WHEN CAN COUNTERFACTUAL AUDITS CERTIFY A FINE-TUNING GAIN?

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual post-training expands supervision through generated edits, but neither more edits nor greater predictive consistency establishes a gain over factual training. Certifying that gain requires an independent comparison with enough power to detect it under the declared nuisance shifts. We study this evidence requirement through paired losses: common model responses can cancel, while additional edits cannot remove uncertainty across task anchors. Shared-anchor certificates and finite-sample power bounds distinguish these two sources of cost. An independent pilot–plan–certify procedure then selects the audit size, sibling count, estimator, and confidence allocation from a frozen menu, accounting for uncertainty in the planning moments themselves. Its guarantees separate false certification, pilot reliability, and final power conditional on a feasible plan. Controlled experiments isolate the savings from pairing and anchor reuse; across three planning laws, 109 of 144 pilots return feasible plans, each successful in all 256 fresh replications, but their budgets remain conservative. A prospective Qwen/ReCoRD audit illustrates covered-shift certification with a different, finite-catalog design. The practical lesson is to budget for uncertainty in the audit plan, not just in the final result: a valid certificate and an economical, adequately powered audit are distinct achievements.

1 INTRODUCTION

Large language models have advanced rapidly in understanding, generation, and reasoning (Brown et al., 2020; Qwen Team, 2024). Fine-tuning translates these capabilities into task behavior using costly demonstrations and human feedback (Wei et al., 2022; Ouyang et al., 2022). The effectiveness of small, carefully curated datasets (Zhou et al., 2023) motivates extending scarce, high-quality supervision without proportional annotation costs. Counterfactual augmentation uses targeted edits for sentiment and inference (Kaushik et al., 2020; 2021), gender-bias mitigation (Zmigrod et al., 2019), and retrieval-guided question answering (Paranjape et al., 2022). CORE and DISCO use large language models to generate useful edits (Dixit et al., 2022; Chen et al., 2023), while causal augmentation improves caregiver-invariant clinical prediction (Feder et al., 2023). Prompting studies also identify task-dependent generation limitations (Li et al., 2024). We study *counterfactual post-training* (CFPT) with edits to nuisance attributes intended to preserve the task and label, rather than label-changing edits. An exploratory binary MultiNLI study (Figure 1) shows that more generated data need not improve performance. Across 16 generation and training pools sharing one evaluation split, mean Brier risk rises beyond its early minimum, while classification error is nonmonotone. Within the $n > 500$ regime, predictive KL decreases even as mean Brier risk worsens. Audited label-disagreement rates do not track risk monotonically either. These descriptive patterns motivate a direct comparison of performance: training volume, predictive agreement, and marginal label quality cannot substitute for evidence of a gain.

Certification also faces a sampling problem: sibling edits are not independent task anchors, and controlling false claims need not provide enough power to detect a gain. We ask: *when can counterfactual comparisons certify a gain under nuisance shift, and how much independent evidence should the audit collect?* We compare frozen models on matched anchors and edits, allowing common task difficulty to cancel. Classical robust mean–variance geometry (Duchi & Namkoong, 2019; 2021) controls the gain under covered nuisance shifts through the sensitivity of the *loss difference*. Shared-anchor confidence bounds distinguish the anchor uncertainty floor from within-

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.

