(\varepsilon, s)\|^2}_{\textit{hacking term}}. \tag{3}$$

For the sequence-level score the first two terms are exactly the diagonal part $\mathbb{E}\sum_i (P\varepsilon)_i^2\|s_i\|^2$ and the third is exactly the off-diagonal part. The same combination in five completion-level moments is (6) *(App. B).*

Why it is Hoeffding. Pathwise, $\sum_{i\neq j}\frac{1}{2}(s_i - s_j)(\varepsilon_i - \varepsilon_j) = G\,SP\varepsilon$. So $G\,\Delta g = (G-1)\,U$, where U is the degree-two U-statistic with kernel $h = \frac{1}{2}(s_1 - s_2)(\varepsilon_1 - \varepsilon_2)$, whose mean is $\mathrm{Cov}(\varepsilon, s)$ (Figure 2). Hoeffding's formula splits the second moment into a squared bias and a variance, $G^2\mathbb{E}\|\Delta g\|^2 = (G-1)^2\|\mathrm{Cov}(\varepsilon, s)\|^2 + \frac{G-1}{G}[4(G-2)\zeta_1 + 2\zeta_2]$, with ζ_1, ζ_2 the traces of the kernel's first- and second-order variances. Collecting terms gives (3). Two facts follow. The hacking term is the squared bias minus $\frac{(G-1)(G-2)}{G}\|\mathrm{Cov}(\varepsilon, s)\|^2$ of variance bookkeeping. And the score mean drops out, because P also annihilates a common shift of the *scores*: $\|SP\varepsilon\|^2 = \|(S - \mu_s\mathbf{1}^\top)P\varepsilon\|^2$ pathwise. Moments of the uncentered score miss an exact correction, nonzero even when $\varepsilon \perp s$ (App. D). Every coefficient is verified in exact rational arithmetic.

Reading the channels. If the error is independent of the score, only the MSW term survives. The *coupling* term is nonzero when errors are larger, or smaller, on completions with atypical score norms — for example, long ones. The *hacking* term is the squared reward-hacking gradient: $\mathbb{E}[\Delta g \mid x] = \frac{G-1}{G}\mathrm{Cov}(\varepsilon, s)$, and for the sequence-level score $\mathrm{Cov}(\varepsilon, s) = \nabla_\theta \mathbb{E}_{\pi_\theta}[\varepsilon]$, the direction in which the policy would increase its own verifier error (Proposition 19, App. F). Under Assumption 1 no slot-level geometry survives averaging over groups (Corollary 9, App. B): structure reaches the update only when it is tied to a *feature* of the completion, such as format, length or strategy.

Example: a format checker. A deterministic checker that rejects correct answers written in a non-canonical format has $\varepsilon(y) = -f(y)$ with $f = \mathbf{1}\{y \text{ correct, non-canonical}\}$ and $p = \Pr_\pi(f = 1)$. For the sequence-level score, (3) reads

$$G^2\mathbb{E}\|\Delta g\|^2 = (G-1)\,p(1-p)\,\mathbb{E}\|s\|^2 + \tfrac{(G-1)^2}{G}(1-2p)\mathrm{Cov}\big(f, \|s\|^2\big) + \tfrac{(G-1)(G^2-2G+2)}{G}\big\|\nabla_\theta \Pr_{\pi_\theta}(f=1)\big\|^2.$$

An ICC audit sees p and how f clusters by prompt. It does not see whether the rejected completions have atypical score norms, and it does not see the hacking gradient, which moves probability away from correct non-canonical answers — the format drift that the toy policy of App. P.1 exhibits. The coupling carries the factor $1 - 2p$. The ratio of an error-weighted to an unweighted mean score norm (A/B, Definition 8) therefore overstates it by $1/(1-2p)$ and has the wrong sign for $p > \frac{1}{2}$, so we report the coupling term itself.

Theorem 3 (Regime (iii): shared verifier randomness)**.** *Let $\varepsilon_i = \varepsilon^{(0)}(y_i) + \omega\,u(y_i)$ with ω independent of $(x, y_{1:G})$, $\mathbb{E}\omega = 0$, and u a second completion-level function. Then $G^2\mathbb{E}[\|\Delta g\|^2 \mid x]$ is the right-hand side of* (3) *for $\varepsilon^{(0)}$ plus $\mathrm{Var}(\omega)$ times the right-hand side of* (3) *for u.*

In this additive single-factor model, shared randomness reaches the update through the same three channels, applied to what is shared. A per-group leniency shift (u constant) raises the ICC and lowers the ESS but contributes *nothing*. A shared rubric that favours one answer format contributes $\mathrm{Var}(\omega)$ times that format indicator's own three terms. Plugging the total error into Theorem 2 would be wrong: it violates Assumption 1 with ω_0 trivial and silently drops part of the shared term.

Proposition 4 (What the ICC/MSW pipeline sees)**.** ***(a)** For groups $y_g = a_g\mathbf{1} + \varepsilon_g$ with i.i.d. group effects uncorrelated with the whole vector ε_g, $\mathbb{E}\varepsilon_g \propto \mathbf{1}$ and common covariance Σ with mean diagonal $\bar{d}$ and mean off-diagonal $\bar{o}$, $\mathbb{E}[\mathrm{MSW}] = \bar{d} - \bar{o}$ and the ANOVA ICC converges to $(\sigma_a^2 + \bar{o})/(\sigma_a^2 + \bar{d})$ (Shrout & Fleiss, 1979). **(b)** Under Assumption 1 (ω_0 trivial) the ICC-implied baseline $B_{\mathrm{ICC}} := G^{-2}\mathbb{E}[\mathrm{MSW}]\mathrm{tr}(P\bar{K})$, $\bar{K} = \mathbb{E}[S^\top S \mid x]$, equals G^{-2} times the MSW term of* (3) *for any score mean, and so does the pooled plug-in $R_{\mathrm{pool}} := G^{-2}\mathrm{tr}(P\,\mathbb{E}[\varepsilon\varepsilon^\top \mid x]\,P\bar{K})$. Hence $\mathbb{E}\|\Delta g\|^2 - B_{\mathrm{ICC}}$ is exactly the coupling plus the hacking term, and no function of* $(\mathrm{ICC}, \mathrm{ESS}, \mathrm{MSW})$*, as population functionals, can detect either.*

The batch level. The hacking term is where the bias lives, and a bias behaves differently from a variance when a batch averages B prompts. Let $\bar{\Theta} = G^2\mathbb{E}\|\Delta g\|^2$, let $b = \mathbb{E}_x G^2\|\mathbb{E}[\Delta g \mid x]\|^2$ be the per-prompt bias and $\mathcal{S} = G^2\|\mathbb{E}_x\mathbb{E}[\Delta g \mid x]\|^2 \leq b$ its part shared across prompts. The share of a

batch perturbation that is systematic given its B prompts is

$$\pi_B = \frac{b + (B-1)\mathcal{S}}{\bar{\Theta} + (B-1)\mathcal{S}} \leq \frac{b}{b + (\bar{\Theta} - b)/B}, \tag{4}$$

with equality iff every prompt has the same hacking gradient (App. D). A per-group share is π_1. The variance shrinks like $1/B$ and the shared bias $\mathcal{S}$ does not, so a hacking channel that is negligible per group can matter for a large batch. The cross-group U-statistic of the $\langle z_j, z_{j'} \rangle$ estimates $\mathcal{S}$ (Proposition 12, App. B.4).

4 What classical audits capture and miss

Conditionally on the sampled completions, the verifier's contribution is exact and needs no assumption on the error covariance. Write

$$\mathbb{E}[\varepsilon \mid S, r] = \mu, \quad \mathrm{Cov}(\varepsilon \mid S, r) = \Sigma_{S,r}, \qquad \Sigma_S := \mathrm{Cov}(\varepsilon \mid S) = \mathbb{E}[\Sigma_{S,r} \mid S] + \mathrm{Cov}(\mu \mid S), \tag{5}$$

and $K = S^\top S$. In regime (i) $\Sigma_{S,r} = 0$ and everything sits in μ.

Theorem 5 (Exact verifier-noise gradient covariance)**.** *With Σ arbitrary:* ***(i)*** $\mathbb{E}[\Delta g \mid S, r] = G^{-1}SP\mu$*;* ***(ii)*** $\mathrm{Cov}(\Delta g \mid S, r) = G^{-2}SP\Sigma_{S,r}PS^\top$*;* ***(iii)*** *with no assumption on the bias,* $\mathrm{Cov}(\Delta g \mid S) = G^{-2}SP\Sigma_S PS^\top$*;* ***(iv)*** $\mathbb{E}[\|\Delta g\|^2 \mid S, r] = G^{-2}\mathrm{tr}(PM_{S,r}PK)$ *with* $M_{S,r} = \Sigma_{S,r} + \mu\mu^\top$.

Corollary 6 (Common-mode annihilation)**.** *The kernel of $\Sigma \mapsto P\Sigma P$ on symmetric matrices is $\{\mathbf{1}a^\top + a\mathbf{1}^\top\}$ and meets the PSD cone in $\{\tau^2\mathbf{1}\mathbf{1}^\top\}$: a covariance injects no gradient noise for every S iff the error is purely common mode.*

Corollary 7 (The ESS reversal, in both parametrisations)**.** *Let $\Sigma = \sigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top]$.* ***(a)*** *At fixed marginal variance σ^2, $\mathrm{trCov}(\Delta g \mid S, r) = G^{-2}\sigma^2(1-\rho)\mathrm{tr}(PK)$ decreases strictly in ρ while $\mathrm{ESS}(\rho)$ also decreases: the two rank exchangeable verifiers oppositely.* ***(b)*** *At fixed within-group variance $\mathbb{E}[\mathrm{MSW}] = \sigma^2(1-\rho)$, $\mathrm{trCov} = G^{-2}\mathbb{E}[\mathrm{MSW}]\mathrm{tr}(PK)$ does not depend on ρ at all.*

The reversal is a statement about *which* scalar is held fixed. It is real between ESS and gradient noise at fixed σ^2, but an audit that reports MSW is not fooled by ρ. Both corollaries are the within-cluster mirror image of the Moulton factor $1 + (G-1)\rho$ for cluster means (Moulton, 1990; Cameron & Miller, 2015), and common-mode annihilation is the baseline invariance of REINFORCE; neither is new as statistics.

Operators. Under $S \to SD$ with an (S, r)-measurable diagonal D, as for `ln`, every pathwise statement holds verbatim. The population statements hold exactly with every moment taken from the centered length-weighted score $s/|o| - \mathbb{E}[s/|o|]$. On our audits the terms that uncentered moments omit are within $\pm 5.4\%$ of the total, and all shares we report use centered moments (App. B). A past-measurable preconditioner is covered like D; Adam only as $\beta_2 \to 1$, and gradient clipping not at all (Corollary 17).

The z-score divisor. Standard GRPO's division by the group standard deviation adds, to first order and for mean-zero errors with a finite normalised fourth moment, a second projection (App. B.3). Binary verifiers violate both hypotheses, and the predicted suppression is absent at every flip rate in the pre-specified simulations ($\mathrm{SR} = 0.74$–0.99 against a predicted 0, with up to 4.8% of groups degenerate and dropped). Those simulations verify every identity to 0.0176 relative Frobenius error. Of their 4 falsifiable hypotheses 1 passes, and all 5 literal failures are reported in App. K.

5 Measuring the channels on language-model policies

The audits. The *pretrained audit* (E12; exploratory, predictions pre-specified) samples `Qwen2.5-1.5B-Instruct` on the first 300 GSM8K test prompts at $G = 8$ and scores the completions with three deterministic string matchers: `rule_strict` (format false negatives), `contains` (false positives) and `cmshift` (a pure common-mode shift). Exact full-parameter scores are count-sketched to $k = 16{,}384$ (App. P.6). The *training-time audit* (E14, same status) trains `Qwen2.5-0.5B-Instruct` by GRPO for 80 steps on `rule_strict` and audits four checkpoints on 120 of those prompts (App. P.7). The channel estimates below are a post hoc re-analysis of their stored records under both operators (E17; Apps. P.9, P.10). A cell is an audit and a verifier, e.g. "E14/40 · `contains`" is E14's step-40 checkpoint scored by `contains`.

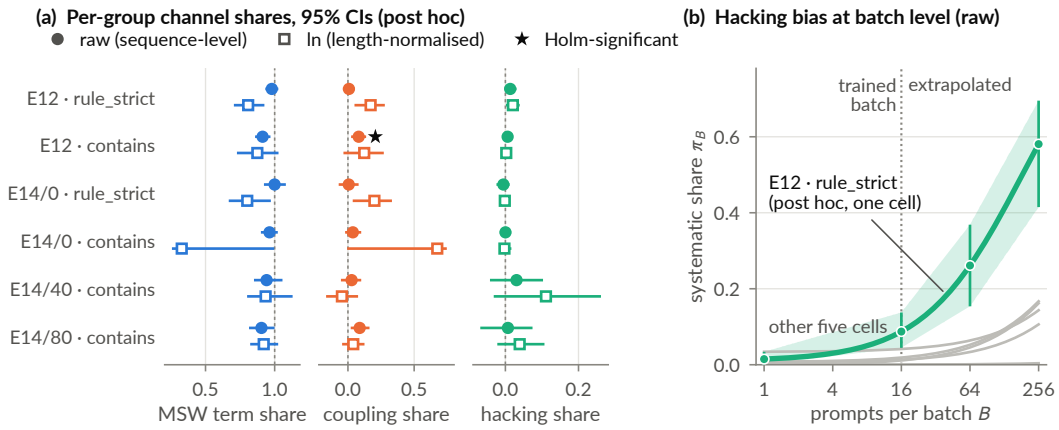

Figure 3: **Per group the MSW term carries most of the verifier noise; at batch level the hacking bias of one format checker does not average out.** **(a)** Shares of $G^2\widehat{T}$ carried by the three terms of (3) (centered moments; they sum to 1) in the six post hoc cells, sequence-level (filled) and length-normalised (open) operator; 95% percentile bootstrap intervals over prompts, not corrected for multiplicity; $\star$: Holm-adjusted $p \leq 0.05$ over the 12 intervals. **(b)** Share π_B of a B-prompt batch perturbation that is systematic given its prompts, (4), sequence-level operator. Curves evaluate (4) at each cell's stored estimates; markers and band are the stored estimates and 95% intervals at $B = 1, 16, 64, 256$. Dotted: the 16-prompt batch of the E14 GRPO run. Only the E12 `rule_strict` cell has a prompt-shared hacking direction whose interval excludes 0; its intervals condition on the one count sketch shared across groups.

The pre-specified predictions mostly failed. In the pretrained audit **all** 4 **of the** 4 **directional predictions failed**, and 2 of 7 against-the-paper items fired. In the training-time audit 2 of the 2 evaluable directional predictions held (geometry trends), 5 were not evaluable because training against `rule_strict` left 1–2 error-carrying groups, and 2 of 9 against-the-paper items fired. Over the 6 interpretable cells $\widehat{T}/\widehat{B_{\mathrm{ICC}}}$ stayed between 0.88 and 1.21 under the sequence-level operator (intervals 0.78–1.32); on the pretrained audit's untruncated groups it is 1.43 $[1.24, 1.63]$ for `contains` (post hoc, App. P.9). Table 2 (App. B) gives every statistic.

Per group, the MSW term dominates (Figure 3a). From centered moments the three terms sum to $G^2\widehat{T}$ exactly. Under the sequence-level operator the MSW term is 0.90–1.00 of the total in the 6 cells. The coupling share lies between 0.006 and 0.089 (E14 step 80 / `contains` / `raw`, $[0.021, 0.163]$). Its interval excludes 0 in 2 of 6 cells, and after Holm correction over the 6 in 1 (E12 / `contains` / `raw`). The cells are not independent: they share completions, prompts and one training trajectory. Under length normalization the coupling share reaches 0.20, and 0.67 $[-0.01, 0.75]$ at E14 step 0 / `contains` / `ln`, where most completions hit the token cap. The per-group hacking share is at most 0.030 under `raw` and 0.111 in absolute value overall. It is a U-statistic of a nonnegative quantity, yet 3 of 12 point estimates are negative; the variance of the count sketch, which all groups of an audit share and our intervals do not propagate, may explain this. The uncentered ratio A/B excluded 1 in 6 of 12 cells but has the coupling's sign in only 8: it is not the coupling (§3). The coupling's sign change between operators, predicted from a fitted length elasticity, occurred in 1 of 6 cells (refuted; the comonotonicity it assumes is absent within groups, App. P.10). The operator matters: an audit computed for one operator does not transfer to the other.

At batch level, the hacking bias need not be small (Figure 3b). For E12 / `rule_strict` / `raw` the prompt-averaged hacking direction is measurably nonzero, $\mathcal{S}/\bar{\Theta} = 0.0053$ $[0.0027, 0.0086]$ (interval above 0 in 1 of 6 cells). By (4) the systematic part grows from 0.015 $[-0.003, 0.034]$ per group to 0.09 $[0.04, 0.14]$, 0.26 $[0.15, 0.37]$ and 0.58 $[0.41, 0.70]$ of the perturbation at $B = 16$, 64 and 256 prompts, consistent with format false negatives pushing every prompt the same way. This is post hoc and one cell. $B = 64$ extrapolates beyond the trained batch, and the shared

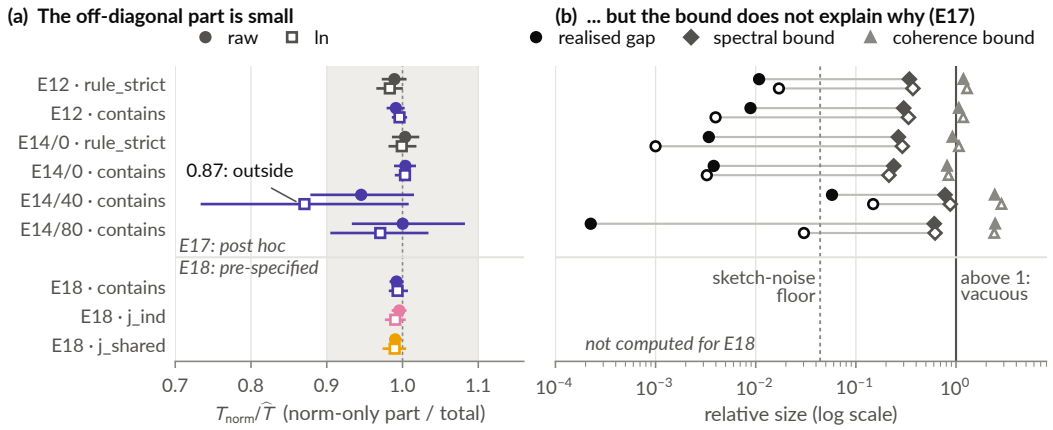

Figure 4: **The off-diagonal part is small on every LLM audit, but the available bound does not explain why. (a)** $T_{\text{norm}}/\widehat{T}$ (norm-only part over total) per cell with 95% intervals — E17: prompt bootstrap, post hoc; E18: two-level bootstrap over seeds and prompts, prediction H1 pre-specified — against the $\pm 10\%$ band fixed in advance (grey). Colours are verifiers (`rule_strict` grey, `contains` violet, `j_ind` magenta, `j_shared` yellow); filled/open: `raw`/`ln`. **(b)** E17 cells: the realised gap $|\widehat{T} - \widehat{T_{\text{norm}}}|/\widehat{T_{\text{norm}}}$, the spectral bound (Theorem 13) and the coherence bound, log scale. The spectral bound holds in every cell but is 5.9–2,690 times the gap; the coherence bound is vacuous.

count sketch correlates groups, so its variance is not propagated and Proposition 12's independence hypothesis holds only approximately. In the other sequence-level cells π_{64} is 0.001–0.06. Under `ln` no cross-group Gram was stored, and the upper bound in (4) (0.00–0.90 at $B = 64$) is uninformative.

Near-orthogonality is a regularity, not an explanation (Figure 4). By Theorem 2, the gap between the total and its norm-only part $T_{\text{norm}} = G^{-2}\mathbb{E}\sum_i (P\varepsilon)_i^2 \|s_i\|^2$ is the hacking term. Empirically $T_{\text{norm}}/\widehat{T} \in [0.90, 1.10]$ in 11 of 12 cells. The exception, E14 step 40 / `contains` / `ln`, gives 0.871 $[0.73, 1.01]$, outside a band fixed in advance. For one group the gap is a quadratic form in the hollow cosine matrix of the scores. It is bounded by that matrix's operator norm, and by Gershgorin in turn by $G - 1$ times the largest score cosine (Theorem 13, App. B). On the stored LLM Grams the median largest cosine is 0.12–0.39 against $1/(G-1) = 0.143$, so the coherence bound is 0.81–2.84, vacuous as declared in advance. The spectral bound is valid — 0.214–0.875, below 1 in 12 of 12 cells and far above the sketch-noise floor 0.044 — but 5.9 to 2,690 times the realized gap. Why the realized quotient lands near 0 depends on how the error pattern aligns with the matrix's eigenvectors, which we cannot model. A forward-only proxy that replaces $\|s_i\|^2$ by a power of the completion length was refuted as well (App. P.10). Near-orthogonality fails where scores share few directions: on a 57,902-parameter toy policy the ICC-implied baseline was off by 1.34–4.82× (exploratory, App. P.1), and the designed 47× construction needs same-strategy score cosines of 0.896 (App. G).

6 A PRE-SPECIFIED MULTI-SEED AUDIT WITH LLM JUDGES

Design. The *multi-seed audit* (E18; exploratory, predictions H1–H4 pre-specified in a hashed plan before any measured run) runs `Qwen2.5-0.5B-Instruct` on GSM8K and on the integer-answer subset of MATH-500: 90 prompts, 3 sampling seeds, a 1,024-token cap and both operators. Besides the string matchers it uses two stochastic LLM judges that never see the label: `j_ind`, one independent call per completion (regime (ii)), and `j_shared`, one sampled reference solution shared by a group's calls (regime (iii); App. P.11). By the frozen reportability rules only the GSM8K cells of `contains`, `j_ind`, `j_shared` are reportable: at least 30 error-carrying groups per seed, judge parse failures at most 1.0% and truncation at most 0.42%. `rule_strict` has too few errors, and MATH-500 truncates 7.6%–9.0%, above a 5% limit.

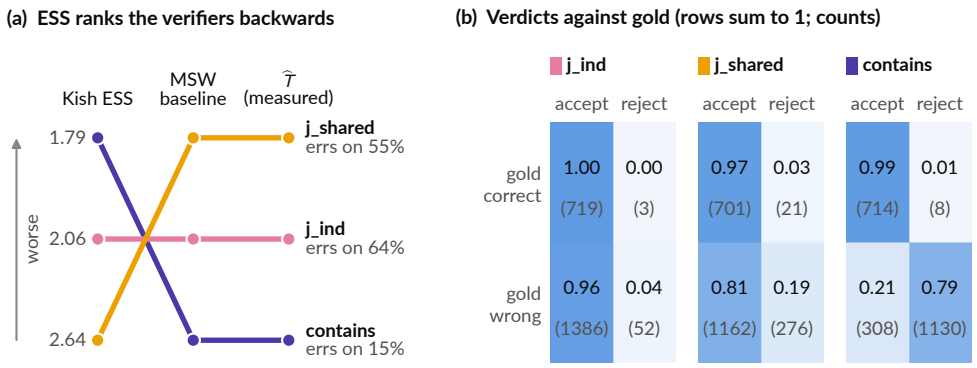

Figure 5: **With verifiers of very different error scale, Kish ESS ranks them in the reverse order of the measured perturbation (H4, pre-specified), and the two judges are lenient.** E18, `Qwen2.5-0.5B-Instruct`, GSM8K, 3 seeds. **(a)** Rankings, worst at the top, by Kish ESS (values at left), by the MSW-based ICC-implied baseline and by $\widehat{T}$ (sequence-level); error rates are per completion, against gold. ESS is scale-free: it depends only on the ICC, and `contains` has the most clustered but by far the rarest errors. The reversal therefore shows that a scale-free summary cannot rank such verifiers, not that the decomposition is right. **(b)** Verdicts against gold pooled over seeds (2,160 completions), row-normalised, with counts. In the same audit H2 (a coupling sign change between operators) and H3 (a shared reference raises the ICC but not $\widehat{T}$) failed.

The judges are weak and lenient (Figure 5b). Pooled over seeds (2,160 GSM8K completions, gold accuracy 0.33), their true-negative rates are 0.04 (`j_ind`) and 0.19 (`j_shared`), against 0.79 for `contains`. They disagree with gold on 0.64 and 0.55 of completions, more than a judge that always answers "incorrect" (0.33); `contains` disagrees on 0.15.

Results. **H1 held**: $T_{\text{norm}}/\widehat{T} \in [0.990, 0.996]$ over all 6 reportable cells, every interval inside the band (Figure 4a). **H2 failed**: for `contains` the coupling statistic κ_{norm} is 0.160 $[0.075, 0.271]$ under `raw` but 0.010 $[-0.075, 0.104]$ under `ln`, so there is no sign change. **H3 failed.** It predicted that the reference-sharing judge would raise the ICC but not $\widehat{T}$. Instead `j_shared` had the *lower* ICC (0.29 against 0.41) and the *higher* $\widehat{T}$ (ratio 1.53 $[1.36, 1.72]$ under `raw`), and its shared component of the conditional error covariance, 0.068 $[-0.032, 0.165]$, has an interval containing 0. The reference-sharing design produced at most a small shared component, so our evidence on regime (iii) is weak. **H4 held** (Figure 5a). Kish ESS ranks the three verifiers, worst first, `contains` > `j_ind` > `j_shared`; $\widehat{T}$ ranks them `j_shared` > `j_ind` > `contains`, the exact reverse (Kendall $\tau = -1$ under both operators). All 6 pairwise reversals have interval support ($\widehat{T}$ ratios 0.30–0.75, upper limits at most 0.85), and the MSW-based ICC-implied baseline ranks the verifiers as $\widehat{T}$ does. This demonstrates that a scale-free dependence summary cannot rank verifiers of different error scale. It is consistent with, not a test of, the dominance of the MSW term: ESS depends only on the ICC, `contains` has the most clustered errors (ICC 0.50) but errs far less often, and $\widehat{T}$ follows the MSW term (MSW 0.064, 0.137, 0.191). One model, one task, three verifiers, three seeds. Against-the-paper items 4 of 11 fired: (iii), (v), (vi), (vii). The common-mode check held ($\widehat{T}$(`cmshift`)/$\widehat{T}$(`rule_strict`) $\leq 2.5 \times 10^{-31}$). *Deviation:* the non-Qwen policy (SmolLM2-1.7B-Instruct) was not completed. Contention on a shared GPU left 74 and 58 of 90 groups in two seeds, so this audit covers one model family.

What the measurements do and do not support. Conditional on $G = 8$, one model family and mostly GSM8K: per group and under the sequence-level operator the MSW term is most of the total; the per-group share is the wrong unit for the hacking channel; the operator matters; and with verifiers of very different error scale an ESS ranking can be backwards.

7 Related work

Xin (2026) is concurrent: it measures within-group verifier-error correlation on $24{,}998$ Qwen2.5 rollout groups and replays group-relative advantages to count sign disagreements; it does not claim that ESS predicts gradient noise. El Mansouri et al. (2025) debias GRPO gradients under *independent* Bernoulli flips — the first moment, with a diagonal error covariance — whereas we characterise the second moment and which error moments it contains. **What is classical.** The ICC (Shrout & Fleiss, 1979); the Kish design effect and its unequal-weighting factor $1 + \mathrm{CV}^2$ (Kish, 1965); the Moulton factor (Moulton, 1990; Cameron & Miller, 2015) and its within-cluster mirror image, which the reversal is; the CR0 sandwich meat of GEE (Liang & Zeger, 1986), to which $\widehat{T}$ corresponds; compound symmetry of exchangeable covariances, which Corollary 9 is; mutual-coherence and Gershgorin bounds, which the last link of Theorem 13 is; and Hoeffding's variance decomposition of U-statistics (Hoeffding, 1948), of which Theorem 2 is an instance for the covariance kernel. None of the completion-level moments is new. **What we add** is the reading of that decomposition as three audit channels of which a scalar pipeline sees one, with the hacking channel's bias separated at the batch level; the centered-score treatment under length normalization; the regime-(iii) functional; and measurements of the channels on real policies under both operators. Operator variants (Liu et al., 2025; Yu et al., 2025; Ahmadian et al., 2024; Zhou et al., 2026), error-channel models (Cai et al., 2025; Wang et al., 2018) and noise-injection studies (Plesner et al., 2026; Zhu & Kang, 2026; Shao et al., 2025) are discussed in App. R.

8 Limitations and conclusion

No link to training outcomes. Our quantities concern one update's verifier-induced perturbation, conditional on the sampled completions; policy-sampling variance is a separate, typically larger term. The one learning test we ran, on a toy, is a powered null (App. O). A training-outcome experiment on a language model (E19) was designed and pre-specified but not run within the compute budget. **Regime (iii).** Our designed shared-randomness judge produced no clear shared error component, so the regime-(iii) theorem is essentially untested. **Breadth.** All completed LLM policies are Qwen2.5 (0.5B, 1.5B), mostly on GSM8K, at $G = 8$. The pretrained and training-time audits are one run each with binding token caps (16.6%, and 52.1% at step 0), and the multi-seed audit's reportable cells are one model on one task. Scores are count-sketched, with one sketch shared across groups whose variance our intervals do not include and which makes groups only approximately independent. **Multiplicity.** The cells come from two sampling runs and one training trajectory; after Holm correction 1 of 12 coupling intervals excludes 0. **Operator.** Real GRPO adds the z-score divisor, clipping, a KL term and Adam. The identities are exact only for the linear update, and the z-scored theory is first order and fails for binary rewards.

We propose that a verifier audit for group-relative RL report, under the operator actually trained, the MSW term, the coupling share, and the hacking share at the trained batch size — the batch-level share (4), not the per-group one. Whether these channels predict training damage is untested (E19 designed, not run). An ICC or an ESS is not enough: both are scale-free and see none of the three terms. On these policies the MSW term dominates per group, but in one post hoc, extrapolated cell a format checker's hacking bias is a substantial part of a batch update, and on a small toy policy the gap to the scalar baseline was large.

AI use statement

This paper was prepared with assistance from large language model agents (Claude Opus models). *Uses requiring disclosure:* writing the experiment infrastructure code (training, audit, analysis and figure scripts, and the containerised environment) and executing it; editing the manuscript for grammar and language; acting as simulated reviewers that critiqued successive drafts.

Ethics statement

This is a theoretical and numerical study of an optimizer's noise; it involves no human subjects, no personal data and no new models or datasets. We identify verifier-error components that are *invisible* to the group-relative update, which could also be read as guidance on which verifier failures

leave no training-time signal; we judge disclosure net positive, since an auditor who knows that a common-mode leniency shift is invisible to the update is better placed to detect it. A small audited perturbation does not mean a verifier is correct: a verifier can be badly biased in the common-mode direction while injecting no gradient noise at all.

Reproducibility statement

Every result is proved in the appendix (Apps. B–I), with all notation in App. A and construction parameters as exact integers (App. J). The simulation designs are in Apps. K–L, the LLM audits' designs, frozen plans and deviations in Apps. P.6–P.11. The simulations are CPU-only; the GPU audits are one-command Docker packages with recorded code hashes, and every re-analysis recomputes from saved per-group records without a GPU. Every experimental number in the text and in the figures is generated from the experiments' result files by a script, and App. Q lists the scripts and commands; code and plan files with self-reported SHA-256 digests are in the supplementary material.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://arxiv.org/abs/2402.14740.

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers. *arXiv preprint arXiv:2510.00915*, 2025. URL https://arxiv.org/abs/2510.00915.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317. URL https://doi.org/10.3368/jhr.50.2.317.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected GRPO: From noisy rewards to unbiased gradients. *arXiv preprint arXiv:2510.18924*, 2025. URL https://arxiv.org/abs/2510.18924.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022. URL https://arxiv.org/abs/2210.10760.

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

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. LLMs gaming verifiers: RLVR can lead to reward hacking. *arXiv preprint arXiv:2604.15149*, 2026. URL https://arxiv.org/abs/2604.15149.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

Yuzhen Huang, Weihao Zeng, Xingshan Zeng, Qi Zhu, and Junxian He. From accuracy to robustness: A study of rule- and model-based verifiers in mathematical reasoning. *arXiv preprint arXiv:2505.22203*, 2025a. URL https://arxiv.org/abs/2505.22203.

Zixun Huang, Jiayi Sheng, and Zeyu Zheng. Variance-aware baselines and adaptive learning rates for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2511.23310*, 2025b. URL https://arxiv.org/abs/2511.23310.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. *arXiv preprint arXiv:2506.07962*, 2025. URL https://arxiv.org/abs/2506.07962.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Leslie Kish. Sampling organizations and groups of unequal sizes. *American Sociological Review*, 30 (4):564–572, 1965. doi: 10.2307/2091346. URL https://doi.org/10.2307/2091346.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026. URL https://arxiv.org/abs/2605.29800.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13. URL https://doi.org/10.1093/biomet/73.1.13.

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.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024. URL https://arxiv.org/abs/2411.00640.

Brent R. Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 72(2):334–338, 1990. doi: 10.2307/2109724. URL https://doi.org/10.2307/2109724.

Adam Noonan. The exceedance design effect: Effective sample size for thresholds under clustering. *arXiv preprint arXiv:2608.21262*, 2026. URL https://arxiv.org/abs/2608.21262.

Andreas Plesner, Francisco Guzmán, and Anish Athalye. An imperfect verifier is good enough: Learning with noisy rewards. *arXiv preprint arXiv:2604.07666*, 2026. URL https://arxiv.org/abs/2604.07666.

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. *arXiv preprint arXiv:2506.10947*, 2025. URL https://arxiv.org/abs/2506.10947.

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

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420. URL https://doi.org/10.1037/0033-2909.86.2.420.

Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. *arXiv preprint arXiv:1810.01032*, 2018. URL https://arxiv.org/abs/1810.01032.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 538–545, 2001. URL https://arxiv.org/abs/1301.2315.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Esther Xin. Are verifier errors independent within a GRPO group? evidence from Qwen2.5 rollouts. *arXiv preprint arXiv:2609.06386*, 2026. URL https://arxiv.org/abs/2609.06386.

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

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

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.

Appendix: supplementary material

This appendix follows the references, as permitted by the call. Sections A–I give notation, the statements that the main text summarises (App. B: the five-moment form, the finite-sample identities, the operators, the z-score divisor, the audit estimator, the spectral bound and the verdict tables), the probability space and every proof, including the moment decomposition (App. D), the spectral bound (App. E) and the designed extremes (App. G); Sections J–O the construction parameters, the pre-specified simulation suite, its design, the full criteria table and the per-study detail; Section P the exploratory studies, including the three language-model audits and their re-analyses; and Section Q the reproduction commands. Code and plan files are in the supplementary archive.

A NOTATION

Table 1: Symbols used in the paper. Where a symbol is reused with a different meaning, the two uses are separated by context and both are listed.

symbol	meaning
G, d	group size; parameter dimension
x, y_i, ω	prompt; completion i; verifier randomness
r, $\tilde{r}$, ε	clean rewards; verifier scores; residual $\tilde{r} - r$
s_i, S, K	score vector $\nabla_\theta \log \pi_\theta(y_i \mid x)$; $[s_1 \dots s_G]$; $S^\top S$
P, $\mathbf{1}$	centering projector $I - \mathbf{1}\mathbf{1}^\top/G$; all-ones vector
μ, $\Sigma_{S,r}$, Σ_S	$\mathbb{E}[\varepsilon \mid S, r]$; $\mathrm{Cov}(\varepsilon \mid S, r)$; $\mathrm{Cov}(\varepsilon \mid S)$
g, Δg, C	clean update; verifier perturbation; $\mathrm{Cov}(\Delta g \mid S, r)$
λ_k, u_k	eigenvalues/eigenvectors of $P\Sigma P$ on $\mathbf{1}^\perp$
ν_k	eigenvalues of PKP on $\mathbf{1}^\perp$
$\bar{d}$, $\bar{o}$	mean diagonal and mean off-diagonal of Σ
ICC, ESS, MSW	$\bar{o}/\bar{d}$; $G/(1+(G-1)\rho)$; within-group mean square
α, β, γ	weights in $\Sigma(u) = \alpha I + \beta \mathbf{1}\mathbf{1}^\top + \gamma u u^\top$
κ, λ, h, u	idiosyncratic and shared score energy; strategy sign vector; error sign vector
a_g, σ_a^2	ANOVA group effect and its variance (Proposition 4)
t	prompt-category coupling strength (App. P.4)
c, $\hat{c}$, Q	Pr; $c/\|c\|$; $I - \hat{c}\hat{c}^\top$
A, $\mathcal{R}$, ϱ, η, κ_4	advantage map; remainder; $\|P\varepsilon\|$; $\sqrt{\mathrm{tr}P\Sigma P}/\|Pr\|$; normalized 4th moment
SR, $\gamma_\perp$	suppression ratio (15); surviving noise fraction
R, R_{pool}, $T_{x,S}$	population risk; pooled plug-in; per-prompt risk $G^{-2}\mathrm{tr}(P\Sigma_x P K_x)$
$\widehat{T}$, $\widehat{T}_\Omega$, $\widehat{T}_w$	audit estimator; sketched; directional
m, Ω, k, w, z_j, e_j	audit size; sketch; sketch dimension; probe; $S_j P e_j$; audit residual
n_+, T_{iso}	correct completions in a group; isotropic comparator (App. P.1)
$M_{S,r}$	$\mathbb{E}[\varepsilon\varepsilon^\top \mid S, r] = \Sigma_{S,r} + \mu\mu^\top$
A, B, C, D, F; Θ_G	completion-level moments (Definition 8); their combination (6)
k_i, T_{norm}, N_j	$\|s_i\|^2$; norm-only part $G^{-2}\mathbb{E}\sum_i (P\varepsilon)_i^2 k_i$; its per-group value $\sum_i v_{j,i}^2 k_{j,i}$
$\mathcal{C}$, $\mathcal{H}$, δ	cosine matrix of the scores; $\mathcal{C} - I$; coherence $\max_{a \neq b} \lvert\mathcal{C}_{ab}\rvert$
κ_{within}, κ_{scale}, $\hat{\theta}$, φ	the channels of (10) (Proposition 14)
B_{ICC}, R_{pool}	ICC-implied baseline; pooled plug-in (Proposition 4)
`raw`, `ln`	operators K and DKD, $D = \mathrm{diag}(1/\lvert o_i\rvert)$
$\mathcal{T}$, V	$\mathrm{tr}(P\Sigma PK)$; $\mathrm{tr}((P\Sigma PK)^2)$ (Proposition 27)
Φ, Pr	standard normal CDF; probability

B STATEMENTS THE MAIN TEXT SUMMARISES

This section collects, in full, the statements and the bookkeeping that the main text only summarises. Proofs are in the sections that follow.

B.1 THE DECOMPOSITION IN FIVE COMPLETION-LEVEL MOMENTS

Definition 8 (Completion-level moments)**.** For one completion $y \sim \pi_\theta(\cdot \mid x)$ with error $\varepsilon = \varepsilon(y)$, score $s = s(y, x)$, score mean $\mu_s = \mathbb{E}[s \mid x]$ and centered score $\bar{s} = s - \mu_s$, let $\mathsf{A} = \mathbb{E}[\varepsilon^2 \|\bar{s}\|^2]$, $\mathsf{B} = \mathbb{E}[\varepsilon^2]\,\mathbb{E}\|\bar{s}\|^2$, $\mathsf{C} = \|\mathbb{E}[\varepsilon\bar{s}]\|^2$, $\mathsf{D} = \mathbb{E}[\varepsilon]\,\mathbb{E}[\varepsilon\|\bar{s}\|^2]$ and $\mathsf{F} = \mathbb{E}[\varepsilon]^2\,\mathbb{E}\|\bar{s}\|^2$. For the sequence-level score $\mu_s = 0$ (the score identity), so $\bar{s} = s$.

A and B read the error and the score norm off the same completion or off two independent ones, so $\mathsf{A}/\mathsf{B} = 1$ when $\varepsilon \perp \|\bar{s}\|^2$. $\mathsf{C} = \|\mathrm{Cov}(\varepsilon, s)\|^2$ is the only moment that sees the *direction* of s: $\mathbb{E}[\Delta g \mid x] = \frac{G-1}{G}\mathrm{Cov}(\varepsilon, s)$, and for the sequence-level score $\mathrm{Cov}(\varepsilon, s) = \mathbb{E}[\varepsilon s] = \nabla_\theta \mathbb{E}_{\pi_\theta}[\varepsilon]$ is the reward-hacking gradient (Proposition 19, App. F). D, F vanish when $\mathbb{E}\varepsilon = 0$. Under Assumption 1 with ω_0 trivial, and with no assumption on μ_s, Theorem 2 is equivalent to

$$\begin{aligned} G^2\,\mathbb{E}\big[\|\Delta g\|^2 \mid x\big] &= \Theta_G(\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{F}) \\ &:= \tfrac{G-1}{G}\Big[(G-1)\mathsf{A} + \mathsf{B} + (G^2 - 2G + 2)\mathsf{C} - 2(G-1)\mathsf{D} + (G-2)\mathsf{F}\Big]. \end{aligned} \tag{6}$$

In Hoeffding's form (main text, §3) the centered-score kernel variances are $4\zeta_1 = \mathsf{A} - 2\mathsf{D} + \mathsf{F} - \mathsf{C}$ and $2\zeta_2 = \mathsf{A} + \mathsf{B} - 2\mathsf{D}$; collecting terms is (6). For $\varepsilon \perp s$ only the MSW term survives, and a completion-independent shift contributes nothing. Everything is verified in exact rational arithmetic (App. D).

Corollary 9 (Exchangeability: no slot-level eigenstructure)**.** *Under Assumption 1 (all three regimes), every pooled slot-space summary — any permutation-equivariant matrix functional of the group, averaged — equals $aI + b\mathbf{1}\mathbf{1}^\top$; in particular every fixed $u \perp \mathbf{1}$ with $\|u\| = 1$ has population alignment $(G-1)u^\top \bar{K} u / \mathrm{tr}(P\bar{K}P) = 1$ exactly. If moreover $\mu_s = 0$ (for `1n`: K formed from centered scores) and the verifier is in regime (i)–(ii) or in the additive regime (iii) of Theorem 3, the $G \times G$ off-diagonal structure of K enters the population risk only through C (and C_u); for $\mu_s \neq 0$ the off-diagonal part also carries Ξ.*

Structure that survives is tied to a *feature* of the completion (format, length, strategy), and it enters through A and C, not through slot geometry.

Lemma 10 (Finite-sample identity for one realized group)**.** *For errors e, Gram K and $v = Pe$, with $S_e = \sum_i e_i$, $S_{ek} = \sum_i e_i k_i$, $S_k = \sum_i k_i$, $\hat{\mathsf{A}} = \sum_i e_i^2 k_i$, $\hat{\mathsf{C}} = \sum_{i \neq j} e_i e_j K_{ij}$, $\chi = \sum_{i \neq j} e_i K_{ij}$ and $K_{\mathrm{tot}} = \mathbf{1}^\top K \mathbf{1}$, exactly, $N := \sum_i v_i^2 k_i = \hat{\mathsf{A}} - \frac{2}{G} S_e S_{ek} + \frac{1}{G^2} S_e^2 S_k$ and $v^\top K v = \hat{\mathsf{A}} + \hat{\mathsf{C}} - \frac{2}{G} S_e(S_{ek} + \chi) + \frac{1}{G^2} S_e^2 K_{\mathrm{tot}}$. The same raw sums give per-group U-statistics of the uncentered moments; Θ_G of them differs from $v^\top K v$ group by group by exactly the in-sample μ_s-terms, $-\Xi$ evaluated at four further U-statistics, so that $\Theta_G + \Xi = v^\top K v$ exactly (App. D).*

B.2 OPERATORS

Under $S \to SD$ for an (S, r)-measurable diagonal D, with $K \to DKD$, the pathwise statements hold verbatim: Theorem 5, its corollaries, Lemma 10 and the pooled identity (10) (App. D); the cosine matrix of the scores is invariant (App. E). The population statements need centering, because the length-weighted score $s/|o|$ has a nonzero mean: Theorem 2 and Proposition 4 hold exactly with every moment taken from $s/|o| - \mathbb{E}[s/|o|]$ (MSW term $(G-1)\,\mathrm{Var}(\varepsilon)\mathrm{trCov}(s/|o|)$, hacking term $\|\mathrm{Cov}(\varepsilon, s/|o|)\|^2$), and the diagonal/off-diagonal attribution in Theorem 2 needs $\mu_s = 0$. On our audits the μ_s-terms that uncentered moments omit are within $\pm 5.4\%$ of the risk (App. P.10); all shares we report use centered moments. A past-measurable preconditioner is covered like D; Adam only as $\beta_2 \to 1$, gradient clipping not at all (Corollary 17, App. F).

B.3 THE Z-SCORE DIVISOR

Standard GRPO's advantage is exactly $\sqrt{G}\,P\tilde{r}/\|P\tilde{r}\|$ ($\sqrt{G-1}$ with the unbiased sample sd); to first order, with $\mathbb{E}[\varepsilon \mid S, r] = 0$ and a finite normalised fourth moment, it adds a second projection onto $\{\mathbf{1}, Pr\}^\perp$ (Theorem 24, App. H). Binary verifiers violate both hypotheses, and the predicted suppression is absent at every flip rate in the pre-specified simulations ($\mathrm{SR} = 0.74$–0.99 against a predicted 0), with up to 4.8% of groups degenerate and dropped. Those simulations (App. K) verify every identity to 0.0176 relative Frobenius error; of their 4 falsifiable hypotheses 1 passes, and all 5 literal failures are reported there.

B.4 THE AUDIT ESTIMATOR AND THE BATCH-LEVEL SHARE

For m gold-labelled groups record $e_j = \tilde{r}_j - r_j^{\text{gold}}$ and set $\widehat{T} = m^{-1}\sum_j G^{-2}\|z_j\|^2$ with $z_j = S_j P e_j = \nabla_\theta \sum_i (e_{j,i} - \bar{e}_j)\log \pi_\theta(y_{j,i} \mid x_j)$: the usual surrogate gradient with advantages Pe_j, one extra backward pass per group (Algorithm 1). It corresponds to the cluster-robust (CR0) sandwich *meat* (Liang & Zeger, 1986) with design $S_j P$; it is not a new statistic.

Theorem 11 (Unbiasedness)**.** *If the gold labels are error-free and the audit groups are drawn i.i.d. from the deployment distribution, $\mathbb{E}[\widehat{T}] = \mathbb{E}_{x,S}\mathbb{E}\|\Delta g\|^2$ for arbitrary unknown Σ_x and noise law; under Assumption 1 (ω_0 trivial) this is $\mathbb{E}_x[G^{-2}\Theta_G(\mathsf{A}_x, \ldots, \mathsf{F}_x)]$.*

This is immediate, because the errors are observed: $\widehat{T}$ is a sample mean of the target. We use prompt-bootstrap intervals (a closed-form variance needs mean-zero Gaussian errors, App. I). Our audits use one count sketch for all groups, and the bootstrap conditions on it; the unpropagated sketch variance may explain why 3 of 12 estimates of the nonnegative hacking share are negative.

Proposition 12 (Systematic versus remainder)**.** *The cross-group U-statistic $\widehat{T}^{\text{sys}} = (m(m-1))^{-1}\sum_{j\neq j'} G^{-2}\langle z_j, z_{j'}\rangle$ is unbiased for $\|\mathbb{E}\Delta g\|^2$, the squared (prompt-averaged) hacking direction; $\widehat{T} - \widehat{T}^{\text{sys}}$ is unbiased for the trace of the* marginal *covariance and still contains the across-prompt spread of prompt-specific hacking.*

Algorithm 1 Channel-resolved verifier audit (under the operator actually trained)

Require: policy π_θ; prompts $x_{1:m}$; gold labeller; group size G; operator weights $d_{j,i}$ (1 for `raw`, $1/|o_{j,i}|$ for `ln`)

1: **for** $j = 1, \ldots, m$ **do**
2: sample $y_{j,1:G} \sim \pi_\theta(\cdot \mid x_j)$; $e_j \leftarrow \tilde{r}_j - r_j^{\text{gold}}$; skip if $Pe_j = 0$ (it contributes exactly 0)
3: $z_j \leftarrow \nabla_\theta \sum_i (e_{j,i} - \bar{e}_j)\, d_{j,i} \log \pi_\theta(y_{j,i} \mid x_j)$ ▷ one backward pass; $= S_j D_j P e_j$
4: optional: K_j of the weighted scores $d_{j,i}s_{j,i}$ (G passes, or a per-group sketch)
5: **end for**
6: **return** $\widehat{T} = m^{-1}\sum_j G^{-2}\|z_j\|^2$ and $\widehat{T}^{\text{sys}}$ (Proposition 12); with K_j: the MSW, coupling and hacking shares from the centered U-statistics of Lemma 10; with $\widehat{T}^{\text{sys}}$: the batch share (4)

What an audit must compute. $\widehat{T}$ needs one extra error-weighted backward pass per group. The channel split needs per-completion scores: G backward passes per group and either exact per-completion Grams K_j or a fresh sketch per group (Proposition 27); gold labels on $m \times G$ completions remain the dominant cost. The batch-level share needs the cross-group inner products $\langle z_j, z_{j'}\rangle$, which a sketch shared across groups provides; $\mathcal{S}$ in (4) is estimated by $G^2\widehat{T}^{\text{sys}}$. An *unstratified* pooled plug-in is biased whenever Σ_x and K_x covary across prompts, while a *stratified* one can beat $\widehat{T}$ (App. P.4); the designed $47\times$ and $25\times$ gaps live outside Assumption 1 (App. G).

B.5 THE SPECTRAL BOUND

For one realized group, with cosine matrix $\mathcal{C} = D_k^{-1/2} K D_k^{-1/2}$, hollow part $\mathcal{H} = \mathcal{C} - I$ and $w = D_k^{1/2} v$, $v^\top K v - N = w^\top \mathcal{H} w$ exactly.