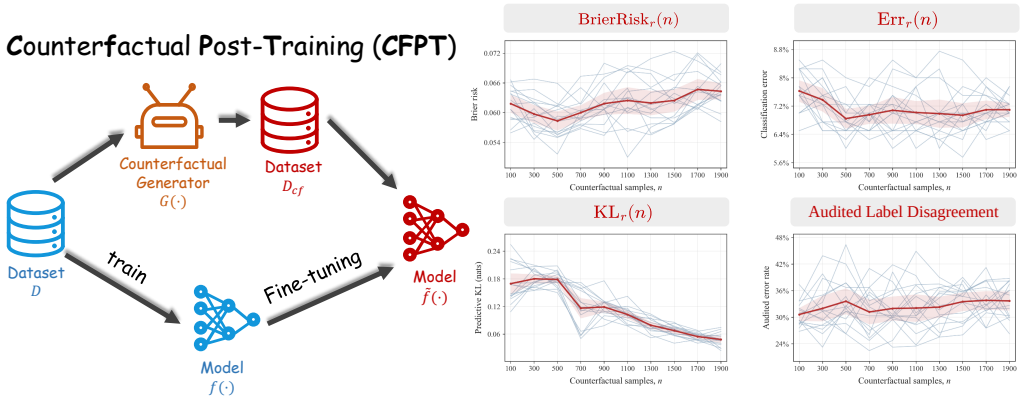


Figure 1: **More counterfactual data need not improve performance. Left:** the original counterfactual generation and fine-tuning workflow. **Right:** Brier risk, classification error, predictive KL, and label disagreement for 16 binary MultiNLI generation/training pools sharing a 500/400/400 train/development/exploratory-test split; test reuses an earlier development set. Each model uses 500 factual and $n = 100, 300, \ldots, 1900$ generated examples. Pale curves are pools, red curves their means, and bands pointwise 95% t-intervals conditional on this shared split. The KL discussion in the text concerns the $n > 500$ segment. Metric definitions, model, training, and sampling details are in Appendix H.1. Label-disagreement rates describe cached judgments, not human semantic certification.

anchor noise. Additional siblings reduce only the latter; they cannot replace independent anchors. Finite-sample power bounds then connect uncertainty to audit cost. Our *pilot–plan–certify* procedure uses independent pilot data to select a budget, sibling count, estimator, and confidence allocation from a frozen menu before collecting a fresh final audit. Simultaneous pilot bounds separate safety, planning reliability, and conditional power for returned plans. The insight is simple: *budget for uncertainty in the budget estimate, not just the final result*. These guarantees concern frozen models and declared shifts, not end-to-end optimization. On three controlled, known-support laws, all 109 returned plans certify in all 256 final replications, but 35 of 144 pilots return infeasible and budgets remain conservative. Plug-in planning can miss target power. This prospective study evaluates planning; the descriptive training curves in Figure 1 motivate the audit question.

Contributions. We make three main contributions.

- **Paired certification.** Shared-anchor bounds and variance decompositions quantify the independent evidence needed to certify gains.
- **Pilot-based planning.** Power conditions support valid menu selection, with a local information benchmark that includes pilot costs.
- **Prospective cost evaluation.** Frozen-budget experiments reveal reliable certification on returned plans, incomplete feasibility, and conservative costs.

2 Related Work

Generated supervision and counterfactual augmentation. Instruction tuning and preference learning rely on curated supervision (Wei et al., 2022; Ouyang et al., 2022; Zhou et al., 2023). Counterfactual NLP data target lexical shortcuts, gender associations, and question-answering failures (Kaushik et al., 2020; 2021; Zmigrod et al., 2019; Paranjape et al., 2022). LLM-based generation and filtering extend these interventions (Dixit et al., 2022; Chen et al., 2023; Feder et al., 2023; Li et al., 2024). Causal representations supply a broader language for separating task content from interventions (Johansson et al., 2016; Shalit et al., 2017; Khemakhem et al., 2020; Pawlowski

et al., 2020; Schölkopf et al., 2021). These approaches motivate CFPT; our question concerns the independent evidence needed to compare a trained CFPT model with its factual comparator.

Invariance, transfer, and robust risk. Domain-adaptation bounds relate risk changes to distributional discrepancy (Ben-David et al., 2010); domain-generalization methods seek invariance across training environments (Muandet et al., 2013). Invariant prediction, environment perturbation, and robust objectives seek stable responses (Peters et al., 2016; Shankar et al., 2018; Arjovsky et al., 2019; Ahuja et al., 2020; Krueger et al., 2021; Sagawa et al., 2020); benchmark studies emphasize the importance of evaluation design (Gulrajani & Lopez-Paz, 2021). The mean–variance geometry of Pearson-χ^2 robustness is established (Duchi & Namkoong, 2019; 2021). We apply it to the *paired loss difference* under conditional nuisance shifts, so shared model responses cancel before uncertainty is bounded. The contribution is the resulting comparison and its finite-information cost, rather than a new divergence duality. The appendix also relates predictive consistency to covered-shift risk using the pairwise-KL decomposition of Malinin & Gales (2021).

Dependence and finite-sample inference. Group-structured augmentation can reduce variance (Chen et al., 2020), while information-theoretic generalization bounds relate error to dependence between training data and a learned model (Russo & Zou, 2016; Xu & Raginsky, 2017; Bu et al., 2020; Steinke & Zakynthinou, 2020). Our main results condition on the trained models and treat complete anchor–sibling clusters as independent observations. Empirical Bernstein bounds (Maurer & Pontil, 2009) and without-replacement concentration (Bardenet & Maillard, 2015) support different audit designs. Reusing anchors for several statistics does not invalidate a simultaneous confidence statement; it does require retaining the covariance relevant to its power.