Theorem 13 (Spectral bound)**.** *For centered v, $(v^\top K v - N)/N$ ranges exactly over $[\lambda_{\min}, \lambda_{\max}]$ of the compression of $\mathcal{H}$ to $D_k^{1/2}(\mathbf{1}^\perp)$. This range interlaces the spectrum of $\mathcal{H}$ and can be sign-definite: for equal norms and equal cosines c it is $\{-c\}$. Over groups, with $\langle\cdot\rangle_N$ the N_j-weighted mean and $\delta_j = \max_{a\neq b}|\mathcal{C}_{j,ab}|$,*

$$|\widehat{T} - \widehat{T_{\text{norm}}}|/\widehat{T_{\text{norm}}} \;\le\; \langle \|\mathcal{H}_j\|_{\text{op}}\rangle_N \;\le\; (G-1)\,\langle \delta_j\rangle_N,$$

the last link by Gershgorin; $\mathcal{H}_j$ is invariant under $S \to SD$, so the two operators differ only through the weights.

The Gershgorin (mutual-coherence) link is attained in the worst case (Proposition 15) and needs $\delta \ll 1/(G-1)$ to say anything. On the stored LLM Grams (post hoc, App. P.10) the coherence

bound is vacuous and the spectral bound valid but loose (§5). Why the realized Rayleigh quotient lands near 0 is a statement about how the error pattern aligns with the eigenvectors of the compressed $\mathcal{H}_j$, which we cannot model.

B.6 THE LANGUAGE-MODEL AUDITS: VERDICT TABLES

Table 2 gives the pre-specified E12/E14 statistics as first reported; Table 3 the ratios and channel shares per cell under both operators. The E17 verdicts are listed in App. P.10 and the E18 verdicts in App. P.11.

Table 2: **The pre-specified E12/E14 statistics, as first reported** (sequence-level operator). E12: `Qwen2.5-1.5B-Instruct`, $m = 300$; E14: `Qwen2.5-0.5B-Instruct` at GRPO steps 0 and 80, $m = 120$. Brackets: 95% bootstrap intervals over prompts; nulls 1, 1 and $1/(G-1) = 0.143$; PR: participation ratio of PKP (isotropic 7); n: groups with $Pe_j \neq 0$. Every `rule_strict` interval contains its null or lies below it. † Not interpretable: training removed the verifier's own errors ($n = 2$); rows with $n < 30$ show no interval. E12 truncates 16.6% of completions; levels are not comparable across E12 and E14.

setting	$\widehat{T}$/ICC-impl. (P1)	directional (P2)	mean $\cos^2$ (P3)	PR	n
E12 `rule_strict`	0.94 [0.88, 1.00]	1.02 [0.98, 1.07]	0.147 [0.109, 0.192]	6.11	131
E12 `contains`	1.20 [1.11, 1.29]	1.08 [1.03, 1.14]	0.240 [0.184, 0.300]	6.11	115
E14 `rule_strict`, step 0	0.88 [0.78, 0.97]	1.00 [0.94, 1.07]	0.170 [0.101, 0.245]	6.31	49
E14 `rule_strict`, step 80†	1.09 n=2, no CI	0.92 n=2, no CI	0.048 n=2, no CI	5.52	2
E14 `contains`, step 0	1.04 [0.95, 1.14]	1.03 [0.98, 1.09]	0.150 [0.097, 0.209]	6.31	48
E14 `contains`, step 80	1.21 [1.10, 1.32]	1.09 [1.00, 1.19]	0.195 [0.116, 0.285]	5.52	42

Table 3: **Ratios and channel shares per audit cell, both operators.** n: groups with $Pe_j \neq 0$ (E18: minimum over seeds). Shares are of $G^2\widehat{T}$, from centered moments (the MSW share is 1 minus the two). *: the 95% interval excludes the null (1 for ratios, 0 for shares), uncorrected for multiplicity; full intervals in Tables P.10 and P.11. E12/E14 rows are post hoc; the E18 rows (GSM8K) are from its pre-specified analysis.

			$\widehat{T}/\widehat{B_{\mathrm{ICC}}}$		$\widehat{T_{\mathrm{norm}}}/\widehat{T}$		coupling share		hacking share	
audit	verifier	n	`raw`	`ln`	`raw`	`ln`	`raw`	`ln`	`raw`	`ln`
E12	`rule_strict`	131	0.94	1.15	0.989	0.983	0.007	0.171*	0.013	0.020*
E12	`contains`	115	1.20*	1.18*	0.991	0.996	0.081*	0.123	0.006	0.002
E14 step 0	`rule_strict`	49	0.88*	0.62	1.003	0.999	0.006	0.200*	-0.005	-0.002
E14 step 0	`contains`	48	1.04	3.75	1.004	1.003	0.037	0.675	0.000	-0.003
E14 step 40	`contains`	30	1.12	1.12	0.945	0.871	0.028	-0.046	0.030	0.111
E14 step 80	`contains`	42	1.21*	1.21*	1.000	0.971	0.089*	0.040	0.007	0.039
E18 (3 seeds)	`contains`	33	1.26*	0.98	0.992	0.993	–	–	–	–
E18 (3 seeds)	`j_ind`	60	0.96	1.13*	0.996	0.990	–	–	–	–
E18 (3 seeds)	`j_shared`	73	1.05*	1.09	0.990*	0.990	–	–	–	–

C THE PROBABILITY SPACE AND THE THREE VERIFIER REGIMES

A draw is $(x, y_{1:G}, \omega)$ — the prompt, the G completions, and the verifier's own randomness ω (the sampling of a judge model, a rubric, a reference solution). Both r and S are functions of $(x, y_{1:G})$, and $\varepsilon = \tilde{r} - r$ of $(x, y_{1:G}, \omega)$. The two conditionings of (5) then behave as follows.

(i) Deterministic verifier. A rule-based checker is a function of y_i alone, so $\Sigma_{S,r} = 0$ once the completions are fixed. The whole effect sits in μ: the perturbation is the deterministic $G^{-1}SP\mu$ of squared size $G^{-2}\|SP\mu\|^2$, and conditioning only on S leaves $\Sigma_S = \mathrm{Cov}(\mu \mid S)$. *If the score map $y \mapsto s(y)$ is injective* — generic once $d \gg G$, and certainly at the $d = 57{,}902$ of §P.1 — then S determines $y_{1:G}$, hence r, ε and μ, so $\Sigma_S = 0$ as well and Theorem 5(ii), (iii), Proposition 18, Corollary 7, Theorem 21 and Theorem 23 all read $0 = 0$ in the covariance form. Only the second-moment form below has content there. Note also that a pooling audit does *not* measure any S-conditional object: it measures the marginal second moment $\mathbb{E}[\varepsilon\varepsilon^\top]$ over $(x, y_{1:G}, \omega)$. This is the

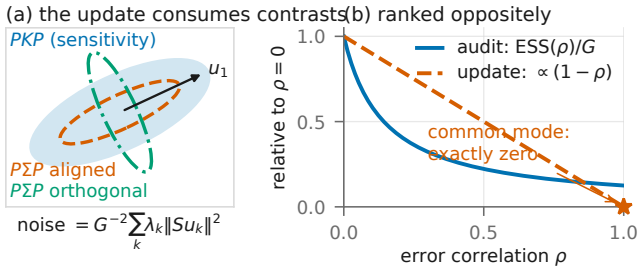

Figure 6: The change of question (**schematic**; measured in Figure P.2). **(a)** the update discards the group mean an ESS describes and consumes $P\tilde{r}$ weighted by S. For one *realized* group the injected noise is $\sum_k \lambda_k \|Su_k\|^2$ by (11), so it depends on how the contrast eigendirections of the error covariance meet the scores. This slot-level alignment is not a population mechanism: under i.i.d. completions it averages out (Corollary 9), and only structure tied to a feature of the completion survives, through the coupling and hacking terms of Theorem 2. **(b)** ESS and the gradient-noise trace are not monotone in each other (Corollaries 6, 7).

regime of §P.1, and the one in which the bias/variance split of Proposition 12 is indispensable: a scalar $\widehat{T}$ there is $\|\mathbb{E}\Delta g\|^2$ plus the across-prompt spread of the same systematic term.

The second-moment reading. Put $M_{S,r} = \mathbb{E}[\varepsilon\varepsilon^\top \mid S, r] = \Sigma_{S,r} + \mu\mu^\top$, which is PSD. Then $\mathbb{E}[\|\Delta g\|^2 \mid S, r] = G^{-2}\mathbb{E}[\varepsilon^\top PKP\varepsilon \mid S, r] = G^{-2}\mathrm{tr}(PM_{S,r}PK)$ by the same one-line computation as Theorem 5(ii)/(iv), since $\mathbb{E}[\varepsilon^\top A\varepsilon] = \mathrm{tr}(A\,\mathbb{E}[\varepsilon\varepsilon^\top])$ for symmetric A. Every consequence of the covariance identity that uses only "$\mathrm{tr}(P \cdot PK)$ with a PSD argument" therefore holds verbatim with (Σ, trCov) replaced by $(M, \mathbb{E}\|\Delta g\|^2)$: *(a)* the spectral form (11), by diagonalising PMP instead of $P\Sigma P$; *(b)* the Rayleigh bounds and the von Neumann range statement of Proposition 18, whose proof uses only PSD-ness of the two arguments; *(c)* common-mode annihilation (Corollary 6), since the kernel computation is a statement about the linear map $A \mapsto PAP$; *(d)* Theorem 21, because $\Sigma(u)$ enters only through $P\Sigma(u)P$, and a deterministic error with $\mu = \sqrt{\gamma}\,u$ gives $M = \gamma uu^\top$ exactly; *(e)* Theorem 23, by bilinearity of $(A, B) \mapsto \mathrm{tr}(PAPB)$ applied to M_x in place of Σ_x. In regime (i) the second reading is the only non-trivial one, and it is the one §P.1 measures: there $\widehat{T}$ estimates $\mathbb{E}_{x,S}\mathbb{E}\|\Delta g\|^2$ exactly (Theorem 11, second half).

Slot exchangeability, and what a pooled audit cannot see. Given x the completions $y_1, \ldots, y_G$ are i.i.d. from $\pi_\theta(\cdot \mid x)$, so the joint law of $(y_{1:G}, \varepsilon, S)$ is invariant under permuting the G *slots*. Any *pooled* slot-space summary — an object formed by averaging a per-group $G \times G$ matrix over groups and then read slot by slot — therefore inherits that invariance in the population: writing Π for a permutation matrix, $\mathbb{E}[\Pi A\Pi^\top] = \mathbb{E}[A]$ for every Π, and the only symmetric matrices fixed by the whole symmetric group are

$$\bar{A} \;=\; a\,I + b\,\mathbf{1}\mathbf{1}^\top, \qquad a = \bar{d} - \bar{o}, \quad b = \bar{o}, \tag{7}$$

with $\bar{d}, \bar{o}$ the mean diagonal and mean off-diagonal. Three consequences. *(1)* $P\bar{\Sigma}P = a_\Sigma P$ and $P\bar{K}P = a_K P$: a pooled slot-space audit sees a *multiple of the projector* in contrast space and therefore **cannot see contrast anisotropy at all** — its $\mathcal{T}$ is $a_\Sigma a_K(G-1)$ whatever the underlying alignment. *(2)* Consequently the population value of $u^\top P\bar{K}Pu/(\mathrm{tr}(P\bar{K}P)/(G-1))$ for any contrast direction u held *fixed across slots* is exactly 1; any measured excess over 1 is finite-m noise, and if the statistic is formed as $\lambda_{\max}$ of a sample mean it is biased *upward* by Jensen, $\lambda_{\max}$ being convex (App. P.9, Q8). *(3)* What survives the symmetrisation is what is tied to a *feature* of the completion rather than to its slot index — answer format, length, solution strategy — because such a direction is carried by the completion through any relabelling. This is why $\widehat{T}$ evaluates the bilinear form at the *realized* pair (Σ_x, K_x) (§B.4) and why the pooled comparator of the $25\times$ construction is a population-level device rather than something a slot-exchangeable audit could report.

(ii) Independent stochastic judge calls. One independent call per completion gives a diagonal, heteroscedastic $\Sigma_{S,r}$ whose entries depend on y_i; off-diagonal structure appears only after marginalizing over prompts, i.e. in Σ_S and not in $\Sigma_{S,r}$.

(iii) Shared verifier randomness. This is the only regime with genuine off-diagonal *conditional* covariance. Suppose the grader draws one rubric or reference solution per group and that draw favours one of two answer formats. Write $u_i = \pm 1$ for the format completion i used and let a per-group latent ζ with $\mathbb{E}\zeta = 0$, $\operatorname{Var}\zeta = \gamma$ shift every score by ζu_i. Then $\operatorname{Cov}(\varepsilon \mid S, r) = \gamma\, uu^\top$ exactly — the rank-one contrast term of the designed families of App. G, which are therefore a mechanism rather than an assumption. A per-group leniency shift (the same ζ added to every score) gives $\beta \mathbf{1}\mathbf{1}^\top$ instead, and is annihilated by Corollary 6. In the completion-level language of Theorem 3 the first is $u(y)$ a format indicator, the second $u \equiv 1$.

Binary verifiers. The motivating case has $\mu \neq 0$: for $r \in \{0,1\}^G$ and rates $p_{\mathrm{FN}}, p_{\mathrm{FP}}$ that do not depend on the completion,

$$\mu = -(p_{\mathrm{FN}} + p_{\mathrm{FP}})\, r + p_{\mathrm{FP}} \mathbf{1}, \qquad P\mu = -(p_{\mathrm{FN}} + p_{\mathrm{FP}})\, Pr, \tag{8}$$

a pure *shrinkage* of g by $1 - (p_{\mathrm{FN}} + p_{\mathrm{FP}})$ rather than a rotation. Format-dependent false negatives make $P\mu$ non-parallel to Pr, and App. P.1's format drift is that rotation. Also $\Sigma_{S,r}$ depends on r, so the S-conditional statement must be read with Σ_S.

D THE MOMENT DECOMPOSITION: STATEMENTS IN FULL AND PROOFS

Throughout, $v := P\varepsilon$, $T := \mathbb{E}[\|\Delta g\|^2 \mid x] = G^{-2}\mathbb{E}[v^\top K v \mid x]$, and for slot-indexed vectors $\operatorname{Cov}_U(a,b) = G^{-1}\sum_i (a_i - \bar a)(b_i - \bar b)$. Every statement in this appendix carries at least one numerical assertion in `theory/near_orthogonal/check_near_orthogonal.py` (shipped in the supplement), which runs in the paper's container.

Proof of Theorem 2. Reduction to a mean-zero score. $SP\varepsilon = \sum_i s_i v_i$ with $\sum_i v_i = \mathbf{1}^\top P\varepsilon = 0$, so replacing every s_i by $s_i - c$ for a common c leaves $SP\varepsilon$ unchanged. With $c = \mu_s$, $\|SP\varepsilon\|^2 = \|\bar S P\varepsilon\|^2$ pathwise, where $\bar S = [\bar s_1, \dots, \bar s_G]$ and $\bar s_i = s(y_i, x) - \mu_s$ is still a fixed function of y_i with $\mathbb{E}[\bar s \mid x] = 0$. It therefore suffices to prove the theorem for a mean-zero score, which we do now, writing s for $\bar s$. With no hypothesis, $\|SP\varepsilon\|^2 = \varepsilon^\top PKP\varepsilon = \sum_{i,j,k,l} P_{ik} P_{jl}\, \varepsilon_k \varepsilon_l\, s_i^\top s_j$. The weights $P_{ab} = \delta_{ab} - 1/G$ are deterministic, so it suffices to evaluate $\mathbb{E}[\varepsilon_k \varepsilon_l\, s_i^\top s_j]$ for each partition of $\{i,j,k,l\}$ into equal values. Under Assumption 1 with ω_0 trivial the pairs (ε_a, s_a) are i.i.d. across a, so the value depends only on the partition, and any block containing a score index with no error index in the same block contributes a factor $\mathbb{E}[s] = 0$. Six sub-cases survive: $i = j = k = l$ (weight $((G-1)/G)^2$, count G, value A); $i = j$, $k = l$, $i \neq k$ (weight G^{-2}, count $G(G-1)$, value B); $i = j$ and k, l distinct from i and from each other (weight G^{-2}, count $G(G-1)(G-2)$, value F); $i = j = k \neq l$ and $i = j = l \neq k$ (weight $-(G-1)/G^2$ each, count $G(G-1)$ each, value D, total $-2(G-1)^2\mathsf{D}/G$); $i = k \neq j = l$ and $i = l \neq j = k$ (weights $((G-1)/G)^2$ and G^{-2}, count $G(G-1)$ each, value $\mathbb{E}[\varepsilon s]^\top \mathbb{E}[\varepsilon s] = \mathsf{C}$, total $(G-1)[(G-1)^2 + 1]\mathsf{C}/G$). Every other partition leaves an isolated score block: with $i \neq j$ and $k = l$ the factor is $\mathbb{E}[\varepsilon^2]\mathbb{E}[s_i]^\top \mathbb{E}[s_j]$ or $\mathbb{E}[\varepsilon^2 s]^\top \mathbb{E}[s]$; with $i \neq j$, $k \neq l$ and fewer than two coincidences between $\{i,j\}$ and $\{k,l\}$, some score is alone. Summing gives (6). The only surviving patterns with $i \neq j$ are the two C patterns, which proves that the off-diagonal part of $\mathbb{E}[v^\top K v]$ is exactly the C term and the diagonal part $\mathbb{E}\sum_i v_i^2 k_i$ is the rest.

For (3), write $(G-1)\mathsf{A} = (G-1)(\mathsf{A} - \mathsf{B}) + (G-1)\mathsf{B}$ and $-2(G-1)\mathsf{D} = -2(G-1)(\mathsf{D} - \mathsf{F}) - 2(G-1)\mathsf{F}$, so that $\frac{G}{G-1}\Theta_G = (G-1)[(\mathsf{A} - \mathsf{B}) - 2(\mathsf{D} - \mathsf{F})] + G(\mathsf{B} - \mathsf{F}) + (G^2 - 2G + 2)\mathsf{C}$. Now $\mathsf{B} - \mathsf{F} = \operatorname{Var}(\varepsilon)\mathbb{E}\|s\|^2$, $\mathsf{A} - \mathsf{B} = \operatorname{Cov}(\varepsilon^2, \|s\|^2)$ and $\mathsf{D} - \mathsf{F} = \bar\varepsilon \operatorname{Cov}(\varepsilon, \|s\|^2)$, and $\operatorname{Cov}((\varepsilon - \bar\varepsilon)^2, \|s\|^2) = \operatorname{Cov}(\varepsilon^2, \|s\|^2) - 2\bar\varepsilon \operatorname{Cov}(\varepsilon, \|s\|^2)$. Returning to a general score, $\mathbb{E}\|\bar s\|^2 = \operatorname{trCov}(s)$ and $\mathbb{E}[\varepsilon \bar s] = \operatorname{Cov}(\varepsilon, s)$, which gives (3). The coefficients and both forms were checked by exact rational enumeration of finite completion laws, and by Monte Carlo on synthetic laws that include a length-normalised score with $\mu_s \neq 0$. □

The same result as Hoeffding's variance formula. Pathwise, $\sum_{i \neq j} \frac{1}{2}(s_i - s_j)(\varepsilon_i - \varepsilon_j) = G\sum_i s_i \varepsilon_i - (\sum_i s_i)(\sum_j \varepsilon_j) = G\, SP\varepsilon$, so $U := SP\varepsilon/(G-1)$ is the degree-two U-statistic with kernel $h(z_1, z_2) = \frac{1}{2}(s_1 - s_2)(\varepsilon_1 - \varepsilon_2)$, $z = (\varepsilon, s)$, and $G^2\|\Delta g\|^2 = (G-1)^2\|U\|^2$. With $\theta = \mathbb{E}h = \operatorname{Cov}(\varepsilon, s)$, $\zeta_1 = \operatorname{trCov}(\mathbb{E}[h(Z_1, Z_2) \mid Z_1])$ and $\zeta_2 = \operatorname{trCov}(h)$, Hoeffding's formula (Hoeffding, 1948) $\mathbb{E}\|U\|^2 = \|\theta\|^2 + \binom{G}{2}^{-1}[2(G-2)\zeta_1 + \zeta_2]$ gives

$$G^2\mathbb{E}[\|\Delta g\|^2 \mid x] = (G-1)^2\|\operatorname{Cov}(\varepsilon, s)\|^2 + \tfrac{G-1}{G}\big[4(G-2)\zeta_1 + 2\zeta_2\big].$$

Because h is invariant under a common shift of the scores, take s centered: then $\mathbb{E}[h \mid Z_1] = \frac{1}{2}(s_1(\varepsilon_1-\bar{\varepsilon})+\theta)$, so $4\zeta_1 = \mathbb{E}[(\varepsilon-\bar{\varepsilon})^2\|s\|^2]-\|\theta\|^2 = \mathsf{A}-2\mathsf{D}+\mathsf{F}-\mathsf{C}$, and $\mathbb{E}\|h\|^2 = \frac{1}{2}(\mathsf{A}+\mathsf{B}-2\mathsf{D})+\mathsf{C}$, so $2\zeta_2 = \mathsf{A}+\mathsf{B}-2\mathsf{D}$. Substituting gives (6). The first term is the squared bias $G^2\|\mathbb{E}[\Delta g \mid x]\|^2$ ($\mathbb{E}[SP\varepsilon \mid x] = (G-1)\mathrm{Cov}(\varepsilon, s)$); the second is a variance, hence nonnegative, and it contains $-\frac{(G-1)(G-2)}{G}\mathsf{C}$, the difference between the hacking term of (3) and the bias. Assertions A92–A96 of the checking script verify each step in exact arithmetic.

The batch formula. For B prompts drawn i.i.d., one group each with independent completions, and $\Delta\bar{g} = B^{-1}\sum_b \Delta g_b$: $\mathbb{E}\|\Delta\bar{g}\|^2 = B^{-1}\mathbb{E}\|\Delta g\|^2 + (1-B^{-1})\|\mathbb{E}\Delta g\|^2$, since distinct prompts' perturbations are independent. With $m_x = \mathbb{E}[\Delta g \mid x]$, the part of $\Delta\bar{g}$ that is systematic given the prompts, $B^{-1}\sum_b m_{x_b}$, has $\mathbb{E}\|B^{-1}\sum_b m_{x_b}\|^2 = B^{-1}\mathbb{E}\|m_x\|^2 + (1-B^{-1})\|\mathbb{E}m_x\|^2$; dividing gives (4) with $b = G^2\mathbb{E}\|m_x\|^2$ and $\mathcal{S} = G^2\|\mathbb{E}m_x\|^2 \le b$ (Jensen). π_B increases in $\mathcal{S}$, so $\mathcal{S} = b$ (the same m_x for every prompt) gives the upper bound. The part that does not average out over batches has share $\sigma_B = B\mathcal{S}/(\bar{\Theta} + (B-1)\mathcal{S}) \le \pi_B$. Assertion A97 checks the formula by exact enumeration.

The uncentered parametrisation. If the five moments are computed from the raw score s instead of $\bar{s}$ (call them $\mathsf{A}^{\mathrm{r}}, \dots, \mathsf{F}^{\mathrm{r}}$), then with $\Gamma = \mathbb{E}[\varepsilon^2 s]^\top \mu_s$, $M_2 = \mathbb{E}[\varepsilon^2]\|\mu_s\|^2$, $H = \mathbb{E}[\varepsilon]\mathbb{E}[\varepsilon s]^\top \mu_s$, $M_1 = \mathbb{E}[\varepsilon]^2\|\mu_s\|^2$, exactly

$$G^2\mathbb{E}[\|\Delta g\|^2 \mid x] = \Theta_G(\mathsf{A}^{\mathrm{r}}, \dots, \mathsf{F}^{\mathrm{r}}) + \Xi,$$
$$\Xi = -\tfrac{2(G-1)^2}{G}\Gamma + \tfrac{(G-1)(G-2)}{G}M_2 - \tfrac{2(G-1)(G-2)^2}{G}H + \tfrac{(G-1)(G-2)(G-3)}{G}M_1.$$

This follows by substituting $\mathsf{A} = \mathsf{A}^{\mathrm{r}} - 2\Gamma + M_2$, $\mathsf{B} = \mathsf{B}^{\mathrm{r}} - M_2$, $\mathsf{C} = \mathsf{C}^{\mathrm{r}} - 2H + M_1$, $\mathsf{D} = \mathsf{D}^{\mathrm{r}} - 2H + M_1$, $\mathsf{F} = \mathsf{F}^{\mathrm{r}} - M_1$ into Θ_G. It also follows independently from the index expansion without discarding isolated score blocks. Every μ_s-term is off-diagonal: the diagonal part $\mathbb{E}\sum_i v_i^2 k_i$ equals $(G-1)\,\mathrm{Var}(\varepsilon)\mathbb{E}\|s\|^2 + \frac{(G-1)^2}{G}\mathrm{Cov}((\varepsilon-\bar{\varepsilon})^2, \|s\|^2)$ in raw moments for any μ_s. Ξ vanishes only when $\mu_s = 0$. Under $\varepsilon \perp s$ the off-diagonal part is $-(G-1)\,\mathrm{Var}(\varepsilon)\|\mu_s\|^2$, and the risk is $(G-1)\,\mathrm{Var}(\varepsilon)\mathrm{trCov}(s)$, not $(G-1)\,\mathrm{Var}(\varepsilon)\mathbb{E}\|s\|^2$.