Safety, planning, and corruption. Risk-controlling calibration (Bates et al., 2021; Angelopoulos et al., 2025) and safe policy improvement (Thomas et al., 2015; Petrik et al., 2016) formalize risk limits or baseline-relative gains. Our question is the audit budget needed to detect a covered-shift gain, not safety alone. Internal pilots update nuisance estimates (Wittes & Brittain, 1990); external-pilot design trades precision against pilot-plus-trial cost (Whitehead et al., 2016). We require a simultaneous finite-sample power witness for bounded, clustered robust endpoints, with a separate final audit preserving safety conditional on every pilot. Splitting alone or an estimated asymptotic variance does not supply this planning guarantee. Finite-sample normal approximation (Shevtsova, 2011) gives a covariance-aware witness; a sequential change-of-measure argument (Kaufmann et al., 2016) charges all anchors, including the pilot. Semantic validity is a separate requirement: label-noise and loss-correction theory (Natarajan et al., 2013; Patrini et al., 2016; Menon et al., 2018) does not certify automatically screened edits. Our audits state their label, coverage, and sampling conditions explicitly.

3 Paired Certification Interface

Condition on J trained CFPT models and their factual comparator. Each fresh task anchor has semantic state C and valid label Y; changing nuisance S preserves the conditional label law $p^*(Y \mid C)$. All losses lie in $[0, B]$. The reference nuisance law $\nu(\cdot \mid C)$ and task marginal μ are fixed. Define the label-averaged paired response and its moments by

$$g_h(c,s) = \mathbb{E}[\ell(w_{C,h}; c, s, Y) - \ell(w_F; c, s, Y) \mid C = c],$$
$$\bar{g}_h(c) = \mathbb{E}_\nu[g_h(c, S) \mid c], \quad \Delta_{\nu,h} = \mathbb{E}_\mu \bar{g}_h(C), \quad V_h = \mathbb{E}_\mu \operatorname{Var}_\nu(g_h(C, S) \mid C).$$

Negative risk difference means CFPT improves on factual training. A covered law changes only $S \mid C$, has density $r = d\pi/d\nu \geq 0$, and satisfies $\mathbb{E}_\nu[r \mid C] = 1$ and $\mathbb{E}_{\mu\nu}(r-1)^2 \leq \rho$. Write this class as $\mathcal{E}_\rho(\nu)$. The centered change-of-measure identity of Theorem B.3 yields

$$\Psi_h(\rho) := \sup_{\pi \in \mathcal{E}_\rho(\nu)} \Delta_{\pi,h} \leq \Delta_{\nu,h} + \sqrt{\rho V_h}.$$

The envelope is locally exact when its maximizing density stays nonnegative (Proposition B.3). The separate-model envelope instead adds $\sqrt{\rho}(\sqrt{V_C} + \sqrt{V_F})$, at least as large by the triangle inequality. Thus two models may each respond strongly to an edit while their *difference* barely changes. The audit should pay for the latter when deciding which model is better.

For example, task difficulty may raise both models' losses on the same anchors. A separate analysis pays for that variation twice; the paired response removes its common part. Cancellation is different from invariance: the individual predictors may remain sensitive even when their relative performance is stable. Conversely, a small reference gain can be overturned when the difference itself reacts strongly to the nuisance. Centering within each task state distinguishes those effects from a change in which tasks are tested. The common task marginal and covered nuisance class are therefore substantive parts of the comparison, not optional regularity conditions.

The auditor freezes candidates, loss, ν, ρ, δ, sampling design, and selection rule before collecting audit anchors (either shared or split across roles). A policy is certified only when a valid simultaneous upper bound is negative; otherwise the rule returns factual training. The power analysis assumes valid deterministic labels and an exact reference design, and states separately how mismatch or contamination consumes margin. Full training-policy definitions, assumptions, and supporting transfer and information results appear in Appendices A–B; the proof guide links each result to its complete argument.

3.1 A CERTIFICATE FROM SHARED ANCHORS

For deterministic valid labels and even K IID nuisance siblings, define $M_i = K^{-1}\sum_k Z_{ik}$ and $Q_i = (K/2)^{-1}\sum_{k=1}^{K/2}(Z_{i,2k-1} - Z_{i,2k})^2/2$. Then $\mathbb{E}M = \Delta$, $\mathbb{E}Q = V$, $M \in [-B, B]$, and $Q \in [0, 2B^2]$. With exact nuisance enumeration, instead use $M = \mathbb{E}_\nu Z$ and $Q = \mathrm{Var}_\nu Z \le B^2$. For scalar inputs, $D_{\mathrm{KL}}(p\|q)$ means Bernoulli KL. The bounded operator is

$$H_c(\bar{Q}; N, e) = c\sup\{u \in [\bar{Q}/c, 1] : D_{\mathrm{KL}}(\bar{Q}/c\|u) \le \log(1/e)/N\}.$$

This is a confidence inversion, not the nuisance shift class.