Remark 1 (Two exact consistency checks). If $\varepsilon \perp s$ then $\mathsf{A} = \mathsf{B}$, $\mathsf{D} = \mathsf{F}$, $\mathsf{C} = \|\mathrm{Cov}(\varepsilon, s)\|^2 = 0$, and (3) is $(G-1)\,\mathrm{Var}(\varepsilon)\mathrm{trCov}(s)$, the isotropic formula $\kappa(G-1)\mathbb{E}[\mathrm{MSW}]/G^2$ for $K = \kappa I$. If $u \equiv 1$, the five moments of u are $\mathbb{E}\|\bar{s}\|^2, \mathbb{E}\|\bar{s}\|^2, 0, \mathbb{E}\|\bar{s}\|^2, \mathbb{E}\|\bar{s}\|^2$ and the coefficient sum $(G-1)^2/G + (G-1)/G - 2(G-1)^2/G + (G-1)(G-2)/G$ vanishes identically in G: common-mode annihilation (Corollary 6) recovered as a cancellation.

Proof of Theorem 3. $\|SP\varepsilon\|^2 = \|SP\varepsilon^{(0)}\|^2 + 2\omega(SP\varepsilon^{(0)})^\top(SPu) + \omega^2\|SPu\|^2$. Since $\omega \perp y_{1:G}$ and $\mathbb{E}\omega = 0$ the cross term has expectation 0, and $\mathbb{E}[\omega^2\|SPu\|^2] = \mathrm{Var}(\omega)\mathbb{E}\|SPu\|^2$. The vectors $\varepsilon^{(0)}$ and u are completion-level functions, so Theorem 2 applies to each separately. □

Why the total error cannot be substituted. The shared ω correlates ε_i across i, so Assumption 1 with ω_0 trivial fails for the total ε. Computing the five moments of the total error directly gives $\mathsf{A} = \mathsf{A}^{(0)} + \mathrm{Var}(\omega)\mathsf{A}_u$ and $\mathsf{B} = \mathsf{B}^{(0)} + \mathrm{Var}(\omega)\mathsf{B}_u$ but leaves $\mathsf{C}, \mathsf{D}, \mathsf{F}$ at their $\varepsilon^{(0)}$ values, silently dropping $\mathrm{Var}(\omega)$ times the $\mathsf{C}, \mathsf{D}, \mathsf{F}$ part of $\Theta_G(u)$.

Exchangeability. Under Assumption 1, for every permutation matrix Π_σ one has $(\Pi_\sigma\varepsilon, \Pi_\sigma K\Pi_\sigma^\top) \overset{d}{=} (\varepsilon, K)$ given x: condition on (x, ω_0), note that the pairs (y_i, ω_i) are i.i.d. and that ε is a coordinatewise and K a pairwise-symmetric image of them, then integrate over ω_0. A symmetric matrix fixed by every Π_σ is $aI + b\mathbf{1}\mathbf{1}^\top$ (transpositions equate the diagonal; S_G acts transitively on ordered pairs of distinct indices).

Proof of Corollary 9. For an equivariant functional A, $\bar{A} = \mathbb{E}[\Pi_\sigma A\Pi_\sigma^\top] = \Pi_\sigma\bar{A}\Pi_\sigma^\top$, hence $\bar{A} = aI + b\mathbf{1}\mathbf{1}^\top$. For fixed $u \perp \mathbf{1}$, $\|u\| = 1$: $u^\top\bar{K}u = a_K$ and $\mathrm{tr}(P\bar{K}P) = a_K(G-1)$. For the trace-normalised $N := PKP/\mathrm{tr}(PKP)$, equivariance and $N = PNP$ give $\mathbb{E}N = aP$ and $\mathrm{tr}N = 1$ gives $a = 1/(G-1)$, so the largest eigenvalue of any average of N_j's is at least $1/(G-1)$: a "fixed direction" statistic built as $(G-1)\lambda_{\max}(m^{-1}\sum_j N_j)$ has population value 1 and is ≥ 1 almost surely (App. P.9, Q8). The last sentence is the off-diagonal statement of Theorem 2. □

Proof of Proposition 4. (a) is Proposition 4's computation (App. G). (b) Under Assumption 1 with ω_0 trivial the errors are i.i.d., so $\mathbb{E}\|P\varepsilon\|^2 = (G-1)\,\mathrm{Var}(\varepsilon)$ and $\mathbb{E}[\mathrm{MSW}] = \mathrm{Var}(\varepsilon)$; and $\mathrm{tr}(P\bar{K}) = G\mathbb{E}\|s\|^2 - G^{-1}\mathbf{1}^\top\bar{K}\mathbf{1} = (G-1)(\mathbb{E}\|s\|^2 - \mathbb{E}\langle s_1, s_2\rangle) = (G-1)\mathrm{trCov}(s)$ because $\mathbb{E}\langle s_1, s_2\rangle = \|\mu_s\|^2$. Hence $G^2 B_{\mathrm{ICC}} = (G-1)\,\mathrm{Var}(\varepsilon)\mathrm{trCov}(s)$, the MSW term of (3), for any score mean. For the pooled plug-in, more generally and with no hypothesis, write $A := P\bar{M}P = \bar{a}P + \tilde{A}$ and $B := P\bar{K}P = \bar{b}P + \tilde{B}$ with $\bar{a} = \mathrm{tr}A/(G-1)$, $\bar{b} = \mathrm{tr}B/(G-1)$ and traceless $\tilde{A}, \tilde{B}$ supported on $\mathbf{1}^\perp$; expanding $\mathrm{tr}(AB)$ gives

$$R_{\mathrm{pool}} - B_{\mathrm{ICC}} = G^{-2}\langle \tilde{A}, \tilde{B}\rangle_F, \tag{9}$$

which vanishes whenever either $\bar{M}$ or $\bar{K}$ is compound symmetric, in particular under Assumption 1 by Corollary 9. Both baselines are functions of $(\bar{d}, \bar{o})$ and $\mathrm{trCov}(s)$ only, so they cannot see the coupling or the hacking term. □

The same fact, pathwise. Without Assumption 1, $T - R_{\mathrm{pool}} = G^{-2}\sum_{a,b}\mathrm{Cov}\big((P\varepsilon\varepsilon^\top P)_{ab}, K_{ab}\big)$ by $v^\top K v = \sum_{a,b}(P\varepsilon\varepsilon^\top P)_{ab}K_{ab}$: all deviation from the pooled scalar is covariance between the verifier's error and the completions' own score geometry, and it is zero whenever ε is independent of $y_{1:G}$ given x.

Proof of Lemma 10, and the U-statistics. $v^\top K v = (e - \bar{e}\mathbf{1})^\top K(e - \bar{e}\mathbf{1}) = e^\top K e - 2\bar{e}\,\mathbf{1}^\top K e + \bar{e}^2\,\mathbf{1}^\top K\mathbf{1}$ with $\bar{e} = S_e/G$; split $e^\top K e = \hat{\mathsf{A}} + \hat{\mathsf{C}}$ and $\mathbf{1}^\top K e = S_{ek} + \chi$. The expression for N is the same computation with K replaced by $\mathrm{diag}(k)$. With $S_{e2} = \sum_i e_i^2$, the per-group U-statistics

$$\hat{\mathsf{A}}_j = \tfrac{\hat{\mathsf{A}}}{G},\quad \hat{\mathsf{B}}_j = \tfrac{S_{e2}S_k - \hat{\mathsf{A}}}{G(G-1)},\quad \hat{\mathsf{C}}_j = \tfrac{\hat{\mathsf{C}}}{G(G-1)},\quad \hat{\mathsf{D}}_j = \tfrac{S_e S_{ek} - \hat{\mathsf{A}}}{G(G-1)},\quad \hat{\mathsf{F}}_j = \tfrac{(S_e^2 - S_{e2})S_k - 2S_e S_{ek} + 2\hat{\mathsf{A}}}{G(G-1)(G-2)}$$

average products over distinct slots of the raw score and are therefore unbiased for the uncentered moments $\mathsf{A}^{\mathrm{r}}, \ldots, \mathsf{F}^{\mathrm{r}}$ under Assumption 1. Averaging likewise $\hat{\Gamma}_j = \sum_{a\neq b} e_a^2 K_{ab}/(G(G-1))$, $\hat{M}_{2,j} = \sum_{a,b,r\neq} e_r^2 K_{ab}/(G(G-1)(G-2))$, $\hat{H}_j = \sum_{a,b,r\neq} e_a e_r K_{ab}/(G(G-1)(G-2))$ and $\hat{M}_{1,j} = \sum_{a,b,r,t\neq} e_r e_t K_{ab}/(G(G-1)(G-2)(G-3))$ (sums over pairwise distinct indices) gives unbiased estimates of Γ, M_2, H, M_1 for any μ_s. Then, *exactly and group by group*,

$$\Theta_G(\hat{\mathsf{A}}_j, \ldots, \hat{\mathsf{F}}_j) + \Xi(\hat{\Gamma}_j, \hat{M}_{2,j}, \hat{H}_j, \hat{M}_{1,j}) = v_j^\top K_j v_j,$$

with the $\hat{\mathsf{A}}, \hat{\mathsf{B}}, \hat{\mathsf{D}}, \hat{\mathsf{F}}$ part alone equal to N_j. So the gap between the uncentered reconstruction $\Theta_G(\hat{\mathsf{A}}_j, \ldots)$ and $v_j^\top K_j v_j$ is precisely the in-sample μ_s-terms $-\Xi(\hat{\Gamma}_j, \ldots)$, and E17's pooled ratio $\Theta_G(\hat{\mathsf{A}}, \ldots)/(G^2\widehat{T})$ equals $1 - \hat{\Xi}/(G^2\widehat{T})$ identically. On the stored audits the per-group identity holds to 5.0×10^{-16} (App. P.10). □

The per-group coupling identity and the four pooled channels. Fix a group with $v \neq 0$, $k_i > 0$, $\bar{k} = G^{-1}\sum_i k_i$, $\bar{o} = (G(G-1))^{-1}\sum_{a\neq b} K_{ab}$, $\nu = \mathrm{tr}(PKP)/(G-1) = \bar{k} - \bar{o}$. With $\alpha_i = G v_i^2/\|v\|^2$ (mean 1) and $\kappa := \mathrm{Cov}_U(\alpha, k)/\bar{k}$, exactly $N = \sum_i v_i^2 k_i = \|v\|^2\bar{k}(1+\kappa)$, because $\sum_i v_i^2 k_i = \|v\|^2\,\mathrm{mean}_U(\alpha k)$. Hence: $\kappa = 0$ iff the squared centered error is uncorrelated across slots with the score norm; $1 + \kappa = \sum_i \pi_i k_i/\bar{k}$ with $\pi_i = v_i^2/\|v\|^2$, and for centered v the *sharp* envelope is $1 + \kappa \in [\mu_{G-1}, \mu_1]/\bar{k}$, the extreme eigenvalues of $P\,\mathrm{diag}(k)\,P$ on $\mathbf{1}^\perp$ (a Rayleigh quotient restricted to $\mathbf{1}^\perp$); and if $\alpha_i = k_i/\bar{k}$ then $1 + \kappa = 1 + \mathrm{CV}^2(k)$, Kish's unequal-weighting factor (Kish, 1965) acting on contrasts.

The pooled bookkeeping. Across m audited groups, with $\widehat{T} = \mathrm{mean}_j G^{-2} v_j^\top K_j v_j$, $\widehat{T_{\mathrm{norm}}} = \mathrm{mean}_j G^{-2} N_j$ and $\widehat{B_{\mathrm{ICC}}} = \mathrm{MSW}\cdot\mathrm{tr}(P\bar{K})/G^2$, the identity

$$\widehat{T}/\widehat{B_{\mathrm{ICC}}} = (1+\hat{\theta})\,(1 + \kappa_{\mathrm{scale}} + \kappa_{\mathrm{within}})/(1-\varphi) \tag{10}$$

holds exactly (Proposition 14): $\hat{\theta} = \widehat{T}/\widehat{T_{\mathrm{norm}}} - 1$ is the off-diagonal channel, whose population value is the hacking term; κ_{within} is the within-group covariance of squared centered error with score norm, the finite-group counterpart of the coupling term; κ_{scale} is the classical across-prompt Jensen gap; and $|\varphi| \leq \max_j \delta_j$ is the baseline's own off-diagonal correction.

Proposition 14 (Pooled form: four exact channels). *For audited groups $j = 1, \dots, m$ let $\widehat{T_{\mathrm{norm}}} = \mathrm{mean}_j G^{-2} N_j$, $\widehat{B_{\mathrm{ICC}}} = G^{-2}(\mathrm{mean}_j \|v_j\|^2)(\mathrm{mean}_j \nu_j)$ (which is exactly $\mathrm{MSWtr}(P\bar{K})/G^2$), $q_j = \|v_j\|^2/\mathrm{mean}_j\|v_j\|^2$, $\kappa_{\mathrm{scale}} = \mathrm{Cov}_j(q_j, \bar{k}_j)/\mathrm{mean}_j \bar{k}_j$, $\kappa_{\mathrm{within}} = \mathrm{mean}_j[q_j \bar{k}_j \kappa_j]/\mathrm{mean}_j \bar{k}_j$, $\varphi = \mathrm{mean}_j \bar{o}_j/\mathrm{mean}_j \bar{k}_j$ and $\hat{\theta} = \widehat{T}/\widehat{T_{\mathrm{norm}}} - 1$. Then $\widehat{T_{\mathrm{norm}}}/\widehat{B_{\mathrm{ICC}}} = (1 + \kappa_{\mathrm{scale}} + \kappa_{\mathrm{within}})/(1 - \varphi)$ and (10) hold exactly, and $|\varphi| \le \max_j \delta_j$.*

Proof. $\mathrm{mean}_j[q_j \bar{k}_j(1 + \kappa_j)] = \mathrm{mean}_j \bar{k}_j + \mathrm{Cov}_j(q_j, \bar{k}_j) + \mathrm{mean}_j[q_j \bar{k}_j \kappa_j]$ because $\mathrm{mean}_j q_j = 1$, and $\mathrm{mean}_j \nu_j = (1 - \varphi)\,\mathrm{mean}_j \bar{k}_j$; multiply by $\widehat{T}/\widehat{T_{\mathrm{norm}}}$. Finally $|\bar{o}_j| \le \delta_j (G(G-1))^{-1} \sum_{a \ne b} \sqrt{k_a k_b} \le \delta_j \bar{k}_j$. □

Reading. $\hat{\theta}$ is the finite-sample off-diagonal channel, whose population value is the hacking term; κ_{within} is the within-group coupling, the finite-group counterpart of $\mathrm{Cov}((\varepsilon - \bar{\varepsilon})^2, \|s\|^2)$ (they are not algebraically identical: κ_j is built from $v = Pe$ in one group, A, B are single-completion moments); κ_{scale} is the classical across-prompt Jensen gap $\mathbb{E}[XY] \ne \mathbb{E}X\,\mathbb{E}Y$; φ is the baseline's own use of $\mathrm{tr}(P\bar{K})$. On the stored LLM audits the three identities hold to 5.4×10^{-16} relative (App. P.10).

Operator dependence. Under $S \to SD$, $D = \mathrm{diag}(d_i)$, $d_i > 0$: $k_i \to d_i^2 k_i$, v and α are unchanged, $\kappa^D = \mathrm{Cov}_U(\alpha, d^2 k)/\mathrm{mean}_U(d^2 k)$, and the cosine matrix $\mathcal{C} = D_k^{-1/2} K D_k^{-1/2}$ is invariant. If within a group $k_i = c\,\ell_i^{\beta}$ in the length $\ell_i = |o_i|$, $d_i = \ell_i^{-\gamma}$, and α is comonotone and non-constant in ℓ, then $\mathrm{sign}\,\kappa^{(\gamma)} = \mathrm{sign}(\beta - 2\gamma)$ (Chebyshev's sum inequality), so the sequence-level and length-normalised couplings have opposite signs iff $0 < \beta < 2$. On the stored audits the comonotonicity hypothesis is essentially absent within groups and the predicted flip was observed in 1 of 6 cells (App. P.10): the sign of the coupling is operator-dependent but not predictable from a fitted elasticity at $G = 8$.

E THE NEAR-ORTHOGONAL REGIME: STATEMENTS IN FULL AND PROOFS

Fix a group with $k_i > 0$; $\mathcal{C} = D_k^{-1/2} K D_k^{-1/2}$ ($\mathcal{C}_{ii} = 1$), $\mathcal{H} = \mathcal{C} - I$ (hollow, $\mathrm{tr}\mathcal{H} = 0$), $w = D_k^{1/2} v$, $N = \|w\|^2$ and $\delta = \max_{a \ne b} |\mathcal{C}_{ab}|$.

Proposition 15 (Coherence form). ***(i)*** *$v^\top K v = N + \sum_{a \ne b} w_a w_b \mathcal{C}_{ab}$.* ***(ii)*** *$|v^\top K v - N| \le \delta(\|w\|_1^2 - \|w\|_2^2) \le \delta(p-1)N \le \delta(G-1)N$, p the support size of w.* ***(iii)*** *(tightness) For G even, $k_i \equiv 1$, balanced $v \in \{\pm 1\}^G$ and $\mathcal{C} = (1-\delta)I + \delta v v^\top$ — a valid correlation matrix of coherence exactly δ — every inequality in (ii) is an equality.* ***(iv)*** *(common mode) With $\bar{c}$ the mean off-diagonal cosine, $v^\top K v = (1 - \bar{c})N + \bar{c}(\mathbf{1}^\top w)^2 + w^\top R' w$ with R' the residual off-diagonal part; for equal score norms and $\mathbf{1}^\top v = 0$ the common mode is a uniform shrinkage by $1 - \bar{c}$, never an alignment effect.*

Proof. (i) $v^\top K v = w^\top \mathcal{C} w$ and $\mathcal{C}_{aa} = 1$. (ii) the triangle inequality and Cauchy–Schwarz on the support. (iii) $\mathcal{C}$ has eigenvalues $1 - \delta$ and $1 - \delta + \delta G$; $w = v$, $N = G$ and $v^\top \mathcal{C} v - N = \delta(G^2 - G) = \delta(G-1)N$. (iv) substitute $\mathcal{C} = I + \bar{c}(\mathbf{1}\mathbf{1}^\top - I) + R'$ and use $\mathbf{1}^\top w = \sum_i v_i \sqrt{k_i}$, which is 0 for equal norms when $\mathbf{1}^\top v = 0$. □

Theorem 16 (Spectral form, in full). ***(i)*** *$v^\top K v - N = w^\top \mathcal{H} w$.* ***(ii)*** *$\lambda_{\min}(\mathcal{H})N \le v^\top K v - N \le \lambda_{\max}(\mathcal{H})N$, both attained at $v = D_k^{-1/2} u$ for the corresponding eigenvectors u; since $\mathrm{tr}\mathcal{H} = 0$, $\mathcal{H} \ne 0$ forces $\lambda_{\min} < 0 < \lambda_{\max}$.* ***(iii)*** *Under $\mathbf{1}^\top v = 0$, the case of the update, the sharp range is $[\lambda_{\min}, \lambda_{\max}]$ of the compression of $\mathcal{H}$ to $D_k^{1/2}(\mathbf{1}^\perp)$, which interlaces the spectrum of $\mathcal{H}$ and need not contain 0: for equal norms and $\mathcal{C} = (1-c)I + c\mathbf{1}\mathbf{1}^\top$ it is $\{-c\}$, so (ii)'s extremes are not attained and the channel has a sign.* ***(iv)*** *$\|\mathcal{H}\|_{\mathrm{op}} \le \|\mathcal{H}\|_F \le \delta\sqrt{G(G-1)}$ and $\|\mathcal{H}\|_{\mathrm{op}} \le (G-1)\delta$; the tempting $\|\mathcal{H}\|_F \le (G-1)\delta$ is false (the equicorrelated matrix gives $\delta\sqrt{G(G-1)}$).* ***(v)*** *For hollow $\mathcal{H}$, $\|\mathcal{H}\|_F/\sqrt{G} \le \|\mathcal{H}\|_{\mathrm{op}} \le \sqrt{(G-1)/G}\,\|\mathcal{H}\|_F$, the upper end attained by $c(\mathbf{1}\mathbf{1}^\top - I)$.* ***(vi)*** *With $r_j = w_j^\top \mathcal{H}_j w_j / N_j$, $(\widehat{T} - \widehat{T_{\mathrm{norm}}})/\widehat{T_{\mathrm{norm}}} = \langle r \rangle_N$ is an identity and Theorem 13's chain is a bound.* ***(vii)*** *$\mathcal{H}$ is invariant under $S \to SD$.*

Proof. (i) $w^\top(I+\mathcal{H})w$. (ii) a Rayleigh quotient, with w ranging over $\mathbb{R}^G$ because $D_k^{1/2}$ is invertible. (iii) the same restricted to a $(G-1)$-dimensional subspace, and Cauchy interlacing. (iv) $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$, Gershgorin, and entrywise. (v) with spectrum λ, $\|\lambda\|_\infty \ge \|\lambda\|_2/\sqrt{G}$, and maximising $\|\lambda\|_\infty/\|\lambda\|_2$ subject to $\sum\lambda = 0$ gives $\lambda \propto (G-1,-1,\dots,-1)$. (vi) average (i)–(ii) with weights N_j. (vii) D cancels in $D_k^{-1/2}KD_k^{-1/2}$. □

Remark 2 (What the spectral form does and does not explain). The identity in (vi) re-expresses the gap and is not evidence. The content is the chain $|\langle r\rangle_N| \le \langle\|\mathcal{H}\|_{\mathrm{op}}\rangle_N \le (G-1)\langle\delta\rangle_N$ (Gershgorin; also $\le \langle\|\mathcal{H}\|_F\rangle_N \le \sqrt{G(G-1)}\langle\delta\rangle_N$, which is looser), checked as assertions A98–A99, and the observation that most slack sits in the last link: δ is a maximum over $G(G-1)/2$ pairs, while the quadratic form sees a spectrum, and a hollow matrix with sign-cancelling entries has a much flatter spectrum than its largest entry. On the stored audits the spectral bound improves on the coherence bound by a factor 3.1–4.1, and is operator-invariant to 2.0×10^{-15} as (vii) requires. It does *not* explain why the realized Rayleigh quotient r_j lands near 0 rather than near an extreme eigenvalue: that is the alignment of $D_k^{1/2}Pe_j$ with the eigenvectors of $\mathcal{H}_j$ — the feature-tied structure the pathwise identity of App. D says cannot be assumed away. The spectral form turns a vacuous bound into a loose one; it does not close the gap.

A benchmark, not a model. If $s_1,\dots,s_G$ were i.i.d. with a rotationally invariant law on $\mathbb{R}^d$, then $\Pr(\delta \ge t) \le G(G-1)e^{-dt^2/2}$ (spherical concentration and a union bound), i.e. $\delta \lesssim 4.16/\sqrt{d}$ at $G=8$ with probability 0.99. Real score Grams are far more coherent than this benchmark (App. P.8), yet their signed mean cosine is close to 0, so the off-diagonal contribution is a sign-cancelling sum rather than a worst case. The benchmark fails, as it should, for tiny policies with heavy parameter sharing (the toy of App. P.1) and for designed constructions: the $47\times$ instance has same-strategy cosine $(\lambda/G)/(\kappa+\lambda/G) = 8.625/9.625 = 0.896$.

F Proof of Theorem 5 and its corollaries

Three statements used in the main text are given here in full.

Corollary 17 (Preconditioned updates: $S \to D_tS$)**.** *At step t write $\mathcal{G}_t = \mathcal{F}_{t-1} \vee \sigma(S_t, r_t)$ for the past together with the current group, and let the applied step be $D_t\tilde{g}_t$ with D_t $\mathcal{G}_t$-measurable (a learning-rate schedule, a fixed diagonal, or a second-moment estimate built from earlier steps). Assume the conditional law of ε_t given (S_t, r_t) is unchanged by conditioning further on $\mathcal{F}_{t-1}$ — automatic for a deterministic verifier. Then Theorem 5 and Proposition 18 hold verbatim with S replaced by D_tS; in particular* $\mathrm{trCov}(D_t\Delta g_t \mid \mathcal{G}_t) = G^{-2}\mathrm{tr}(P\Sigma_{S,r}PK^D)$ *with* $K^D = S^\top D_t^\top D_tS$.