Proposition 1 (Shared-anchor moment certificate)**.** *Freeze J candidates, ρ, and the sampling law. For $N \ge 2$, set $e \le \delta/(2J)$ and*

$$U_{\mathrm{sh},j} = \bar{M}_j + r_j + \sqrt{\rho V_{U,j}}, \qquad V_{U,j} = \min\{B^2, H_{c_Q}(\bar{Q}_j; N, e)\}. \tag{1}$$

For IID anchors with sampled siblings, use $c_Q = 2B^2$ and $r_j = \sqrt{2s^2_{M,j}\log(2/e)/N} + 14B\log(2/e)/[3(N-1)]$, where $s^2_{M,j}$ is unbiased. For a simple random sample without replacement from a fixed catalog with enumerated nuisances, use $c_Q = B^2$ and $r_j = 2B\sqrt{\log(1/e)/(2N)}$. Then $\Pr\{\forall j : \Psi_j \le U_{\mathrm{sh},j}\} \ge 1 - \delta$. Selecting any negative endpoint, otherwise returning factual training, has family-wise false acceptance at most δ.

Both statistics use the same anchors. Empirical Bernstein or the without-replacement MGF comparison controls the mean; bounded KL inversion controls sensitivity. A union bound over $2J$ marginal events suffices: estimator independence is unnecessary. Splitting the same budget between mean and sensitivity estimation is valid but reduces the observations available to each statistic.

For comparison, a separate-model certificate adds the two mean radii and two sensitivity radii. Proposition D.1 gives the matched confidence allocation. Crossing paired/separate responses with shared/split anchors separates two savings that are often conflated: covariance cancellation and reuse of observations.

An exact covariance-only contrast. In the balanced primary controlled law, $Q_C = Q_F = Q_\Delta/4$. Homogeneity of H_c makes the summed separate sensitivity radius equal to the paired radius, for shared or split roles. With the same e and mean-block size n, the EB range terms also match, leaving exactly

$$U_{\mathrm{sep}} - U_{\mathrm{pair}} = \sqrt{2\log(2/e)/n}\,(s_C + s_F - s_\Delta), \qquad s^2_\Delta = s^2_C + s^2_F - 2s_{CF}.$$

Here s_{CF} is the sample covariance. Thus the primary factorial's pairing contrast isolates variance aggregation; general laws need not share this cancellation.

3.2 A JOINT ENDPOINT WHEN NUISANCES ARE SAMPLED

Pairing is available to a generic robust estimator as well. For any nonnegative sensitivity bound V_U,

$$\widehat{\Delta} + r + \sqrt{\rho V_U} = \inf_{\eta > 0}\{\widehat{\Delta} + r + \eta\rho + V_U/(4\eta)\}. \tag{2}$$

The explicit moment endpoint estimates the two terms separately. Alternatively, combining them *within each anchor* retains their covariance during uncertainty estimation.

Proposition 2 (A generic certificate from finite siblings)**.** *Under the deterministic-label protocol, let N independent anchors each have an even number K of IID nuisance siblings, with paired losses $Z_{ij} \in [-B, B]$. Use the preceding M_i, Q_i, and freeze a positive grid $\mathcal{H}$ of size H before outcomes. Define*

$$Y_{i,\eta} = M_i + Q_i/(4\eta), \qquad R_\eta = B + \max\{B, B^2/(2\eta)\}.$$

Writing s_η^2 for the unbiased sample variance of these N anchor responses and $t = \log(2JH/\delta)$, the endpoint

$$U_{\mathrm{GD}} = \min_{\eta \in \mathcal{H}} \left\{ \eta\rho + \bar{Y}_\eta + \sqrt{2s_\eta^2 t/N} + \frac{7R_\eta t}{3(N-1)} \right\} \tag{3}$$

simultaneously bounds the robust loss increase of all J candidates with probability at least $1-\delta$ for $N \geq 2$. It requires no enumeration, known support size, or separate calibration block.

Here $\mathbb{E}Q_i = V$ includes inner nuisance uncertainty; substituting the naive empirical within-anchor variance would underestimate V by $(K-1)/K$. The empirical Bernstein bound (Maurer & Pontil, 2009) is applied to *whole anchor responses*, including their sampled variance estimates. No optimization over an empirical nuisance support is treated as exact. Equation (3) relaxes the density-positivity constraint; it bounds $\Delta + \sqrt{\rho V}$, and equals the robust target in the local regime, apart from the frozen grid's nonnegative approximation error.