Proof. D_t factors out of a conditional covariance it is measurable for: $\mathrm{Cov}(D_tM\varepsilon \mid \mathcal{G}_t) = D_tM\Sigma_{S,r}M^\top D_t^\top$ with $M = G^{-1}SP$, using the hypothesis to replace $\mathrm{Cov}(\varepsilon_t \mid \mathcal{G}_t)$ by $\Sigma_{S,r}$; and $\mathrm{tr}(D_tSP\Sigma PS^\top D_t^\top) = \mathrm{tr}(P\Sigma P\, S^\top D_t^\top D_tS)$. □

Proposition 18 (Spectral form, bounds, alignment)**.** *Let $P\Sigma P = \sum_{k=1}^{G-1}\lambda_k u_k u_k^\top$ on $\mathbf{1}^\perp$ (the* contrast eigendirections *u_k of the error), with $\lambda_1 \ge \cdots \ge \lambda_{G-1}$ and $\nu_1 \ge \cdots \ge \nu_{G-1}$ the sorted spectra of $P\Sigma P$ and of PKP there. Then*

$$\mathrm{trCov}(\Delta g \mid S, r) = G^{-2}\sum_k \lambda_k\|Su_k\|^2, \qquad w^\top\mathrm{Cov}(\Delta g \mid S, r)w = G^{-2}\sum_k \lambda_k\langle S^\top w, u_k\rangle^2. \tag{11}$$

Hence $\lambda_{\min}\mathrm{tr}(PK) \le G^2\mathrm{trCov} \le \lambda_{\max}\mathrm{tr}(PK)$, both attained; and with both sorted spectra fixed, G^2trCov takes every value in $[\sum_k \lambda_k\nu_{G-k}, \sum_k \lambda_k\nu_k]$ as the alignment varies over $SO(G-1)$.

Proposition 19 (Systematic perturbation: the reward-hacking gradient)**.** *Let the verifier error be a fixed function $\varepsilon(y)$ of the completion and let $y_1,\dots,y_G$ be i.i.d. from $\pi_\theta(\cdot \mid x)$ with $\mathbb{E}_\pi\|\nabla_\theta \log\pi_\theta\|^2 < \infty$, $\mathbb{E}_\pi\varepsilon^2 < \infty$ and the usual domination condition. Then*

$$\mathbb{E}_y[\Delta g \mid x] = \tfrac{G-1}{G}\,\nabla_\theta\,\mathbb{E}_{\pi_\theta}[\varepsilon(y) \mid x] = \tfrac{G-1}{G}\,\mathbb{E}[\varepsilon s], \tag{12}$$

so for the sequence-level score ($\mu_s = 0$) $\mathsf{C} = \|\mathbb{E}[\varepsilon s]\|^2$ of Definition 8 is the squared norm of the reward-hacking gradient. It fails for group-dependent verifiers (Remark 3).

Proof of Theorem 5. (i) By linearity, $\mathbb{E}[G^{-1}SP\varepsilon \mid S,r] = G^{-1}SP\,\mathbb{E}[\varepsilon \mid S,r] = G^{-1}SP\mu$, and $P\mathbf{1} = 0$ annihilates $\mu = b\mathbf{1}$. (ii) $\Delta g - \mathbb{E}[\Delta g \mid S,r] = M(\varepsilon - \mu)$ with $M = G^{-1}SP$ measurable with respect to (S,r); hence $\mathrm{Cov}(M\varepsilon \mid S,r) = M\Sigma_{S,r}M^\top = G^{-2}SP\Sigma_{S,r}P^\top S^\top$, and P is symmetric. For the trace, $\mathrm{tr}(SP\Sigma PS^\top) = \mathrm{tr}(P\Sigma P\,S^\top S) = \mathrm{tr}(P\Sigma PK)$. (iii) $M = G^{-1}SP$ is S-measurable, so $\mathrm{Cov}(M\varepsilon \mid S) = M\,\mathrm{Cov}(\varepsilon \mid S)\,M^\top$ with no assumption on μ; and by the law of total covariance $\mathrm{Cov}(\varepsilon \mid S) = \mathbb{E}[\mathrm{Cov}(\varepsilon \mid S,r) \mid S] + \mathrm{Cov}(\mathbb{E}[\varepsilon \mid S,r] \mid S) = \mathbb{E}[\Sigma_{S,r} \mid S] + \mathrm{Cov}(\mu \mid S) = \Sigma_S$. (iv) $\mathbb{E}\|\Delta g\|^2 = \mathrm{trCov} + \|\mathbb{E}\Delta g\|^2$. □

Only two conditional moments enter, so heavy tails and discreteness are irrelevant *for the identity*. Part (iii) is exact without an unbiasedness assumption; (ii) is finer, needed when bias and variance must be separated. On Corollary 17: an audit reporting $\widehat{T}$ for the raw gradient answers the raw-gradient question, and Adam's own D_t is built from a second-moment estimate that includes the current step's gradient. *With Adam's bias correction* the running second moment at step t is $\hat{v}_t = (1-\beta_2^t)^{-1}\sum_{u\le t}(1-\beta_2)\beta_2^{t-u}g_u^2$, whose weight on the current gradient is $(1-\beta_2)/(1-\beta_2^t)$; this tends to $1/t$ as $\beta_2 \to 1$ at fixed t, so it is *not* negligible at finite t and the corollary applies to Adam only in the joint limit $\beta_2 \to 1$, $t \to \infty$. Separately, **gradient-norm clipping is not covered**: scaling by $\min\{1, \tau/\|\tilde{g}\|\}$ makes the applied step a nonlinear function of $\tilde{r}$, so the multiplier is not $\mathcal{G}_t$-measurable and does not factor out of the conditional covariance.

Proof of Proposition 18. $\mathrm{tr}(SP\Sigma PS^\top) = \sum_k \lambda_k \mathrm{tr}(Su_ku_k^\top S^\top) = \sum_k \lambda_k \|Su_k\|^2$, and $w^\top Su_ku_k^\top S^\top w = (u_k^\top S^\top w)^2$. Both $P\Sigma P$ and PKP are PSD and vanish on $\mathrm{span}(\mathbf{1})$, so $\mathrm{tr}(P\Sigma P \cdot K) = \mathrm{tr}(P\Sigma P \cdot PKP)$ may be computed inside the $(G-1)$-dimensional space $\mathbf{1}^\perp$. On $\mathbf{1}^\perp$ we have $\lambda_{\min} I \preceq P\Sigma P \preceq \lambda_{\max} I$, and for PSD B, $A \preceq A'$ implies $\mathrm{tr}(AB) \le \mathrm{tr}(A'B)$ because $\mathrm{tr}((A'-A)B) = \mathrm{tr}(B^{1/2}(A'-A)B^{1/2}) \ge 0$; taking $A = \lambda_{\min} I$, $A' = \lambda_{\max} I$ and $\mathrm{tr}(I|_{\mathbf{1}^\perp} B) = \mathrm{tr}(PKP)$ gives the Rayleigh bounds, attained when $P\Sigma P$ is a multiple of P on $\mathbf{1}^\perp$ or when the range of PKP lies in the corresponding eigenspace. For the range statement, fix the eigenbases and write the achievable value as $f(U) = \mathrm{tr}(U\Lambda U^\top N)$ over $U \in SO(G-1)$, where Λ and N hold the two spectra. By von Neumann's (Ruhe's) trace inequality $\sum_k \lambda_k \nu_{G-k} \le f(U) \le \sum_k \lambda_k \nu_k$, and both endpoints are attained by the same-order and reversed-order eigenbasis alignments, which may be taken in $SO(G-1)$ by flipping the sign of one eigenvector. Since $SO(G-1)$ is connected for $G \ge 3$ and f is continuous, f attains every intermediate value. □

Proof of Corollary 6. ($\supseteq$) $P(\mathbf{1}a^\top + a\mathbf{1}^\top)P = (P\mathbf{1})a^\top P + (Pa)(\mathbf{1}^\top P) = 0$. ($\subseteq$) Suppose $P\Sigma P = 0$. Write $I = P + J$ with $J = \mathbf{1}\mathbf{1}^\top/G$ and expand $\Sigma = P\Sigma P + P\Sigma J + J\Sigma P + J\Sigma J$. With $a_0 = \Sigma\mathbf{1}/G$ we get $P\Sigma J = (Pa_0)\mathbf{1}^\top$, $J\Sigma P = \mathbf{1}(Pa_0)^\top$ and $J\Sigma J = \beta_0\mathbf{1}\mathbf{1}^\top$ with $\beta_0 = \mathbf{1}^\top a_0/G$; hence $\Sigma = b_0\mathbf{1}^\top + \mathbf{1}b_0^\top$ with $b_0 = Pa_0 + (\beta_0/2)\mathbf{1}$. The map $a \mapsto \mathbf{1}a^\top + a\mathbf{1}^\top$ is injective (its (i,i) entry is $2a_i$), so the kernel has dimension G. For the PSD statement, $P\Sigma P = 0$ says $v^\top \Sigma v = 0$ for all $v \perp \mathbf{1}$; for PSD Σ this forces $\Sigma v = 0$, so $\mathrm{range}(\Sigma) \subseteq \mathrm{span}(\mathbf{1})$ and $\Sigma = \tau^2\mathbf{1}\mathbf{1}^\top$. Pathwise: $P(z\mathbf{1} + \xi) = P\xi$. For a single fixed S, $SP\Sigma PS^\top = 0$ holds iff $\mathrm{range}(P\Sigma P) \subseteq \ker S$, which is weaker; quantifying over S spanning $\mathbb{R}^{d\times G}$ recovers the statement above. □

Proof of Corollary 7. $P[(1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top]P = (1-\rho)P + \rho(P\mathbf{1})(\mathbf{1}^\top P) = (1-\rho)P$, so $\mathrm{Cov} = G^{-2}\sigma^2(1-\rho)SPS^\top$, which is decreasing in the Loewner order, and its trace has slope $-\sigma^2\mathrm{tr}(PK)/G^2 < 0$ whenever $\mathrm{tr}(PK) > 0$ (i.e. unless all score vectors coincide). For this Σ, $\bar{d} = \sigma^2$ and $\bar{o} = \rho\sigma^2$, so ICC $= \rho$; and $\frac{d}{d\rho}\mathrm{ESS}(\rho) = -G(G-1)/(1+(G-1)\rho)^2 < 0$. The PSD range of ρ follows from the eigenvalues $\sigma^2(1-\rho)$ (multiplicity $G-1$) and $\sigma^2(1+(G-1)\rho)$. □

Proof of Proposition 19. Write $\bar{\varepsilon} = G^{-1}\sum_j \varepsilon_j$. Then $\Delta g = G^{-1}\sum_i (\varepsilon_i - \bar{\varepsilon})s_i$ and, taking expectations over i.i.d. $y_1, \dots, y_G$ and using $\mathbb{E}[s] = \mathbb{E}[\nabla_\theta \log \pi_\theta] = 0$,

$$\mathbb{E}[\varepsilon_i s_i] = \mathbb{E}[\varepsilon s], \qquad \mathbb{E}[\varepsilon_j s_i] = \mathbb{E}[\varepsilon]\mathbb{E}[s] = 0 \ (j \ne i),$$

so $\mathbb{E}[(\varepsilon_i - \bar{\varepsilon})s_i] = (1-G^{-1})\mathbb{E}[\varepsilon s]$ and $\mathbb{E}[\Delta g] = G^{-1}\cdot G\cdot(1-G^{-1})\mathbb{E}[\varepsilon s] = \frac{G-1}{G}\mathbb{E}[\varepsilon s]$. Finally $\mathbb{E}[\varepsilon(y)\nabla_\theta \log \pi_\theta(y)] = \nabla_\theta \mathbb{E}_{\pi_\theta}[\varepsilon(y)]$ by the score-function identity, valid under the stated integrability. □

Remark 3 (Scope of Proposition 19, and one heuristic). The identity extends to regimes (ii)–(iii) with $\varepsilon(y)$ replaced by $\mathbb{E}_\omega[\varepsilon \mid y]$ whenever ω is independent of y (condition on $y_{1:G}$ first and apply the same argument), and *fails* for group-dependent verifiers — majority-vote pseudo-labels, or a reference answer built from the sampled completions — because then ε_i is not a function of y_i alone and $\mathbb{E}[\varepsilon_j s_i] \neq 0$ for $j \neq i$. Separately, it is often said that the systematic part accumulates linearly over steps while the zero-mean part averages out at rate $(\text{steps} \times \text{batch})^{-1/2}$. That is a *heuristic*: training is non-stationary and Adam-preconditioned (Kingma & Ba, 2015), and neither rate is established here. What can be computed is the crossover $B\,T^* = \mathbb{E}\mathrm{tr}\mathrm{Cov}(\Delta g)/\|\mathbb{E}\Delta g\|^2$ — the number of group draws beyond which the systematic part dominates the zero-mean one — reported per checkpoint in App. P.3.

Remark 4 (RLOO). $A^{\mathrm{LOO}} = \frac{G}{G-1} P\tilde{r}$ exactly, so $\mathrm{Cov}(\Delta g^{\mathrm{LOO}} \mid S, r) = (G-1)^{-2} SP\Sigma PS^\top$: every statement in the paper transfers verbatim under this constant rescaling, and all ratios, orderings and sign-flip probabilities are unchanged.

Proof of Remark 4. $\sum_{j\neq i} \tilde{r}_j = G\bar{\tilde{r}} - \tilde{r}_i$, so $A_i^{\mathrm{LOO}} = \tilde{r}_i - (G-1)^{-1}(G\bar{\tilde{r}} - \tilde{r}_i) = \frac{G}{G-1}(\tilde{r}_i - \bar{\tilde{r}})$, i.e. $A^{\mathrm{LOO}} = \frac{G}{G-1} P\tilde{r}$. Hence $\tilde{g}^{\mathrm{LOO}} = G^{-1} S A^{\mathrm{LOO}} = (G-1)^{-1} SP\tilde{r}$ and the covariance carries a factor $(G/(G-1))^2$. Ratios and sign-flip probabilities are invariant to positive scaling. □

Proposition 20 (Sign flips). *With $C = G^{-2} SP\Sigma PS^\top$, $g \neq 0$, $\mathrm{SNR} = \|g\|^2/\sqrt{g^\top C g} > 0$ and the hypothesis $g^\top SP\mu = 0$ (no systematic perturbation along g): if $\Delta g \mid S, r \sim \mathcal{N}(0, C)$ then $\Pr(\langle \tilde{g}, g\rangle < 0 \mid S, r) = \Phi(-\mathrm{SNR})$; without any distributional assumption, $\Pr(\langle \tilde{g}, g\rangle < 0 \mid S, r) \leq 1/(1+\mathrm{SNR}^2)$, and this bound is the supremum over mean-zero laws with that variance (attained in the limit by a two-point law with an atom at the boundary; such a law with the prescribed projected variance exists). If $g^\top SP\mu \neq 0$ the Gaussian statement holds with* SNR *replaced by $\mathrm{SNR}' = (\|g\|^2 + G^{-1} g^\top SP\mu)/\sqrt{g^\top C g}$, and the distribution-free bound $1/(1+\mathrm{SNR}'^2)$ holds* provided $\mathrm{SNR}' > 0$ *(Cantelli's inequality is one-sided; at $\mathrm{SNR}' = -2$ the expression 0.2 is not a bound, the true probability being 0.977 for a two-point law). In the binary-shrinkage case* (8)*, $\mathrm{SNR}' > 0$ holds automatically whenever $p_{\mathrm{FN}} + p_{\mathrm{FP}} < 1$. For the z-scored update the same holds to first order with $g_{\mathrm{norm}} = G^{-1/2} S\hat{c}$ and C_{norm} from* (14)*, inheriting the error of Theorem 26.*

Proof of Proposition 20. $\langle \tilde{g}, g\rangle = \|g\|^2 + g^\top \Delta g$ with $\mathbb{E}[g^\top \Delta g \mid S, r] = G^{-1} g^\top SP\mu$, which vanishes under the hypothesis; then $g^\top \Delta g \sim \mathcal{N}(0, g^\top C g)$ in the Gaussian case, giving $\Phi(-\|g\|^2/\sqrt{g^\top C g})$. Without distributional assumptions, let $Y = -g^\top \Delta g$, mean 0 and variance $g^\top C g$; Cantelli's inequality gives $\Pr(Y \geq t) \leq \sigma_Y^2/(\sigma_Y^2 + t^2)$, and $t = \|g\|^2$ yields $1/(1+\mathrm{SNR}^2)$. The two-point law attaining Cantelli's bound places its mass at t and $-\sigma_Y^2/t$; since the event is the strict inequality $Y > t$, the bound is a supremum. The shifted statement follows by recentering Y and applying Cantelli with $t = \|g\|^2 + G^{-1} g^\top SP\mu$, which requires $t > 0$, i.e. $\mathrm{SNR}' > 0$. □

G Designed extremes: the 47× and 25× constructions

Two exact constructions show what the covariance identity permits for arbitrary families of (Σ_x, K_x). Both lie *outside* Assumption 1 — they correlate a slot-labelled sign pattern with the noise or with a cross-prompt pairing — and nothing comparable was observed on any language-model audit (§5). They are illustrations, not estimates of what real verifiers do.

G.1 A family the ICC cannot separate: 47×

For G even and balanced $u \in \{\pm 1\}^G$ set $\Sigma(u) = \alpha I + \beta \mathbf{1}\mathbf{1}^\top + \gamma uu^\top$ ($\alpha > 0$, $\beta, \gamma \geq 0$), the covariance generated by the per-group rubric mechanism of regime (iii) (App. C): every member shares the diagonal, the mean off-diagonal, hence the ICC, and the spectrum, differing only in the eigenvector of the dominant contrast, and $P\Sigma(u)P = \alpha P + \gamma uu^\top$. For scores take orthonormal $q_0, \ldots, q_G$, balanced h and $s_i = \sqrt{\kappa} q_i + \sqrt{\lambda/G}\, h_i q_0$ — idiosyncratic plus a shared "solution strategy" axis — so $K = \kappa I + (\lambda/G) hh^\top$.

Theorem 21 (Closed-form spread). $G^2\mathrm{trCov}(\Delta g \mid S, r) = \alpha[\kappa(G-1)+\lambda]+\gamma[\kappa G+(\lambda/G)(u^\top h)^2]$; *the maximum over balanced u is at $u = h$, the minimum at any balanced $u \perp h$, and*

$$\mathrm{Ratio} = \frac{\max_u \mathrm{trCov}}{\min_u \mathrm{trCov}} = 1 + \frac{(\gamma/\alpha)(\lambda/\kappa)G}{(G-1)+(\lambda/\kappa)+(\gamma/\alpha)G}. \tag{13}$$

Moreover $\mathrm{Ratio} \le \mathrm{cond}(PKP|_{\mathbf{1}^\perp})$ *for any $(\bar{d}, \bar{o})$-matched family, and* $\mathrm{Ratio} \le \min\{\mathrm{cond}(P\Sigma P|_{\mathbf{1}^\perp}), \mathrm{cond}(PKP|_{\mathbf{1}^\perp})\}$ *once the members also share the spectrum of $P\Sigma P$ (Remark 5). The common-mode weight β does not appear.*

In moment language. $\Sigma(u)$ is regime (iii) with $u(y_i) = h_i$ a strategy indicator and $\mathrm{Var}(\omega) = \gamma$ (Theorem 3); the score model makes that indicator coincide with the scores' shared axis, so the shared component's hacking moment $\mathsf{C}_u = \|\mathbb{E}[us]\|^2$ is as large as λ and G allow. The ratio compares "shared component aligned with the score axis" to "orthogonal to it", and it needs a same-strategy score cosine of $(\lambda/G)/(\kappa + \lambda/G) = 0.896$ (App. E).

The $47\times$ instance, in full. Take $G = 8$, $d = 32$, u over the seven non-constant Walsh–Hadamard columns of H_8, $\alpha = 1$, $\beta = 10$, $\gamma = 19$, $h = h^{(1)}$, $\kappa = 1$, $\lambda = 69$ (Table 4). Every member is positive definite with population ICC $17/70 = 0.242857$ and condition numbers $153, 70$; $G^2\mathrm{trCov}$ is 10716 at $u = h$ and 228 at $u \perp h$, a ratio of exactly $47 \le \min\{153, 70\}$, reproduced by Monte Carlo as 46.967 (CI $[46.888, 47.048]$) at estimated ICCs agreeing to 0.36% (Figure 8); over all matched families the ratio is *unbounded* (Proposition 22). Equation (13) says how much of $47\times$ is the construction: $\gamma/\alpha = \lambda/\kappa = 1$ gives $1.50\times$ for *any* G, and $47\times$ needs $\gamma/\alpha = 19$ *and* $\lambda/\kappa = 69$ together; substituting the anisotropy actually *measured* on either policy gives far less (App. P.5).

Proof of Proposition 4. $y_g^\top P y_g = \varepsilon_g^\top P \varepsilon_g$ since $P\mathbf{1} = 0$, and for ε with mean proportional to $\mathbf{1}$ and covariance Σ, $\mathbb{E}[\varepsilon^\top P \varepsilon] = \mathrm{tr}(P\Sigma) = \mathrm{tr}\Sigma - G^{-1}\mathbf{1}^\top\Sigma\mathbf{1} = (G-1)(\bar{d} - \bar{o})$; dividing by $G-1$ gives $\mathbb{E}[\mathrm{MSW}] = \bar{d} - \bar{o} = \mathrm{tr}(P\Sigma P)/(G-1)$ (if $\mathbb{E}\varepsilon = \mu$ is not proportional to $\mathbf{1}$, MSW gains $\mu^\top P\mu/(G-1)$). The group mean $\bar{y}_g = a_g + \mathbf{1}^\top\varepsilon_g/G$ has variance $\sigma_a^2 + G^{-2}\mathbf{1}^\top\Sigma\mathbf{1} = \sigma_a^2 + (\bar{d} + (G-1)\bar{o})/G$ provided a_g is uncorrelated with ε_g (hence with $\mathbf{1}^\top\varepsilon_g$), and MSB is G times an unbiased estimate of it. The a.s. limits follow from the strong law applied to i.i.d. summands, and $(\mathrm{MSB} - \mathrm{MSW})/(\mathrm{MSB} + (G-1)\mathrm{MSW}) \to (G\sigma_a^2 + G\bar{o})/(G\sigma_a^2 + G\bar{d})$. With group-dependent Σ_g the same holds with Σ replaced by $\mathbb{E}_g\Sigma_g$. □

Properties of the sign-vector family and Theorem 21. For balanced $u \in \{\pm 1\}^G$: $\mathbf{1}$ and u are orthogonal eigenvectors of $\Sigma(u) = \alpha I + \beta\mathbf{1}\mathbf{1}^\top + \gamma uu^\top$ with eigenvalues $\alpha + G\beta$ and $\alpha + G\gamma$, and α on $\{\mathbf{1}, u\}^\perp$, so $\Sigma(u) \succeq 0$ iff $\alpha \ge 0$, $\alpha + G\beta \ge 0$, $\alpha + G\gamma \ge 0$. The diagonal is $\alpha+\beta+\gamma u_i^2 = \alpha+\beta+\gamma$, independent of u; $\sum_{i\ne j} u_i u_j = (\mathbf{1}^\top u)^2 - G = -G$ gives $\bar{o} = \beta - \gamma/(G-1)$, also independent of u; hence the ICC is common to the family. Since $Pu = u$ and $P\mathbf{1} = 0$, $P\Sigma(u)P = \alpha P + \gamma uu^\top$. For the scores of §G.1, $K_{ij} = \kappa\delta_{ij} + (\lambda/G)h_i h_j$ by orthonormality, so $\mathrm{tr}(PK) = \kappa(G-1) + \lambda$ and $u^\top K u = \kappa G + (\lambda/G)(u^\top h)^2$. Therefore $\mathrm{tr}(P\Sigma(u)PK) = \alpha\mathrm{tr}(PK) + \gamma\, u^\top K u$. Over balanced u, $(u^\top h)^2$ is maximal ($= G^2$) at $u = \pm h$ and zero for $u \perp h$; substituting and taking the ratio gives (13). For the bound, Proposition 18 gives $\lambda_{\min}\mathrm{tr}(PKP) \le G^2\mathrm{trCov} \le \lambda_{\max}\mathrm{tr}(PKP)$ for every member (all members share the spectrum here), so the ratio is at most $\lambda_{\max}/\lambda_{\min} = \mathrm{cond}(P\Sigma P|_{\mathbf{1}^\perp})$; exchanging the roles of Σ and K in the same argument gives $\mathrm{cond}(PKP|_{\mathbf{1}^\perp})$. □

Remark 5 (The $\mathrm{cond}(P\Sigma P)$ half needs a shared spectrum). For a family whose members share the spectrum of $P\Sigma P$ — the sign-vector family above — Proposition 18 bounds the ratio by $\mathrm{cond}(P\Sigma P|_{\mathbf{1}^\perp})$ as well. For a family merely matched in $(\bar{d}, \bar{o})$ this half can fail, and the following counterexample (checked numerically) shows it. Take $G = 8$, $a, b > 0$, $h \in \{\pm 1\}^G$ balanced, $v = h/\sqrt{G}$, and

$$\Sigma_1 = aI + b\mathbf{1}\mathbf{1}^\top, \qquad \Sigma_2 = (G-1)a\, vv^\top + (a/G + b)\mathbf{1}\mathbf{1}^\top, \qquad K = I + (5/G)hh^\top.$$

Both are PSD, both have $\bar{d} = a + b$ and $\bar{o} = b$, hence the same ICC; $P\Sigma_1 P = aP$ while $P\Sigma_2 P = (G-1)a\, vv^\top$. Then $G^2\mathrm{trCov}$ is $a\mathrm{tr}(PK) = 12a$ for Σ_1 and $(G-1)a\, v^\top K v = 42a$ for Σ_2, a ratio of 3.5, whereas $\mathrm{cond}(P\Sigma_1 P|_{\mathbf{1}^\perp}) = 1$. The $\mathrm{cond}(PKP|_{\mathbf{1}^\perp}) = 6$ half does hold, and does so for *every* $(\bar{d}, \bar{o})$-matched family, because $\mathrm{tr}(P\Sigma P) = (G-1)(\bar{d} - \bar{o})$ is then common to the family and Proposition 18 applied with the roles of Σ and K exchanged gives $\nu_{\min}\mathrm{tr}(P\Sigma P) \le G^2\mathrm{trCov} \le$

$\nu_{\max}\mathrm{tr}(P\Sigma P)$. More generally the ratio over any matched family is at most $\max_i \lambda^{(i)}_{\max}/\min_i \lambda^{(i)}_{\min}$ over its members.