Proposition 3 (Finite-sibling variance anatomy)**.** *Put $\mu(C) = \mathbb{E}[Z \mid C]$, $v(C) = \mathrm{Var}(Z \mid C)$, and let $\kappa_j(C)$ be its centered conditional jth moment. With $V = \mathbb{E}v$,*

$$\tau_K^2 = \mathrm{Var}\,\mu + V/K, \qquad \Omega_K = \mathrm{Var}\,Q = \mathrm{Var}_C\, v + \frac{\mathbb{E}\kappa_4 + \mathbb{E}v^2}{K}, \tag{4}$$

$$\mathrm{Cov}(M, Q) = \mathrm{Cov}_C(\mu, v) + \mathbb{E}\kappa_3/K, \tag{5}$$

$$\sigma_\eta^2 := \mathrm{Var}\,Y_\eta = \tau_K^2 + \frac{\Omega_K}{16\eta^2} + \frac{\mathrm{Cov}(M, Q)}{2\eta}. \tag{6}$$

Moreover $\mathrm{Var}\,\mu = \sigma_{A,C}^2 + \sigma_{A,F}^2 - 2\,\mathrm{Cov}_C(\mu_C, \mu_F)$.

When $V, \rho > 0$, at $\eta_* = \sqrt{V}/(2\sqrt{\rho})$, the nuisance-estimation contribution is $\rho\Omega_K/(4V)$, with an additional covariance term $\sqrt{\rho/V}\,\mathrm{Cov}(M, Q)$. Increasing K reduces sibling noise, but leaves both the anchor floor and $\mathrm{Var}_C\, v$. Neither an anchor covariance saving nor a nuisance-variance saving can be inferred from K alone.

The sign of $\mathrm{Cov}(M, Q)$ can therefore change which endpoint is more economical. Appendix D.4 supplies the proofs and explicit budget comparisons, including the zero-sensitivity case.

4 Predicting Certification Cost

For each candidate (indices suppressed), let $\Psi = -\gamma < 0$, $s_P = \Delta + \sqrt{\rho V} - \Psi$, and $s_S = \Delta + \sqrt{\rho}(\sqrt{V_C} + \sqrt{V_F}) - \Psi$. These population slacks cannot be reduced by more observations. Let $\mathcal{B}_R(\sigma, N; e, \beta)$ bound the upper fluctuation of an empirical-Bernstein mean endpoint, and let $\mathcal{A}_c(V, \Omega, N; e, \beta)$ bound the excess square-root sensitivity radius. Their explicit finite-sample forms, including range corrections and $V = 0$, are given in Appendix D.6. Both scale as $N^{-1/2}$ to leading order; the mean term retains its variance, whereas the sensitivity term also pays for estimating V.

Theorem 1 (Shared-anchor finite-budget prediction)**.** *Assume bounded deterministic losses, IID anchors and even IID siblings, frozen candidates and shift class, and the matched event ledger above. Put $\tau_\Delta^2 = \mathrm{Var}_C\,\mathbb{E}[Z \mid C] + V/K$, $\Omega = \mathrm{Var}\,Q$, $\zeta = \mathrm{Cov}(M, Q)$, and $\tau_X^2 = \mathrm{Var}\,M_X$, $\Omega_X = \mathrm{Var}\,Q_X$. Each of the following is a sufficient power condition:*

$$W_P := \mathcal{B}_{2B}(\tau_\Delta, N; e, \beta) + \sqrt{\rho}\,\mathcal{A}_{2B^2}(V, \Omega, N; e, \beta) < \gamma - s_P, \tag{7}$$

$$W_S := \sum_{X \in \{C,F\}} \left\{ \mathcal{B}_B(\tau_X, N; e, \beta) + \sqrt{\rho}\,\mathcal{A}_{B^2/2}(V_X, \Omega_X, N; e, \beta) \right\} < \gamma - s_S, \tag{8}$$

$$W_{G,\eta} := \mathcal{B}_{R_\eta}(\sigma_\eta, N; e_G, \beta) < \gamma - s_P - q_\eta, \tag{9}$$

where $\sigma_\eta^2 = \tau_\Delta^2 + \Omega/(16\eta^2) + \zeta/(2\eta)$, $q_\eta = \rho\eta + V/(4\eta) - \sqrt{\rho V}$, *and* $e_G \le \delta/(JH)$ *for the grid or* $e_G \le \delta/J$ *for an a priori fixed* η. *If the respective condition holds for every candidate, the family certifies with probability at least* $1 - 3J\beta$, $1 - 3(J+1)\beta$, *or* $1 - 2J\beta$, *respectively. The generic condition needs one outcome-independent grid choice per candidate. Choosing* $\beta = (1-p)/[6(J+1)]$ *suffices for power at least* p *for each procedure. Safety remains the separate* $1 - \delta$ *guarantee. A nonpositive right-hand margin cannot be repaired by these bounds.*

These explicit inequalities can be inverted numerically for N and charged KN evaluations. For split roles substitute the calibration size in $\mathcal{A}$ and audit size in $\mathcal{B}$. The formulas retain the anchor floor, sibling noise, calibration uncertainty, confidence multiplicity, and linear range cost. They are sufficient, not minimum required budgets.

Read the margin before increasing the budget. The right side of each power condition is the gain left after population envelope slack. If it is nonpositive, collecting more anchors cannot make that particular sufficient condition succeed. If it is positive but small, both the $N^{-1/2}$ uncertainty and bounded-range correction matter. Increasing K helps only the within-anchor terms; changing the response from separate losses to their difference changes which variation is paid for. Finally, changing the confidence construction affects multiplicity, range, and covariance costs even when the population robustness geometry is identical. These are distinct audit-design choices, so estimator comparisons must match the observations and confidence allocations before attributing a saving to pairing.

The local Gaussian forecast and paired/separate ratio are retained in Appendix D.8. They use population moments and provide no finite-pilot guarantee. The next section makes that distinction operational.

5 From an Oracle Cost to a Pilot-Based Plan

How many independent anchors should an auditor collect before running a robustness certificate? Population forecasts cannot answer this question without accounting for the uncertainty of the moments used in planning. We split a pilot of $r = r_0 + r_1$ anchors into a design block and an estimation block; the final audit is independent of both. Freeze a finite menu of $(N, K, \text{estimator}, \text{confidence allocation})$, candidates, loss, ρ, and known support enclosures before the pilot. The design block sets centering and tangency points. On the estimation block, construct a simultaneous $1 - \delta_{\text{plan}}$ confidence set $\mathcal{C}_r$ for every scalar moment used by the menu. It includes means, variances and third absolute moments of linear combinations of (M, Q); these combinations preserve $\mathrm{Cov}(M, Q)$ and paired cross-model covariance.

Two finite-sample witnesses assess each menu entry. The first takes the worst value of the power inequalities in Theorem 1 over $\mathcal{C}_r$. The second replaces the concave KL sensitivity radius by a global tangent and applies Berry–Esseen control to the resulting anchor response. For example, a paired endpoint is bounded by a mean of $Y = M + aQ$ plus a known intercept and its empirical-Bernstein radius. Estimating the variance of Y preserves the mean–sensitivity covariance. This is a finite-sample majorant, not a plug-in Gaussian approximation. Appendix E.3 gives the explicit inequality and its variance and remainder conditions.

Theorem 2 (Pilot planning, power, and its oracle limit)**.** *Assume bounded IID anchors, even IID siblings,* $p \ge 1/2$, *valid declared support enclosures, a finite frozen menu, and independent design, estimation, and final blocks. Choose a minimum-*N *menu entry, with frozen tie-breakers, for which one of the two witnesses holds for every frozen candidate; return* infeasible *if none exists. Then:*

1. *Conditional on* every *pilot history, the final selected certificate has false-certification probability at most* δ.
2. *With pilot probability at least* $1 - \delta_{\text{plan}}$, *every returned feasible plan has conditional final family-certification power at least* p. *This does not guarantee that a plan is returned.*
3. *If both pilot blocks grow, moments are consistent, enabled variance denominators stay positive, tangencies are continuous, and all finite-menu decisions are strictly separated from their boundaries, then* $\widehat{N}/N_{\text{wit}} \to 1$ *in probability whenever the same-witness population oracle* N_{wit} *is feasible.*

The third statement is *not* convergence to the Gaussian forecast, the actual minimum-power budget, or an information lower bound. It concerns the final block only: at a fixed law, growing the pilot sends total cost upward. Appendix E explains the corresponding small-margin obstruction for ordinary pilot confidence intervals. Unconditional power is at least $p\,\Pr(\text{good pilot and feasible})$, and is at least $(1-\delta_{\text{plan}})p$ only if a feasible plan is always returned. Pilot-selected duals need no final grid penalty: selection occurs before fresh audit outcomes. The pilot confidence set must nevertheless cover the entire search. Appendix E gives explicit confidence bounds, proofs, and the finite-menu stability argument.