Proposition 22 (The ratio is unbounded)**.** *Over all* $(G$*, families with matched* $(\bar{d}, \bar{o})$*,* $S)$ *the supremum of* $\max_u \mathrm{trCov}/\min_u \mathrm{trCov}$ *is* $+\infty$*: in* (13) *fix* $\gamma = \lambda = 1$*,* $G = 8$ *and let* $\alpha = \kappa = t \to 0^+$ *with* $\beta(t)$ *chosen to hold the ICC at any fixed target; the ratio is* $1 + 8/(7t^2 + 9t) \to \infty$*, both matrices becoming singular in the limit.*

Proof of Proposition 22. With $\gamma = \lambda = 1$, $G = 8$ and $\alpha = \kappa = t$, (13) reads $1+8/(t(7t+1)+8t) = 1+8/(7t^2+9t)$, which diverges as $t \to 0^+$ (values 1.5, 9.25, 89.2, 889.2 at $t = 1, 10^{-1}, 10^{-2}, 10^{-3}$). Positive definiteness and a fixed target ICC are preserved by choosing $\beta = \beta(t) > 0$ appropriately: at $G = 8$, $\beta = \alpha/3 + 11\gamma/21$ holds the ICC at $1/4$, and any target in $(0, 1)$ is achievable similarly. Consistently with (13), both $\mathrm{cond}(P\Sigma P)$ and $\mathrm{cond}(PKP)$ diverge in this limit. □

Per-direction version. With $w = g/\|g\|$ and $g = G^{-1}SPr$, (11) gives $w^\top \mathrm{Cov} w = G^{-2}[\alpha\|PS^\top w\|^2 + \gamma(u^\top S^\top w)^2]$ because $P\Sigma(u)P = \alpha P + \gamma uu^\top$. Since $S^\top w \propto KPr$, the maximizing member is the one aligned with the *clean reward contrast*, not with the score strategy direction h. In the $47\times$ instance with $r = h^{(2)}$ one has $Kh^{(2)} = \kappa h^{(2)}$, so $w^\top \mathrm{Cov} w = 0.015625$ for $u = h^{(1)}$ against 2.390625 for $u = h^{(2)}$: a $153\times$ per-direction spread with the ordering *reversed* relative to the trace criterion. Both are correct; they answer different questions (total gradient noise versus the noise that can flip the update sign, Proposition 20).

G.2 Pooled summaries: $25\times$

Write $\Sigma_x := \mathrm{Cov}(\varepsilon \mid x, S)$ and $K_x := S^\top S$ (in regime (i) read M_x), and set $R = \mathbb{E}_{x,S}[G^{-2}\mathrm{tr}(P\Sigma_x P K_x)]$, $\bar{\Sigma} = \mathbb{E}_{x,S}\Sigma_x$, $\bar{K} = \mathbb{E}_{x,S}K_x$ and $R_{\mathrm{pool}} = G^{-2}\mathrm{tr}(P\bar{\Sigma}P\bar{K})$.

Theorem 23 (Non-identifiability from pooled summaries)**.** $R - R_{\mathrm{pool}} = G^{-2}\mathbb{E}_{x,S}[\mathrm{tr}(P(\Sigma_x - \bar{\Sigma})P(K_x - \bar{K}))]$*: R is not a function of* $(\bar{\Sigma}, \bar{K})$ *but also needs the across-prompt covariance between the centered verifier covariance and the centered score Gram.*

The instance. Take $G = 8$, $\Sigma(u) = I + 4\mathbf{1}\mathbf{1}^\top + 6uu^\top$, $K(h) = I + (55/8)hh^\top$ with $u, h \in \{h^{(1)}, h^{(2)}\}$ and two populations of two equiprobable prompt types, ALIGNED pairing (Σ^A, K^A) with (Σ^B, K^B) and ANTI pairing (Σ^A, K^B) with (Σ^B, K^A). The marginal laws agree, so $\bar{\Sigma}$, $\bar{K}$, the marginal variances and the per-prompt ICC ($2/7 = 0.285714$) coincide; yet $R = 42.96875$ versus 1.71875, exactly $25\times$ (Monte Carlo 25.006, $[24.959, 25.053]$). A pooled plug-in that keeps the slot labels returns 1430 for both; but completions are exchangeable within a prompt (Corollary 9), so a real pooled audit sees the symmetrised summaries and returns 487.14 ($= 3410/7$) for both. The $25\times$ gap between the true risks survives symmetrisation; the slot-fixed comparator does not. By (9) the gap such a pooled comparator can exploit is $\langle \tilde{A}, \tilde{B}\rangle_F$, which vanishes as soon as the score Gram is slot-exchangeable in the mean. *In moment language*, the two populations have identical within-prompt moments $\mathsf{A}, \ldots, \mathsf{F}$ and differ only in how (Σ_x, K_x) are paired across prompts — the across-prompt analogue of κ_{scale} in Proposition 14, a second-order covariance that neither Theorem 2 (which conditions on x) nor a pooled audit sees. The estimator comparison of Table 11 uses the slot-fixed comparator 1430; it was not re-run against the symmetrised comparator, so we claim no advantage over a realisable pooled audit from it (a fairer comparison is App. P.4).

H Proofs for the z-score normalization

Theorem 24 (Radial reformulation and first-order expansion)**.** $A(\tilde{r}) = \sqrt{G}\, P\tilde{r}/\|P\tilde{r}\|$ *whenever* $P\tilde{r} \neq 0$*; with* $c = Pr \neq 0$*,* $\hat{c} = c/\|c\|$*,* $Q = I - \hat{c}\hat{c}^\top$*,* $\eta = \sqrt{\mathrm{tr}(P\Sigma P)}/\|Pr\|$ *and* $\varrho = \|P\varepsilon\|$*: **assume*** $\mathbb{E}[\varepsilon \mid S, r] = 0$ ***and a finite*** $\kappa_4 = \mathbb{E}\varrho^4/(\mathrm{tr}P\Sigma P)^2$*. Then an explicit deterministic remainder bound gives*

$$\mathrm{Cov}(\Delta g_{\mathrm{norm}} \mid S, r) = (G\|c\|^2)^{-1} S\, QP\Sigma PQ\, S^\top + O\big(G\kappa_4^{1/2}\eta^3\big), \tag{14}$$

where QP projects onto $\{\mathbf{1}, c\}^\perp$; hence $QP\hat{c} = 0$ kills first-order noise along $\hat{c}$, and $\mathbb{E}[\Delta g_{\mathrm{norm}} \mid S, r] = O(\eta^2)$ even for an unbiased verifier.

The statement in full. This appendix gives the explicit constants and the $\|S\|_{\mathrm{op}}^2$ transfer to the gradient (Theorem 26), together with the measured validity regime. In full: normalization discards magnitude and keeps only the *direction* of the centered reward vector in $\mathbf{1}^\perp$ (Lemma 25); A linearizes as $A(r+\varepsilon) = \sqrt{G}\hat{c} + \frac{\sqrt{G}}{\|c\|} QP\varepsilon + \mathcal{R}$ with $\|\mathcal{R}\| \le 4\sqrt{G}\varrho^2/\|c\|^2$ for $\varrho = \|P\varepsilon\| \le \frac{1}{2}\|c\|$, so $\Delta g_{\mathrm{norm}} = G^{-1}S(A(r+\varepsilon) - A(r))$ obeys (14); and A being nonlinear is what makes $\mathbb{E}[\Delta g_{\mathrm{norm}} \mid S, r] = O(\eta^2)$ even for an unbiased verifier.

Where it fails, measured. We test the *suppression ratio* (15) — noise along $\hat{c}$ relative to its value without the second projection, predicted to be 0 — against Theorem 26's trace-error bound $26G\kappa_4^{1/2}\eta^3 + 160G\kappa_4\eta^4$, $\kappa_4 = \mathbb{E}\varrho^4/(\mathrm{tr}P\Sigma P)^2$. The relative error of (14) crosses 5% at $\eta = 0.132$ for the anisotropic sign family and 0.263 for isotropic Σ (Table 9). *Binary rewards with flip noise break it*, consistently with that bound: independent flips at rate p with n_+ of G completions correct give $\eta = \sqrt{pG(G-1)/(n_+(G-n_+))}$ and $\kappa_4 \approx 1/(Gp)$, so $\kappa_4^{1/2}\eta$ is constant in p and nothing improves as $p \to 0$ at $n_+ = 1, 4$ ($n_+ = 2$ is non-monotone, Table 10); across all flip cells SR = 0.74–0.99, i.e. no suppression. If all completions are correct or all wrong, $c = 0$ and $A(\tilde{r})$ is pure noise *however small the noise*; dropping such groups (up to 4.8% here) is a selection effect that is not Σ-free, and the $+\delta$ stabilizer makes the second projection partial.

The suppression ratio quoted in §B.3 and in the experiment records is, in full,

$$\mathrm{SR} = \frac{\mathrm{Var}\left(\hat{c}^\top[A(r+\varepsilon) - A(r)]\right)}{\mathrm{Var}\left((\sqrt{G}/\|c\|)\,\hat{c}^\top P\varepsilon\right)} = \frac{\|c\|^2\,\hat{c}^\top \mathrm{Cov}(A(r+\varepsilon))\hat{c}}{G\,\hat{c}^\top\Sigma\hat{c}}, \tag{15}$$

the noise along $\hat{c}$ relative to what it would be without the second projection.

Lemma 25 (Exact radial reformulation)**.** *Whenever $P\tilde{r} \neq 0$, $A(\tilde{r}) = \sqrt{G}\,P\tilde{r}/\|P\tilde{r}\|$, so $\|A(\tilde{r})\| = \sqrt{G}$ identically; when $P\tilde{r} = 0$ we set $A(\tilde{r}) = 0$ (implementations drop the group).*

Proof of Lemma 25. $\mathrm{sd}(\tilde{r}) = \|P\tilde{r}\|/\sqrt{G}$ under the $1/G$ convention, so $A(\tilde{r}) = P\tilde{r}/\,\mathrm{sd}(\tilde{r}) = \sqrt{G}P\tilde{r}/\|P\tilde{r}\|$ and $\|A\| = \sqrt{G}$. Under the $1/(G-1)$ convention every statement holds with $\sqrt{G}$ replaced by $\sqrt{G-1}$ and covariances scaled by $(G-1)/G$; ratios are unchanged. With a stabilizer $\delta > 0$, $A_\delta = P\tilde{r}/(\mathrm{sd}(\tilde{r}) + \delta)$ has derivative $(\mathrm{sd}+\delta)^{-1}[I - \frac{\mathrm{sd}}{\mathrm{sd}+\delta}\hat{c}\hat{c}^\top]P$ at $\tilde{r} = r$: the second projection is partial, with suppression factor $\delta/(\mathrm{sd}+\delta)$, and the $c = 0$ singularity disappears. □

Proof of Theorem 24. Since $\hat{c} \in \mathrm{range}(P)$ we have $P\hat{c} = \hat{c}$, so $PQ = P - \hat{c}\hat{c}^\top = QP$, which is symmetric and idempotent ($(P - \hat{c}\hat{c}^\top)^2 = P - \hat{c}\hat{c}^\top$) with range $\mathbf{1}^\perp \cap c^\perp$. Differentiating $z \mapsto \sqrt{G}z/\|z\|$ at $z = c$ gives $(\sqrt{G}/\|c\|)(I - \hat{c}\hat{c}^\top)$, and composing with P gives the stated linearization.

For the remainder, normalize $\|c\| = 1$ (both sides scale identically under $(c, e) \mapsto (tc, te)$ and A is invariant under positive rescaling of $\tilde{r}$). Write $e = P\varepsilon = a\hat{c} + e_\perp$, $\varrho = \|e\|$, $|a| \le \varrho$, and note $QP\varepsilon = e_\perp$. With $z = c + e$ and $L = \|z\|$, $L^2 = (1+a)^2 + \|e_\perp\|^2 = 1 + 2a + \varrho^2$, and

$$\mathcal{R}/\sqrt{G} = \left(\tfrac{1+a}{L} - 1\right)\hat{c} + \left(\tfrac{1}{L} - 1\right)e_\perp.$$

Assume $\varrho \le 1/2$; then $L \ge 1 - \varrho \ge 1/2$, $1 + a \ge 1/2$ and $L \le 3/2$. Since $(1+a)^2 - L^2 = -\|e_\perp\|^2$, we get $|(1+a)/L - 1| = \|e_\perp\|^2/(L(1+a+L)) \le 2\varrho^2$; and $|1-L| = |2a+\varrho^2|/(1+L)$ gives $|1/L - 1|\,\|e_\perp\| \le \varrho(2\varrho+\varrho^2)/(L(1+L)) \le \frac{4}{3}\varrho^2(2+\varrho) \le \frac{10}{3}\varrho^2$. Hence $\|\mathcal{R}\|/\sqrt{G} \le \varrho^2\sqrt{4 + 100/9} \le 4\varrho^2$. Globally, $\|A(r+\varepsilon)\| = \sqrt{G}$ by Lemma 25, so $\|\mathcal{R}\| \le 2\sqrt{G} + (\sqrt{G}/\|c\|)\|P\varepsilon\|$ with no smallness assumption.

The covariance statement follows by substituting the linearization into $\Delta g_{\mathrm{norm}} = G^{-1}S(A(r+\varepsilon) - A(r))$ and using $(QP)^\top = PQ$. The exact expansion of the trace is $\mathrm{tr}(PQ\Sigma QPK) = \mathrm{tr}(P\Sigma PK) - 2\hat{c}^\top\Sigma PK\hat{c} + (\hat{c}^\top\Sigma\hat{c})(\hat{c}^\top K\hat{c})$, obtained by expanding $Q = I - \hat{c}\hat{c}^\top$ twice and using $P\hat{c} = \hat{c}$. The order-η^2 bias is $G^{-1}S\,\mathbb{E}[\mathcal{R}]$, bounded by the same remainder estimate. □

Theorem 26 (Accuracy of the covariance approximation)**.** *Let $\eta = \sqrt{\mathrm{tr}(P\Sigma P)}/\|Pr\|$, assume $\mathbb{E}[\varepsilon \mid S, r] = 0$ and $\mathbb{E}\varrho^4 \le \kappa_4(\mathrm{tr}P\Sigma P)^2$, and let $C_{\mathrm{true}} = \mathrm{Cov}(X)$, $C_{\mathrm{lin}} = \mathrm{Cov}(L_0) =$*

$(G/\|c\|^2)QP\Sigma PQ$ *for* $X = A(r+\varepsilon) - \sqrt{G}\hat{c}$ *and its linearization* L_0. *Then* $\mathbb{E}\|\mathcal{R}\|^2 \le 160G\kappa_4\eta^4$,

$$\|C_{\text{true}} - C_{\text{lin}}\|_F \le 26G\kappa_4^{1/2}\eta^3 + 160G\kappa_4\eta^4, \qquad |\text{tr}C_{\text{true}} - \text{tr}C_{\text{lin}}| \le 26G\kappa_4^{1/2}\eta^3 + 160G\kappa_4\eta^4,$$

so the relative error of the trace is $O(\eta)$ *provided* $\gamma_\perp = \text{tr}(QP\Sigma PQ)/\text{tr}(P\Sigma P)$ *is bounded away from* 0. *Multiplying by* $G^{-2}\|S\|_{\text{op}}^2$ *transfers the absolute bound to the gradient covariance; the* relative *transfer carries the extra factor* $\|S\|_{\text{op}}^2\text{tr}(C_{\text{lin}})/\text{tr}(SC_{\text{lin}}S^\top)$, *which is not bounded for anisotropic* S. *If the law of* ε *is symmetric about* 0 *then* $\mathbb{E}[L_0\mathcal{R}^\top]$ *vanishes at third order and the rate improves to* $O(\eta^2)$.

Proof of Theorem 26. Let $E = \{\varrho \le \|c\|/2\}$. On E, the remainder bound gives $\mathbb{E}[\mathbf{1}_E\|\mathcal{R}\|^2] \le 16G\,\mathbb{E}\varrho^4/\|c\|^4 \le 16G\kappa_4\eta^4$. On E^c, use the global bound and $(x+y)^2 \le 2x^2+2y^2$: $\mathbb{E}[\mathbf{1}_{E^c}\|\mathcal{R}\|^2] \le 8G\,\Pr(E^c) + 2(G/\|c\|^2)\mathbb{E}[\mathbf{1}_{E^c}\varrho^2]$. Markov at the fourth moment gives $\Pr(E^c) \le 16\mathbb{E}\varrho^4/\|c\|^4 \le 16\kappa_4\eta^4$, and Cauchy–Schwarz gives $\mathbb{E}[\mathbf{1}_{E^c}\varrho^2] \le 4\kappa_4\eta^4\|c\|^2$; hence $\mathbb{E}[\mathbf{1}_{E^c}\|\mathcal{R}\|^2] \le 136G\kappa_4\eta^4$ and the total is at most $160G\kappa_4\eta^4$. For the covariance error, write $X = L_0 + \mathcal{R}$ with $\mathbb{E}L_0 = 0$; then $\text{Cov}(X) - \text{Cov}(L_0) = \mathbb{E}[L_0\mathcal{R}^\top] + \mathbb{E}[\mathcal{R}L_0^\top] + \mathbb{E}[\mathcal{R}\mathcal{R}^\top] - (\mathbb{E}\mathcal{R})(\mathbb{E}\mathcal{R})^\top$, and Cauchy–Schwarz bounds each Frobenius norm, giving $\|C_{\text{true}} - C_{\text{lin}}\|_F \le 2(\mathbb{E}\|L_0\|^2)^{1/2}(\mathbb{E}\|\mathcal{R}\|^2)^{1/2} + \mathbb{E}\|\mathcal{R}\|^2$. Substituting $\mathbb{E}\|L_0\|^2 \le G\eta^2$ yields the stated bound; the trace version follows from the same Cauchy–Schwarz chain applied to traces rather than to Frobenius norms (the alternative route $|\text{tr}M| \le \sqrt{\text{rank}(M)}\|M\|_F$ costs an extra $\sqrt{G-1}$ and does not give the stated constant). Dividing by $\text{tr}C_{\text{lin}} = (G/\|c\|^2)\text{tr}(QP\Sigma PQ) = G\gamma_\perp\eta^2$ gives relative error $O(\eta)$. For a symmetric law, L_0 is odd and the leading term of $\mathcal{R}$ is even in ε, so their cross-moment vanishes at third order, leaving $\mathbb{E}[\mathcal{R}\mathcal{R}^\top] = O(\eta^4)$ and relative error $O(\eta^2)$. □

Binary rewards. For $r \in \{0,1\}^G$ with n_+ correct completions, $c = Pr$ takes the two values $1 - n_+/G$ and $-n_+/G$, so $\|c\|^2 = n_+(G-n_+)/G$, maximal at $n_+ = G/2$ and minimal at $n_+ \in \{1, G-1\}$. Independent flips at rate p give $\varepsilon_i \in \{-1, 0, 1\}$ with $\text{Var}(\varepsilon_i) = p(1-p)$, so $\eta = \sqrt{pG(G-1)/(n_+(G-n_+))}$ up to the $(1-p)$ factor: at $G = 8$, $n_+ = 4$, reaching $\eta \le 0.1$ requires $p \lesssim 0.003$, far below realistic verifier flip rates. Moreover the number of flips in a group is $\text{Bin}(G, p)$, so $\mathbb{E}\varrho^4 \asymp Gp$ while $(\text{tr}P\Sigma P)^2 \asymp ((G-1)p)^2$ for small p, giving $\kappa_4 \asymp 1/(Gp)$ and hence $\kappa_4^{1/2}\eta \to \sqrt{(G-1)/(n_+(G-n_+))}$, a constant. The failure of the approximation in this regime is thus *consistent with* Theorem 26 rather than a counterexample to it: a non-vanishing upper bound cannot predict a failure, it merely fails to exclude one. The direct reason is simpler: conditional on there being any flip, $\|P\varepsilon\|/\|c\|$ is $O(1)$ and never small, so both C_{true} and C_{lin} scale like p and their relative discrepancy is p-free. Finally $\mathbb{E}[\varepsilon \mid S, r] = 0$ also fails here (eq. 8).

I Proofs for pooling and the estimator

Proof of Theorem 23. Write $\Sigma_x = \bar{\Sigma} + D_x$, $K_x = \bar{K} + E_x$ with $\mathbb{E}_{x,S}D_x = 0$, $\mathbb{E}_{x,S}E_x = 0$. By bilinearity of $(A, B) \mapsto \text{tr}(PAPB)$, $\text{tr}(P\Sigma_x PK_x) = \text{tr}(P\bar{\Sigma}P\bar{K}) + \text{tr}(P\bar{\Sigma}PE_x) + \text{tr}(PD_xP\bar{K}) + \text{tr}(PD_xPE_x)$. Taking $\mathbb{E}_{x,S}$, the two middle terms vanish because the maps $E \mapsto \text{tr}(P\bar{\Sigma}PE)$ and $D \mapsto \text{tr}(PDP\bar{K})$ are linear. □

The closed form referred to in Theorem 11 holds **for mean-zero Gaussian** $e \mid x, S$ **only** and reads, with $T_{x,S} = G^{-2}\text{tr}(P\Sigma_x PK_x)$,

$$\text{Var}(\widehat{T}) = m^{-1}\Big[\mathbb{E}_{x,S}\,2\text{tr}\big((P\Sigma_x PK_x)^2\big)/G^4 + \text{Var}_{x,S}(T_{x,S})\Big] \qquad (e \mid x, S \sim \mathcal{N}(0, \Sigma_x)). \quad (16)$$

It applies to none of the deterministic-verifier sections of this paper. In regime (i) the audit residual is a deterministic function of the completions, so $e \mid x, S$ is neither Gaussian nor mean-zero (§2, App. C); §P.1, §5 and every appendix reporting them therefore use bootstrap intervals and never (16). Only the Monte Carlo cells of §K with Gaussian residuals satisfy its hypothesis.