Two different reasons to abstain. An infeasible pilot means that no entry in the declared menu has a power witness strong enough to justify its budget. It does not establish that CFPT is harmful. A feasible pilot, in contrast, authorizes collection of a fresh audit; that audit still returns factual training whenever no upper endpoint is negative. These outcomes should be counted separately. The theorem's family-power statement concerns certifying every member of the prescribed candidate family, so all of its required margins must be positive. Safety does not require that assumption: harmful candidates and failed pilot confidence sets are still protected by the independent final certificate. In reporting an audit, one should therefore give pilot feasibility, final success conditional on a returned plan, and end-to-end success, together with costs on every branch. A single success percentage cannot represent all three probability levels.

A returned plan is therefore a collection instruction, not a recommendation to deploy the CFPT model; deployment waits for the independent certificate. The pilot may select more siblings to stabilize nuisance estimation, or more anchors to reduce task-level uncertainty. Those choices answer different questions. In particular, strong covariance cancellation can make a paired audit economical while an uncertain mean or nuisance response still prevents the pilot from establishing a usable negative margin.

An information benchmark that charges the pilot. For a one-parameter subclass $A \sim \text{Bernoulli}(\theta)$ with θ-independent conditional nuisance/factual laws, choose $\Psi(\theta_1) = -\gamma$ and $\Psi(\theta_0) = 0$. Any procedure with unconditional power $q > \delta$, safety δ, and integrable total anchor count T satisfies

$$\mathbb{E}_{\theta_1} T \;\geq\; \frac{D_{\text{KL}}(q\|\delta)}{D_{\text{KL}}(\theta_1\|\theta_0)}. \tag{10}$$

The count includes design and estimation pilots and aborted plans. This local two-point testing benchmark follows by the likelihood chain rule and binary data processing; it is not a global minimax result. Increasing K cannot add information about θ once the anchor is known. Thus pilot overhead and confidence construction can leave a large gap even when pairing removes common factual variation.

6 Experiments

We test three distinct claims: pairing can reduce the evidence needed for a comparison; a pilot can turn unknown moments into a valid audit plan; and the cost of doing so can erase much of the population-level saving. Figure 2 separates these questions. Complete laws, frozen menus, simultaneous intervals, and negative outcomes appear in Appendix F. Our code is available at `https://anonymous.4open.science/r/counterfactual-audit-certification-6485/README.md`

Pairing and reuse are different sources of savings. Eight bounded-loss geometries vary cross-model correlation, anchor scale, and nuisance heterogeneity, holding the robust margin at $.025$. All four paired/separate and shared/split methods receive identical losses and matched confidence allocations, with $K = 2$. In this construction the sensitivity radii coincide, isolating the mean-covariance effect of pairing. Across the 90% thresholds, a symmetric decomposition of total log saving attributes factors 1.02–1.91 to pairing and 1.81–2.10 to reuse. Confirmation uses 1600 independent replications per budget; simultaneous binomial bands identify first grid crossings without assuming monotone power. Population-moment forecasts are within 3.04% of all 64 confirmed thresholds, whereas rigorous sufficient budgets are larger. This agreement tests cost accounting with known moments, not estimation of those moments from a pilot.