Proof of Theorem 11. Condition on (x, S) and write $\|SPe\|^2 = e^\top PKPe$. For any random vector e with mean μ and covariance Σ and any symmetric M, $\mathbb{E}[e^\top Me] = \text{tr}(M\Sigma) + \mu^\top M\mu$; with $M = PKP$ and $P^2 = P$ this is $\text{tr}(P\Sigma PK) + \|SP\mu\|^2$. Dividing by G^2, taking $\mathbb{E}_{x,S}$ and averaging over the m i.i.d. groups gives both statements. For the variance, if $e \sim \mathcal{N}(0, \Sigma)$ then

$\mathrm{Var}(e^\top Me) = 2\mathrm{tr}(M\Sigma M\Sigma)$, and with $M = PKP$, $\mathrm{tr}(PKP\Sigma PKP\Sigma) = \mathrm{tr}((P\Sigma PK)^2)$; the law of total variance adds the across-prompt term $\mathrm{Var}_{x,S}(T_{x,S})$ and the m^{-1} factor, both finite by the assumed second moments. Finally $\mathrm{tr}(A^2) \le (\mathrm{tr}A)^2$ for PSD A gives $\mathrm{Var}(\widehat{T}) \le 3m^{-1}\mathbb{E}_{x,S}T_{x,S}^2$, so the relative standard error is $O(m^{-1/2})$ with constant $\sqrt{3\mathbb{E}T^2/(\mathbb{E}T)^2}$; equality in the first inequality holds iff $P\Sigma PK$ is rank one. □

Proof of Proposition 12. For $j \neq j'$ the groups are independent, so $\mathbb{E}\langle z_j, z_{j'}\rangle = \|\mathbb{E}z\|^2 = G^2\|\mathbb{E}\Delta g\|^2$, and dividing by G^2 and averaging over the $m(m-1)$ ordered pairs preserves the expectation. Since $\mathbb{E}[G^{-2}\|z\|^2] = \mathbb{E}\|\Delta g\|^2 = \mathrm{trCov}(\Delta g) + \|\mathbb{E}\Delta g\|^2$ with *both* moments taken over the full draw (x, S, ε), the difference $\widehat{T} - \widehat{T}^{\mathrm{sys}}$ has expectation $\mathrm{trCov}(\Delta g)$, the trace of the **marginal** covariance. By the law of total covariance $\mathrm{trCov}(\Delta g) = \mathbb{E}_x\mathrm{trCov}(\Delta g \mid x) + \mathrm{trCov}_x\big(\mathbb{E}[\Delta g \mid x]\big)$, and the second term is the across-prompt spread of the prompt-specific systematic error — by Proposition 19 the spread of the per-prompt reward-hacking gradients. So $\widehat{T} - \widehat{T}^{\mathrm{sys}}$ is *not* "the noise part": it contains all prompt-specific hacking that averages to nothing across prompts, and equals $\mathbb{E}_x\mathrm{trCov}(\Delta g \mid x)$ only when $\mathbb{E}[\Delta g \mid x]$ is the same for every x. (If the audit groups share a prompt, the statement holds conditionally on x, with $\mathbb{E}\Delta g$ replaced by $\mathbb{E}[\Delta g \mid x] = \frac{G-1}{G}\nabla_\theta \mathbb{E}_\pi[\varepsilon \mid x]$; that is the within-prompt variant reported in App. P.3, and it is the one that does isolate conditional noise.) □

Proposition 27 (Sketching, with exact inflation)**.** *Let $\Omega_j \in \mathbb{R}^{k\times d}$ be drawn i.i.d.* per group*, independent of (e, S), with $\mathbb{E}[\Omega^\top\Omega] = I_d$. Then $\widehat{T}_\Omega = m^{-1}\sum_j G^{-2}\|\Omega_j z_j\|^2$ is unbiased for R and consistent in m. For a Gaussian sketch and a prompt with $\mathcal{T} = \mathrm{tr}(P\Sigma PK)$, $V = \mathrm{tr}((P\Sigma PK)^2)$, $G^4\,\mathrm{Var}(\widehat{T}_\Omega) = 2V(1 + 2/k) + 2\mathcal{T}^2/k$, an exact per-group variance inflation $\mathrm{IF}(k) = 1 + (2 + \mathcal{T}^2/V)/k$ bracketed by $1 + 3/k \le \mathrm{IF}(k) \le 1 + (G+1)/k$; the lower end is attained when $P\Sigma PK$ is rank one — the aligned, high-risk case. If instead a* single Ω *is reused for all groups, $\widehat{T}_\Omega$ remains unbiased but is no longer consistent: with $C_{\mathrm{tot}} = G^{-2}\mathbb{E}[zz^\top]$, $\mathrm{Var}_\Omega(\mathbb{E}[\widehat{T}_\Omega \mid \Omega]) = 2\|C_{\mathrm{tot}}\|_F^2/k$, a floor no increase of m removes. Measured over predicted, this floor is $0.90/1.06/1.00/0.92/1.01$ for $k = 1, 2, 4, 16, 64$ (App. P.4).*

Proposition 28 (Directional variant)**.** *For fixed $w \in \mathbb{R}^d$, $\widehat{T}_w = G^{-2}(w^\top SPe)^2$ satisfies $\mathbb{E}[\widehat{T}_w] = w^\top\mathrm{Cov}(\Delta g \mid S)w$ when $\mathbb{E}[\varepsilon \mid S, x] = 0$, and $\mathbb{E}[(w^\top\Delta g)^2]$ in general. It needs only the G scalars $w^\top s_i = \frac{d}{dt}\big|_0 \log\pi_{\theta+tw}(y_i \mid x)$, one JVP per completion — no backward pass, no d-dimensional storage.*

Proof of Proposition 27. Unbiasedness: $\mathbb{E}\|\Omega SPe\|^2 = \mathbb{E}[e^\top PS^\top\mathbb{E}[\Omega^\top\Omega]SPe] = \mathbb{E}\|SPe\|^2$. For the Gaussian sketch $\Omega = W/\sqrt{k}$, condition on $z = SPe$: $Wz \sim \mathcal{N}(0, \|z\|^2 I_k)$, so $\|\Omega z\|^2 = (\|z\|^2/k)\chi^2_k$ with mean $\|z\|^2$ and conditional variance $2\|z\|^4/k$. By the law of total variance $\mathrm{Var}(\|\Omega z\|^2) = (2/k)\mathbb{E}\|z\|^4 + 2V$, and for the Gaussian quadratic form $\|z\|^2 = e^\top PKPe$ we have $\mathbb{E}\|z\|^2 = \mathcal{T}$, $\mathrm{Var}(\|z\|^2) = 2V$, hence $\mathbb{E}\|z\|^4 = \mathcal{T}^2 + 2V$; substituting gives $2V(1 + 2/k) + 2\mathcal{T}^2/k$. The bracket follows from $V \le \mathcal{T}^2 \le (G-1)V$ by Cauchy–Schwarz on the at most $G-1$ nonzero eigenvalues of $P\Sigma PK$. A Rademacher sketch also satisfies $\mathbb{E}[\Omega^\top\Omega] = I_d$, so unbiasedness is preserved. For the shared-sketch statement, put $C_{\mathrm{tot}} = G^{-2}\mathbb{E}[zz^\top]$; then $\mathbb{E}[\widehat{T}_\Omega \mid \Omega] = \mathrm{tr}(\Omega^\top\Omega\, C_{\mathrm{tot}})$ for every m, and for a Gaussian $\Omega = W/\sqrt{k}$ with rows w_a, $\mathrm{tr}(\Omega^\top\Omega C_{\mathrm{tot}}) = k^{-1}\sum_{a\le k} w_a^\top C_{\mathrm{tot}} w_a$, a sum of k i.i.d. quadratic forms each of variance $2\|C_{\mathrm{tot}}\|_F^2$; hence $\mathrm{Var}_\Omega = 2\|C_{\mathrm{tot}}\|_F^2/k$, independent of m. □

Why the $\hat{\Sigma}$-plug-in route is dominated. (1) Σ_x has $G(G+1)/2$ free parameters per prompt type, of which only the $G(G-1)/2$-dimensional quotient is identifiable (Corollary 6); estimating it needs many audited groups *per type*, while $\widehat{T}$ is unbiased from a single group. (2) Practical audits pool, and then Theorem 23 makes the plug-in biased — in the $25\times$ construction the slot-fixed comparator over-estimates one population by $13.0\times$ and under-estimates the other by $1.92\times$ (the symmetrised comparator of App. G.2 is biased too, but by different factors). (3) A covariance estimate discards the $\|SP\mu_x\|^2$ term, which $\widehat{T}$ includes. (4) Cost: $O(dG)$ versus $O(dG^2)$ compute and $O(G^2)$ storage per group.

The sketch must be *fresh* per group (App. P.4), and the dominant cost is gold labels on $m \times G$ completions — a few hundred groups sufficed in the $d = 32$ construction but, on the learned policy, only while the verifier still errs often.

Proof of Proposition 28. $w^\top \Delta g = G^{-1} w^\top SPe$; squaring and taking expectations and using Theorem 5(ii)/(iv) contracted with w gives the claim. The scalars $w^\top s_i$ are directional derivatives of $\log \pi_\theta(y_i \mid x)$ along w, computable by a single JVP per completion. □

J CONSTRUCTION PARAMETERS

Table 4: Exact parameters of the two designed constructions (App. G). Both use Walsh–Hadamard columns of H_8 as sign vectors, and both instantiate the same closed form (13). The slot-fixed pooled comparator is a population-level device; a slot-exchangeable audit forms the symmetrised one.

	47× (§G.1)	25× (§G.2)
G, d	8, 32	8, 32
$\Sigma(u) = \alpha I + \beta \mathbf{1}\mathbf{1}^\top + \gamma u u^\top$	$\alpha = 1, \beta = 10, \gamma = 19$	$\alpha = 1, \beta = 4, \gamma = 6$
eigenvalues of Σ	$1^{(6)}$, 81, 153	$1^{(6)}$, 33, 49
$K = \kappa I + (\lambda/G) h h^\top$	$\kappa = 1, \lambda = 69$	$\kappa = 1, \lambda = 55$
marginal variance $\bar{d}$	30	11
mean off-diagonal $\bar{o}$	$51/7$	$22/7$
population ICC	$17/70 = 0.242857$	$2/7 = 0.285714$
$\mathrm{tr}(PK)$	76	62
$\mathrm{cond}(P\Sigma P\|_{\mathbf{1}^\perp})$, $\mathrm{cond}(PKP\|_{\mathbf{1}^\perp})$	153, 70	49, 56
varies over	$u \in \{h^{(1)}, \ldots, h^{(7)}\}$ at one prompt	joint law of (Σ_x, K_x)
$G^2\times$ risk, high / low	10716 / 228	2750 / 110
pooled plug-in, slot-fixed	—	1430 (both populations)
pooled plug-in, symmetrised	—	487.14 (both populations)
ratio	47 exactly	25 exactly

K THE PRE-SPECIFIED SIMULATION SUITE

Protocol. The claims of §4 and App. G are identities, expansions and exact constructions, so this suite is numerical verification, not benchmark comparison. The unit of replication is the seed (20 seeds, 200,000 draws per cell, over four residual laws and a (G, d) grid; App. L). All 22 criteria, their thresholds and the test level 0.05 were fixed before the runs and nothing was re-thresholded; the protocol files carry self-reported SHA-256 digests and *no external timestamp*, so "pre-specified" means fixed in advance by us, never registered with a third party. 17 criteria are *implementation checks* the theorems already imply (15 pass; the 2 failures, C4b, C6a, are pass rules over many uncorrected simultaneous intervals); only 4 are *falsifiable hypotheses* the theorems do not imply, of which 1 passes. All 5 literal failures are reported, including the 10% suppression threshold on every one of the 6 anisotropic sign-family cells (§B.3). *Deviations:* two Monte Carlo design flaws were corrected post hoc, the pre-specified $\max / \min$ spread estimator was replaced by a ratio-of-means estimand, and one t_5 cell was extended by 40 seeds *after* its interval was seen to miss 47.0 — each flagged post hoc, with its effect on the reported number, in App. P.

What the numbers show. Table 6 has every measurement: the covariance identity recovered to 0.0176 relative Frobenius error over 36 cells, and $\widehat{T}$ beating the slot-fixed pooled plug-in in all 18 RMSE cells — one construction-induced bias, one fact not eighteen, and no win on a fairer baseline (App. P). Whether the one-step geometry predicts *learning* was a pre-specified secondary hypothesis; its falsifiable half is a powered null (App. O).

L SIMULATION DESIGN

In the $47\times$ and $25\times$ studies the construction is deterministic, so the 20 seeds are independent batches of 200,000 draws and the seed-level interval is a batch-means interval. The pooled plug-in of §B.4 is

Table 5: What is random at which level in each study. "per seed" means redrawn once per seed and held fixed across the draws of that seed; "per draw" means redrawn for each of the 200,000 Monte Carlo draws. Intervals are t-intervals over the 20 seed-level estimates (19 degrees of freedom) unless stated otherwise.

study	fixed by construction	per seed	per draw
identity	G, d, Σ type	S (i.i.d. Gaussian entries, then scaled), Σ (Wishart cells)	ε
reversal	G, d, σ^2, ρ grid	S — *redrawn per ρ cell* (see §K)	ε
z-score	$G = 8$, $d = 32$, Σ shape, η grid, r type	S, Σ (Wishart cells)	ε (or the flip pattern)
$47\times$	the full construction	— (seeds are independent batches)	ε
$25\times$	the full construction	prompt-type draws	ε
estimator	the two populations, m grid, k grid	sketches Ω, audit composition	e, audit groups
sign flips	G, d, SNR grid	S, Σ	ε
learning	task, policy class, G, learning rate	policy init., prompt stream, verifier noise	— (one trajectory per seed)

formed by averaging the per-group empirical residual covariance over the m audit groups and the per-group Gram matrix likewise, then evaluating $G^{-2}\mathrm{tr}(P\hat{\tilde{\Sigma}}P\hat{\tilde{K}})$.

M ALL PRE-SPECIFIED CRITERIA

Table 6: The eight studies and their headline measurements. "Criteria" counts those passing *as literally written*; I/F marks implementation checks versus falsifiable hypotheses. Targets, thresholds and Monte Carlo standard errors per criterion: Table 7.

	kind	headline measurement	criteria
covariance identity	I	max rel. Frobenius error 0.0176, 36 cells (thr. 0.02)	1/1
exchangeable reversal	I	slope < 0, CI excludes 0, 12/12; RLOO ratio $(G/(G-1))^2$ to 3.4×10^{-15}	4/4
z-score validity	F	rel. error ≤ 0.033 at $\eta \leq 0.1$ (thr. 0.05), 10/10; suppression fails 6/6	2/3
$47\times$ spread	I	MC 47.095/47.134/47.064 (N/R/t_5) vs 47.0	1/2
$25\times$ pooling gap	I	MC 25.006/25.021/25.043 vs 25.0; pooled identical to 0	2/2
audit estimator	I	RMSE ratio 0.0021–0.477 vs plug-in; inflation matches 32/32	3/4
sign flips	F	Gaussian error 0.00103 (thr. 0.01); Rademacher 0.0811	3/3
learning toy	F	common mode 0.0066 vs contrast 0.0856; aligned vs orthogonal: null	1/3

Two headline measurements of Table 6 are, in full: the Gaussian sign-flip law is recovered to 0.0010 absolute, but Rademacher and t_5 are off by up to 0.0811 and 0.0092, so the flip *formula*, unlike the identity, is not distribution free (Figure 11); and the first-order z-score trace holds within 5% over the regime of §B.3 (Table 9).

Table 7 lists all 22 criteria exactly as fixed before the runs, with the literal verdict, the classification of §K (I = implementation check, F = falsifiable hypothesis, R = report-only) and, where a seed-level interval is available, the implied Monte Carlo standard error. 17 pass and 5 fail as written.

Table 7: The 22 pre-specified success criteria and their literal verdicts. Thresholds and $\alpha = 0.05$ were fixed at design time; nothing was re-thresholded. "s.e." is the Monte Carlo standard error of the seed-level mean where the analysis records an interval.

ID	kind	study	pre-specified criterion	s.e.	verdict
C1	I	identity	rel. Frobenius error of $\widehat{\mathrm{Cov}}(\Delta g)$ vs $G^{-2}SP\Sigma PS^\top \leq 0.02$ in every $(G, \mathrm{law}, \Sigma)$ cell	—	pass
C2a	I	reversal	slope of trCov in ρ negative with 95% CI excluding 0 in every (G, law) cell, while ESS decreases	0.088	pass
C2b	I	reversal	RLOO trace / centered trace $= (G/(G-1))^2$	—	pass
C2c	I	reversal	MC trace matches $G^{-2}\sigma^2(1-\rho)\mathrm{tr}(PK)$	0.0000	pass
C2d	I	reversal	sanity: $\mathrm{ESS}(0.530, G{=}8) = 1.70$; isotropic $\rho{=}0$ trace within 2% of $G^{-2}\sigma^2\mathrm{tr}(PK)$	—	pass
C3a	F	z-score	rel. error of the first-order trace ≤ 0.05 for $\eta \leq 0.1$ on the primary cells	—	pass
C3b	F	z-score	suppression ratio $\mathrm{SR} < 0.1$ for $\eta \leq 0.1$	—	**FAIL**
C3c	I	z-score	$\|\mathcal{R}\| \leq 4\sqrt{G}\varrho^2$ (Thm. 24)	0.0001	pass
C4a	I	47×	$\max_u \mathrm{trCov} / \min_u \mathrm{trCov}$ within ±5% of 47.0	0.042	pass
C4b	I	47×	every pairwise ΔICC 95% CI contains 0, all 21 pairs × 3 laws	—	**FAIL**
C5a	I	25×	$R_{\mathrm{ALIGNED}}/R_{\mathrm{ANTI}}$ within ±5% of 25.0, every law	0.0225	pass
C5b	I	25×	the two populations' pooled $(\bar{\Sigma}, \bar{K})$ agree to 10^{-12}	—	pass
C6a	I	estimator	95% CI of the bias of $\widehat{T}$ contains 0 in every (population, m, law) cell	—	**FAIL**
C6b	I	estimator	$\mathrm{RMSE}(\widehat{T}) < \mathrm{RMSE}(\text{pooled plug-in})$ on both populations	—	pass
C6c	I	estimator	sketch unbiased and inflation matches $1 + (2 + \mathcal{T}^2/V)/k$ to 10% in all 32 cells	—	pass
C6d	I	estimator	$\widehat{T}_w$ unbiased for $w^\top \mathrm{Cov}(\Delta g \mid S) w$ (Prop. 28)	—	pass
C7a	I	sign flips	$\lvert\widehat{\Pr}(\langle \tilde{g}, g\rangle < 0) - \Phi(-\mathrm{SNR})\rvert \leq 0.01$ under Gaussian residuals	—	pass
C7b	I	sign flips	Cantelli bound $1/(1+\mathrm{SNR}^2)$ holds in every cell and law	—	pass
C7c	R	sign flips	report (no threshold) the deviation from $\Phi(-\mathrm{SNR})$ under Rademacher and t_5	—	pass
C8a	F	learning	final clean reward: aligned < orthogonal, paired by seed, 95% CI excluding 0	—	**FAIL**
C8b	F	learning	final clean reward: aligned < common mode, paired by seed	—	**FAIL**
C8c	I	learning	design validity: marginal variance matched across noisy conditions; ICC matched aligned vs orthogonal	—	pass

N ADDITIONAL EXPERIMENTAL DETAIL

This appendix collects the supporting figures and per-cell tables behind §B.3 and §K: the validity regime of the first-order normalized-GRPO covariance (Figure 10, Tables 9 and 10), the pooling construction and the audit estimator (Figure 9, Table 11), the sign-flip law (Figure 11) and the seven-member $47\times$ family (Table 8). In the binary-flip cells the measured relative error of the first-order formula is flat in the flip rate at $n_+ = 1$ (26.4%–31.6%) and $n_+ = 4$ (7.6%–12.6%) over $p \in [0.005, 0.2]$, and non-monotone at $n_+ = 2$ (0.5%–7.6%, dipping at intermediate p), where the first-order *trace* is accurate while the suppression is still absent; $p = 5\%$ at $n_+ = 4$ already means $\eta \approx 0.41$. The stored quantity is an absolute error, which conflates the $k = 2$ dip with genuine accuracy; we say so rather than interpret it.

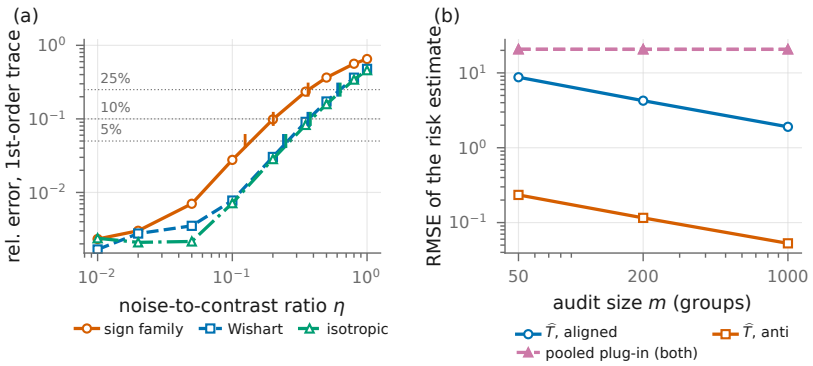

Figure 7: **(a)** measured validity regime of the first-order normalized-GRPO covariance (14): relative error of the predicted trace against η for three Σ shapes, $5/10/25\%$ crossings ticked. **(b)** RMSE of the direct estimator $\widehat{T}$ and of the pooled $\hat{\Sigma}$ plug-in against the audit size m, on the two populations of the $25\times$ construction: $\widehat{T}$ falls at $m^{-1/2}$, the plug-in is stuck at its bias. Both populations are adversarial for the plug-in by construction; a fair-baseline comparison is §P.1.

Table 8: The seven-member sign-vector family at $G = 8$ under Gaussian residuals. All members share the diagonal 30, the mean off-diagonal $51/7$ and the population ICC $17/70$; only the dominant contrast eigenvector differs. CIs are over 20 seeds.

member	closed form trCov	Monte Carlo	95% CI	$\widehat{\text{ICC}}$
$u^{(1)} = h$	167.4375	167.3647	[167.0590, 167.6703]	0.24295
$u^{(2)}$	3.5625	3.5646	[3.5612, 3.5679]	0.24267
$u^{(3)}$	3.5625	3.5626	[3.5589, 3.5662]	0.24318
$u^{(4)}$	3.5625	3.5644	[3.5595, 3.5693]	0.24283
$u^{(5)}$	3.5625	3.5627	[3.5591, 3.5662]	0.24281
$u^{(6)}$	3.5625	3.5630	[3.5595, 3.5664]	0.24250
$u^{(7)}$	3.5625	3.5637	[3.5604, 3.5670]	0.24271

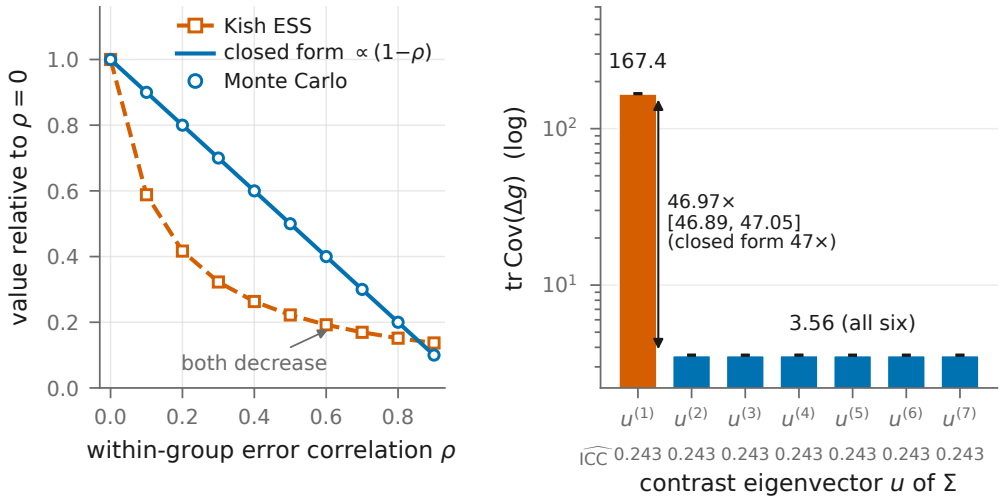

Figure 8: **Left:** at $G = 8$ with Gaussian residuals and fixed marginal variance, the Kish effective sample size $G/(1 + (G - 1)\rho)$ and the gradient-noise trace both *decrease* in ρ (indexed to their value at $\rho = 0$); Monte Carlo points lie on the closed form $\propto (1 - \rho)$. **Right:** the seven-member sign-vector family of §G.1 — identical marginal variances, mean off-diagonal and population ICC $17/70$ (estimated $\widehat{\mathrm{ICC}}$ below each bar) — with a spread in gradient-noise trace of $46.967\times$ (CI $[46.888, 47.048]$) by the corrected ratio-of-means estimand of App. P, against the closed-form $47\times$. Error bars: 95% CIs over 20 seeds.

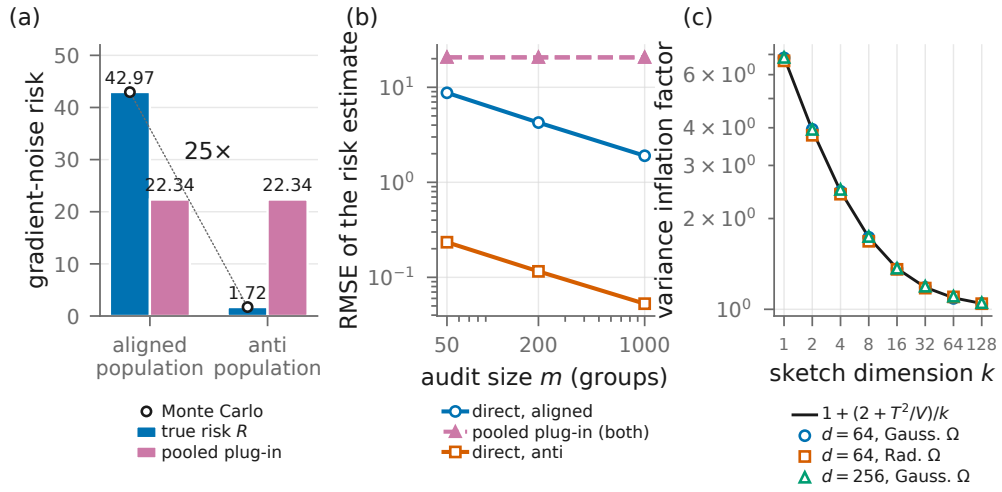

Figure 9: **(a)** two populations with identical pooled $(\bar{\Sigma}, \bar{K})$, per-prompt marginal variances and per-prompt ICC, whose true risks differ by $25\times$; the pooled plug-in returns the same number for both. **(b)** RMSE of $\widehat{T}$ versus the pooled plug-in against audit size m. **(c)** variance inflation of the sketched estimator against $1 + (2 + \mathcal{T}^2/V)/k$.

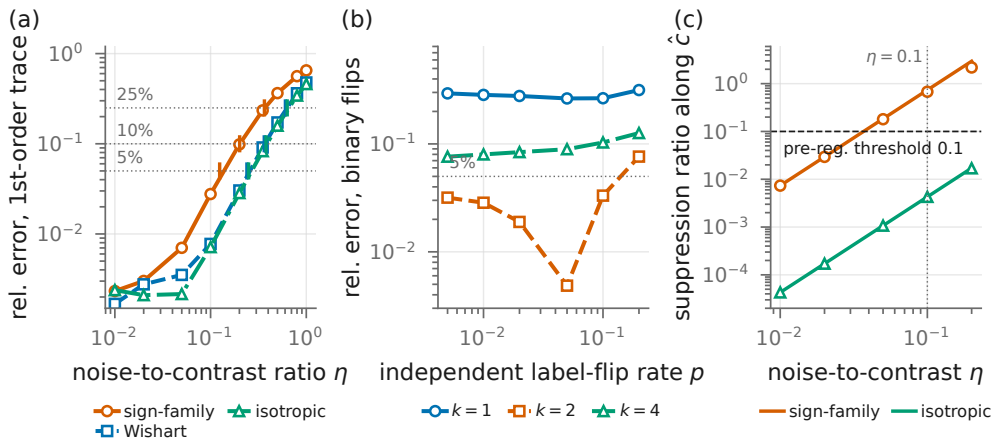

Figure 10: Validity regime of the first-order normalized-GRPO covariance (14). **(a)** relative error versus η for three Σ shapes, $5/10/25\%$ crossings ticked. **(b)** binary rewards with independent label flips: the error is flat in the flip rate at $n_+ = 1$ and $n_+ = 4$, and non-monotone at $n_+ = 2$, rather than shrinking with p. **(c)** suppression ratio (15) along $\hat{c}$: markers are measurements, lines the second-order closed form $\frac{1}{2}\mathrm{tr}(M^2)/(\|c\|^2 \hat{c}^\top \Sigma \hat{c})$, $M = QP\Sigma PQ$. The pre-specified threshold ($\mathrm{SR} < 0.1$ at $\eta = 0.1$) is met by near-isotropic Σ and *not* by the anisotropic sign family — a literal failure of the pre-specified protocol that the closed form explains (§K).

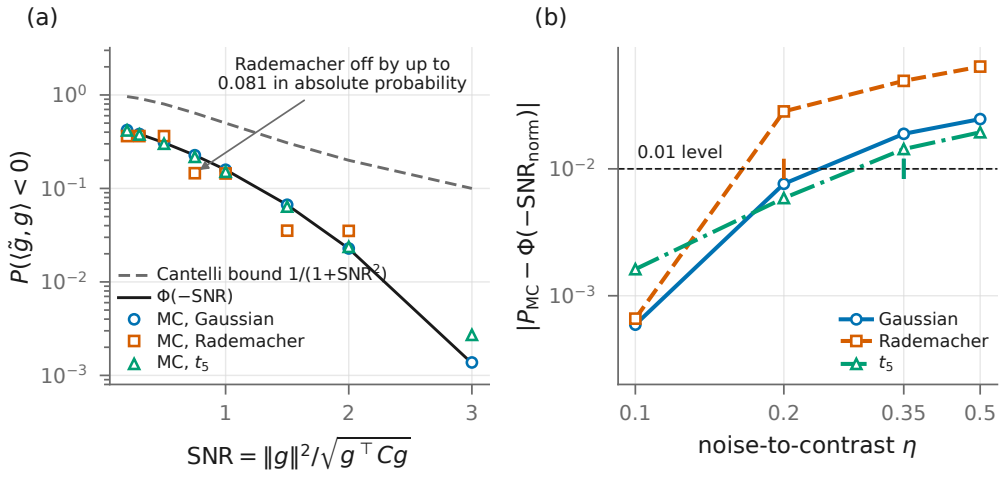

Figure 11: Sign flips. **(a)** empirical $\Pr(\langle \tilde{g}, g \rangle < 0)$ against the Gaussian prediction $\Phi(-\mathrm{SNR})$ (Proposition 20) and the distribution-free Cantelli bound. The covariance identity is distribution free; the flip formula is not — Rademacher residuals deviate by up to 0.0811 in absolute probability, t_5 by 0.0092, while the Gaussian law is recovered to 0.0010. **(b)** normalized-GRPO analogue: absolute error of the first-order prediction against η, with the 0.01 level and the per-law crossing marked ($\eta = 0.35$ for Gaussian and t_5, 0.20 for Rademacher).

Table 9: Noise-to-contrast ratio η at which the relative error of the first-order normalized-GRPO covariance (14) crosses 5, 10 and 25%. The regime is roughly twice as wide for near-isotropic Σ as for the strongly anisotropic sign family.

Σ shape, reward	law	η at 5%	at 10%	at 25%
sign family, binary r ($k = 4$)	Gaussian	0.124	0.201	0.365
	Rademacher	0.129	0.208	0.376
	t_5	0.116	0.180	0.347
sign family, continuous r	Gaussian	0.132	0.209	0.389
	Rademacher	0.137	0.216	0.399
	t_5	0.122	0.196	0.372
Wishart, binary r ($k = 4$)	Gaussian	0.240	0.364	0.605
Wishart, continuous r	Gaussian	0.226	0.348	0.599
isotropic, binary r ($k = 4$)	Gaussian	0.250	0.379	0.632
isotropic, continuous r	Gaussian	0.263	0.399	0.669

Table 10: Binary rewards with n_+ correct completions and independent label flips at rate p. The relative error of the first-order formula does not shrink with p (it is flat at $k = 1, 4$ and non-monotone at $k = 2$), and the suppression ratio (15) along $\hat{c}$ stays near 1. "degenerate" is the fraction of groups with $P\tilde{r} = 0$, which have an undefined z-scored advantage and are dropped.

	p	η	rel. error	SR	degenerate
$k = 1$	0.005	0.199	0.294	0.744	0.48%
	0.01	0.281	0.285	0.769	0.93%
	0.02	0.396	0.278	0.797	1.74%
	0.05	0.616	0.264	0.858	3.49%
	0.1	0.849	0.265	0.867	4.79%
	0.2	1.131	0.316	0.743	4.18%
$k = 2$	0.005	0.152	0.032	0.938	0.00%
	0.01	0.215	0.029	0.934	0.01%
	0.02	0.302	0.019	0.926	0.03%
	0.05	0.471	0.005	0.910	0.19%
	0.1	0.648	0.033	0.906	0.53%
	0.2	0.864	0.076	0.903	1.05%
$k = 4$	0.005	0.132	0.076	0.815	0.00%
	0.01	0.186	0.080	0.822	0.00%
	0.02	0.262	0.084	0.831	0.00%
	0.05	0.408	0.089	0.857	0.00%
	0.1	0.561	0.103	0.902	0.01%
	0.2	0.748	0.126	0.986	0.13%

Table 11: RMSE of the direct estimator $\widehat{T}$ and of the slot-fixed pooled $\hat{\Sigma}$ plug-in, at audit sizes $m = 50/200/1000$, 200 audits per seed. True risks are 42.96875 (aligned) and 1.71875 (anti). The plug-in's error is bias dominated and does not improve with m; both populations are adversarial for it by construction, and the slot-fixed comparator is not what a slot-exchangeable audit can form (App. G.2).

		$\widehat{T}$, $m =$			pooled plug-in, $m =$		
population	law	50	200	1000	50	200	1000
aligned	Gaussian	8.751	4.255	1.904	20.62	20.71	20.65
aligned	Rademacher	7.952	4.086	1.812	20.59	20.58	20.62
aligned	t_5	9.916	5.014	2.228	20.78	20.65	20.62
anti	Gaussian	0.234	0.116	0.053	20.72	20.60	20.62
anti	Rademacher	0.201	0.099	0.044	20.60	20.60	20.62
anti	t_5	0.299	0.151	0.073	20.80	20.63	20.64

O THE LEARNING EXPERIMENT

Does the geometry predict *learning*? A pre-specified secondary hypothesis asks whether the one-step geometry predicts learning in a numpy GRPO toy. One half is confirmed — the common-mode verifier, *worst* by an ESS audit, costs 0.0066 of final clean reward against 0.0856 for an orthogonal-contrast verifier — but is close to true by construction (Corollary 6). **The falsifiable half fails, and post hoc the design was powered against its own predicted effect:** the dose–response predicts an extra cost of 0.0386 ($[0.0272, 0.0524]$) against an observed 0.0034, and in 3 of 6 cells the predicted interval's *lower* end exceeds the minimum detectable difference — a *powered null* (App. P.3).

The toy is a J-way answer $\times$ format completion space with a linear-softmax policy and a rule verifier whose errors are format dependent, run for 300 GRPO steps with 20 seeds, two operators (centered, z-score) and three noise scales $(0.25, 0.50, 1.00)$. The three noisy verifiers are matched in marginal error variance, and the two contrast verifiers are additionally matched in ANOVA ICC; the design-validity checks pass in all 12 cells. Figure 12 and Tables 12–13 show the outcome under both operators.

Common-mode noise, which an ICC/ESS audit ranks as by far the worst verifier ($\widehat{\mathrm{ICC}} \approx 0.950$), tracks the clean verifier under the z-score operator, while contrast noise at $\widehat{\mathrm{ICC}} \approx 0.193$ and identical marginal variance costs an order of magnitude more. What the experiment does *not* support is the finer ordering between aligned and orthogonal contrast noise: the pre-specified criterion fails in all 6 cells (the paired difference is negative in 4 and positive in 2; no CI excludes 0), and App. P.3 shows that the paper's own dose–response predicts a gap several times the minimum detectable difference, so this is a null rather than an absence of power. The second criterion (aligned worse than common mode) fails in the 3 *centered* cells, and Table 13 shows why: under the centered operator every condition, including the clean verifier, ends between 0.950 and 0.953 clean reward, against a minimum detectable difference of 0.0015–0.0031 — at these noise scales the centered update is simply not damaged enough by any of the verifiers for the geometry to matter. The per-direction scalarisation $w^\top \mathrm{Cov}(\Delta g) w$ tabulated alongside the trace orders aligned above orthogonal by 1.39–1.78 over the centered cells, and the dimensionless $g^\top \mathrm{Cov}(\Delta g) g / \|g\|^4$ of Table P.4 by 1.39–1.82 over all cells; neither reverses the prediction here. A post-hoc re-analysis with identical seeds (App. P.4, Table P.4) adds the per-condition verifier share of total update variance and the predicted sign-flip probability, and computes the dose–response argument: in 3 of the 6 cells the design was powered against the implied effect and did not see it.

Table 12: The learning toy under the *z-score* operator: realized marginal variance and ANOVA ICC of the verifier residuals, the predicted one-step noise trace, and the learning outcome after 300 steps.

noise scale	verifier	marg. var.	$\widehat{\mathrm{ICC}}$	$G^{-2}\mathrm{tr}(P\Sigma PK)$	final clean reward	steps to 0.8
$\sigma_\varepsilon = 0.25$	clean verifier	0.000	0.000	0.0000	0.9621	98.5
	common mode	0.062	0.950	0.0005	0.9618	99.5
	orthogonal contrast	0.063	0.195	0.0077	0.9557	109.5
	aligned contrast	0.063	0.195	0.0143	0.9562	108.0
$\sigma_\varepsilon = 0.50$	clean verifier	0.000	0.000	0.0000	0.9621	98.5
	common mode	0.248	0.950	0.0018	0.9603	100.2
	orthogonal contrast	0.250	0.191	0.0332	0.9380	135.3
	aligned contrast	0.251	0.195	0.0595	0.9379	138.5
$\sigma_\varepsilon = 1.00$	clean verifier	0.000	0.000	0.0000	0.9621	98.5
	common mode	1.007	0.950	0.0072	0.9554	108.8
	orthogonal contrast	1.001	0.193	0.1683	0.8765	213.2
	aligned contrast	1.002	0.193	0.2660	0.8731	213.9

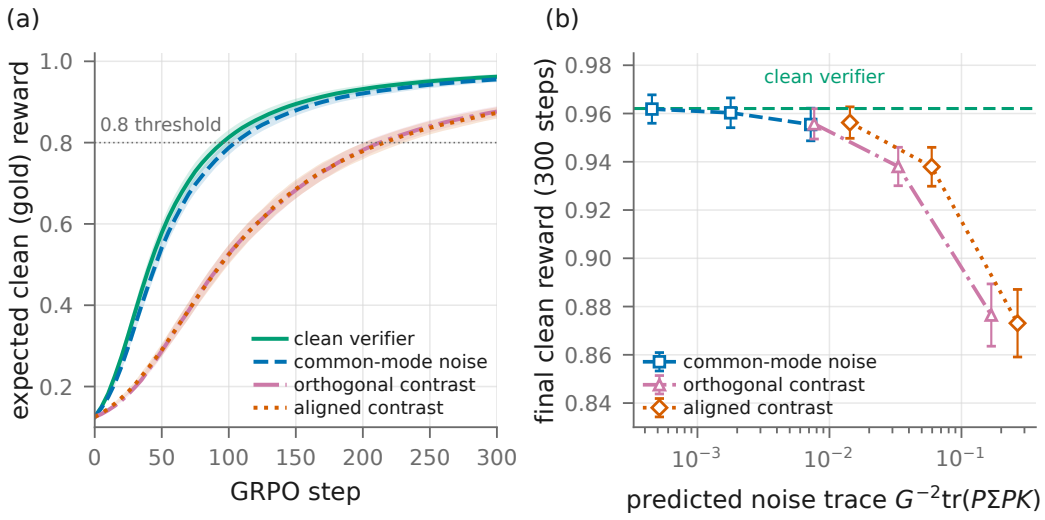

Figure 12: **(a)** clean (gold) reward during training under the z-score operator at the largest noise scale, mean ± s.e. over 20 seeds: common-mode noise (ICC 0.950) tracks the clean verifier; contrast noise (ICC ≈ 0.193) at the same marginal variance costs about 0.09 of final reward. **(b)** final clean reward against the predicted one-step noise trace $G^{-2}\mathrm{tr}(P\Sigma PK)$, error bars ±1 s.e. over the 20 seeds (not 95% intervals; those are in Table P.4): the prediction orders the conditions across more than a decade, but the aligned and orthogonal conditions are separated horizontally without a corresponding vertical separation.

Table 13: The learning toy under the *centered* operator — the case in which Theorem 5 is exact. "along w" is $w^\top \mathrm{Cov}(\Delta g) w$ with w the exact normalised clean-objective gradient; it is *not* the dimensionless $g^\top \mathrm{Cov}(\Delta g) g / \|g\|^4$ of Table P.4, which uses the within-prompt sample-mean clean update (App. P.3). Every condition ends within 0.950–0.953, so nothing separates them; this is where the second learning criterion fails.

noise scale	verifier	marg. var.	$\widehat{\mathrm{ICC}}$	trace	along w	final clean reward	95% CI
$\sigma_\varepsilon = 0.25$	clean verifier	0.000	0.000	0.0000	0.00000	0.9523	[0.9427, 0.9618]
	common mode	0.063	0.950	0.0006	0.00001	0.9526	[0.9431, 0.9620]
	orthogonal contrast	0.063	0.192	0.0091	0.00016	0.9512	[0.9424, 0.9600]
	aligned contrast	0.062	0.192	0.0169	0.00023	0.9520	[0.9426, 0.9614]
$\sigma_\varepsilon = 0.50$	clean verifier	0.000	0.000	0.0000	0.00000	0.9523	[0.9427, 0.9618]
	common mode	0.251	0.950	0.0023	0.00004	0.9524	[0.9436, 0.9613]
	orthogonal contrast	0.249	0.189	0.0313	0.00055	0.9522	[0.9432, 0.9613]
	aligned contrast	0.250	0.195	0.0494	0.00086	0.9520	[0.9422, 0.9618]
$\sigma_\varepsilon = 1.00$	clean verifier	0.000	0.000	0.0000	0.00000	0.9523	[0.9427, 0.9618]
	common mode	1.002	0.950	0.0080	0.00014	0.9520	[0.9425, 0.9614]
	orthogonal contrast	1.000	0.196	0.0822	0.00221	0.9508	[0.9413, 0.9603]
	aligned contrast	0.998	0.196	0.1154	0.00343	0.9499	[0.9393, 0.9605]

P EXPLORATORY STUDIES

The complete record of every study in this section (frozen plans and their hashes, verification checks, per-cell tables, figures and run logs) is in the supplementary material under `code/experiments/`; here we keep one pointer per study. Every number quoted in the main text is generated from those files.

P.1 A LEARNED-POLICY CHECK: A TOY TRAINED BY GRPO

See the corresponding `README.md` and results in the supplementary material.

P.2 DESIGN OF THE LEARNED-POLICY AUDIT

See the corresponding `README.md` and results in the supplementary material.

P.3 RE-ANALYSES OF THE LEARNED-POLICY AUDIT

See the corresponding `README.md` and results in the supplementary material.

P.4 CORRECTED AND FAIRER REDESIGNS

See the corresponding `README.md` and results in the supplementary material.

P.5 SENSITIVITY OF THE CLOSED-FORM RATIO

See the corresponding `README.md` and results in the supplementary material.

P.6 THE PRETRAINED-LLM AUDIT, AND THE POST-HOC RE-ANALYSES OF IT

See the corresponding `README.md` and results in the supplementary material.

P.7 THE TRAINING-TIME AUDIT

See the corresponding `README.md` and results in the supplementary material.

P.8 THE NOISE ENVELOPE OF THE STORED LLM SCORE GEOMETRY

See the corresponding `README.md` and results in the supplementary material.

P.9 TRUNCATION, THE LENGTH-NORMALISED OPERATOR, AND THE SKETCH

See the corresponding `README.md` and results in the supplementary material.

P.10 THE MOMENT DECOMPOSITION AND THE NEAR-ORTHOGONAL REGIME ON THE STORED GRAMS (E17)

See the corresponding `README.md` and results in the supplementary material.

P.11 THE MULTI-SEED AUDIT WITH A LARGER TOKEN CAP (E18)

See the corresponding `README.md` and results in the supplementary material.

Q REPRODUCIBILITY DETAILS

Pipeline. `theory/constructions.py` builds every construction and asserts every closed-form quantity in the paper (203 assertions). Each experiment `experiments/E0x/run.py`

writes a per-seed record file and a `summary.json` containing its pre-specified criterion checks; `analysis/make_criteria.py` assembles `analysis/criteria.json` (the 22 criteria) and `analysis/paper_numbers.json` (a flat key → value table, each entry carrying its source file and JSON path); `paper/make_numbers.py` turns that table into `paper/numbers.tex` and the table bodies in `paper/tables/`; and `paper/figs/fig*.py` generate the figures reading only the experiment and analysis JSON files. No experimental number in the LaTeX source is typed by hand.

Commands.

```
python3 theory/constructions.py     # closed-form assertions
for e in E01 E02 E03 E04 E05 E06 E07 E08
  do python3 experiments/$e/run.py; done
python3 experiments/E09/run.py all      # exploratory
for f in experiments/E10/run_e10?.py; do python3 $f; done
python3 experiments/E11/run_e11_audit.py   # exploratory
python3 experiments/E11/run_e11_reanalysis.py
python3 experiments/E11/aggregate_e11.py
python3 analysis/make_criteria.py   # criteria + paper_numbers
python3 analysis/make_post_numbers.py  # exploratory keys
python3 paper/make_numbers.py       # numbers.tex, tables/*.tex
for f in paper/figs/fig*.py; do python3 $f; done
cd paper && latexmk -pdf main.tex

python3 theory/near_orthogonal/check_near_orthogonal.py
# the three GPU audits, each an unattended container run
bash experiments/E12/docker/run.sh        # e12-llm-audit:1.0
bash experiments/E14/run.sh --max-minutes 115   # e14-grpo:1.0
bash experiments/E18/run.sh --max-minutes 225   # e14-grpo:1.0
cd experiments/E18 && python3 analyze.py --results results
# the CPU re-analyses of what they stored (no GPU)
python3 experiments/E13_posthoc/run_posthoc.py  # ar-paper:1.0
python3 experiments/E15/run_e15.py              # ar-paper:1.0
python3 experiments/E16/run_e16.py              # ar-paper:1.0
python3 experiments/E17/run_e17.py              # ar-paper:1.0
```

Image definitions: `ar-paper:1.0` is built from `docker/Dockerfile` (shipped in the supplement) and carries Python, numpy/scipy and TeX Live; `e12-llm-audit:1.0` is built from `experiments/E12/docker/Dockerfile`; `e14-grpo:1.0` is that same image with its `ENTRYPOINT` cleared. Every command above was run inside one of these three containers; nothing was installed on the host.

Cost and settings. Seeds 0–19 for the pre-specified suite; 0–9 for the learned-policy study and 10–19 for its confirmatory rerun; 200,000 Monte Carlo draws per seed and cell (the audit cells use 200 independent audits per seed instead). All runs are CPU-only and cost no model calls: E01–E08 and E10 are numpy/JAX, the learned-policy study is JAX end to end (`jax` 0.10.2, `numpy` 2.4.4, `scipy` 1.17.1, Python 3.11). Wall-clock time is 324 s for the learned-policy study, 156 s for the corrected redesigns and 816 s for the re-analyses and the confirmatory rerun, on two cores; the whole pipeline is under an hour. The exceptions are the three LLM audits. E18 ran in `e14-grpo:1.0` in several resumable chunks on a shared card (its GPU ledger is in the supplement), and E17 recomputes from stored records on CPU. E12 ran once in `e12-llm-audit:1.0` for 25.1 minutes at 7.2 GB peak GPU memory (seed 0, $B = 2{,}000$ bootstrap resamples over prompts, seeded in the shipped code), and E14 once in `e14-grpo:1.0` for 27.6 minutes (seed 0, paired bootstrap $B = 2{,}000$); E13 and E15 recompute from their saved records on CPU in `ar-paper:1.0`, E15 with $B = 2{,}000$ and the three seeds its plan fixes. Noise laws are standardized to the target Σ by $\varepsilon = \Sigma^{1/2}\zeta$ with $\mathbb{E}\zeta = 0$, $\mathrm{Cov}\zeta = I$, except the binary-flip model where the unstandardized flip process is the object of study.

R ADDITIONAL RELATED WORK

§7 keeps the concurrent measurement, the imported scalar tools and the statistical identification of $\widehat{T}$; the rest of the literature review is here, with nothing dropped.

The concurrent audit, and the reversal on its own number. Taking Xin (2026)'s $\hat{\rho} = 0.530$ at face value *under an exchangeability assumption §G.1 shows need not hold*, Corollary 7 would imply a trace *reduced* by 0.470. That is an illustration of the reversal on a published number, **not** a measurement of their setting: we did not observe their groups, their verifier's errors need not be exchangeable, and their paper makes no claim about gradient noise.

The CR0 correspondence, explicitly. The cluster-robust (CR0) sandwich *meat* of the GEE tradition (Liang & Zeger, 1986) is $\sum_j X_j^\top \hat{e}_j \hat{e}_j^\top X_j$ for cluster design matrices X_j and residuals $\hat{e}_j$. Putting $X_j = S_j P$ and $\hat{e}_j = e_j$, the audit residual of §B.4, gives $\sum_j \|S_j P e_j\|^2$, i.e. $mG^2\widehat{T}$ exactly: the estimator is that meat, not a new statistic. Likewise, the correction §5 finds operative is Kish's unequal-weighting effect $1 + \mathrm{CV}^2$ (Kish, 1965) with the score norms $\|s_i\|$ as the weights — with the one difference that it acts on *contrasts*, so the common-mode part is annihilated first (Corollary 6) before any weighting matters.

Operator variants. Zhou et al. (2026) show the GRPO gradient is a U-statistic under *clean* rewards, where no residual covariance arises. Dr. GRPO (Liu et al., 2025) removes the divisor whose linearization is our Q, DAPO (Yu et al., 2025) changes the operator further, RLOO (Ahmadian et al., 2024) rescales P, and the baseline literature (Williams, 1992; Weaver & Tao, 2001; Huang et al., 2025b) gives variance reduction under *independent* reward noise, which Corollary 7 makes direction dependent.

Verifier-error channels, and what injecting noise measures. Cai et al. (2025) and Wang et al. (2018) formalise conditionally independent error channels and the confusion-matrix treatment; empirically Plesner et al. (2026) inject i.i.d. noise up to 15% with little degradation — noise with no preferred contrast direction — while Zhu & Kang (2026) find 8–12% degradation from wrong labels and Shao et al. (2025) show random rewards move benchmarks, so an aggregate error rate predicts outcomes poorly; Huang et al. (2025a), Helff et al. (2026) and Kim et al. (2025) document the mechanisms making residuals correlated and policy-aligned, and Gao et al. (2022) studies the component Proposition 19 isolates.

Design effects in LLM evaluation. Kohli (2026) and Noonan (2026) bring the ICC and the design effect to LLM evaluation, again for *means* rather than for a score-weighted contrast, which is the distinction §4 turns on.