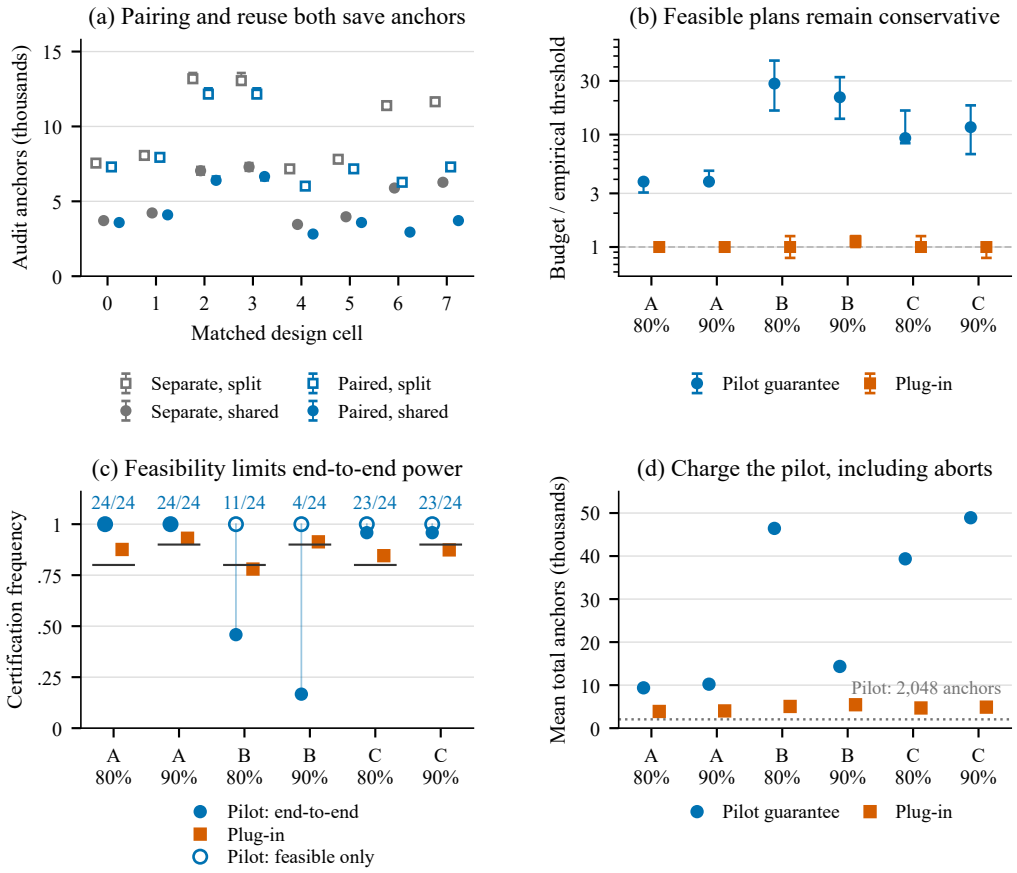


Figure 2: **Where audit savings arise, and where planning spends them.** **(a)** 90% certification thresholds for eight matched geometries; bars are simultaneous, per-geometry confidence sets for the first grid crossing (1600 replications). Pairing and anchor reuse are separate contrasts. **(b)–(d)** Independent-pilot study on laws A–C at 80%/90% target power, with 24 pilots per condition and 256 fresh final replications per returned plan. **(b)** Feasible-only median budgets and IQRs, normalized by each empirical menu threshold. **(c)** Filled circles include infeasible pilots as failures; open circles condition on feasibility. Labels count feasible pilots; black ticks mark target power. **(d)** Mean total anchors, including the 2048-anchor pilot on every branch. Lower total cost caused by more abstentions is not an efficiency improvement at fixed power. Panels (a) and (b)–(d) are different experiments.

An independent pilot exposes the cost of unknown moments. For planning, three different laws have robust margins .035, .025, .020 and equal paired/separate population envelopes. We fix one CF candidate, $\rho = .05$, safety $\delta = .025$, and planning failure .05 per law. Each pilot has 256 design and 1792 estimation anchors, with eight siblings. Known support enclosures are supplied to pilot bounds; final endpoints retain the global loss range. The frozen menu includes $K = 2, 8$, paired and separate KL endpoints, three fixed-dual choices, two confidence allocations, and 32 budgets. All 144 plans are locked before fresh final sampling. Exact multinomial draws preserve the full finite bounded anchor–sibling law, rather than replacing it by a Gaussian approximation.

All 109 feasible plans certify in all 256 final replications, with simultaneous fixed-plan 95% binomial lower bounds above .96. However, 35 pilots are infeasible; law B's end-to-end success frequencies are only .458 and .167. Plug-in planning is cheaper but underestimates some budgets: 29 of 144 plans have simultaneous upper power limits below their targets. Among valid plans, 43 select paired KL and 66 select generic DRO; all use the covariance-aware Berry–Esseen witness. Table 1 shows

Table 1: **Fresh audit budgets and pilot feasibility.** Pilot budgets are feasible-only medians. Empirical thresholds use an independent 1024-replication menu benchmark; the population sufficient bound is Theorem 1. Full grid intervals and evaluation costs are in Appendix F.1.

Law / p	Empirical	Gaussian	Sufficient	Pilot $\widehat{N}$	Feasible
A / 0.8	1,907	1,907	3,725	7,276	24/24
A / 0.9	1,907	1,907	4,657	7,276	24/24
B / 0.8	2,980	2,980	5,821	84,703	11/24
B / 0.9	2,980	2,980	7,276	64,036	4/24
C / 0.8	2,384	2,384	5,821	22,204	23/24
C / 0.9	2,980	2,980	5,821	34,694	23/24

that pilot uncertainty outweighs this witness's population advantage. Charging both $r + \widehat{N}$ anchors and $8r + K\widehat{N}$ paired-loss evaluations makes that overhead explicit.

A real-model certification, under its own sampling design. A prospective Qwen/ReCoRD audit (Qwen Team, 2025; Zhang et al., 2018) uses $1 - \mathrm{F1}$ loss, four enumerated prompt views, and disjoint 5000/10000/2000 calibration/audit/test article samples drawn without replacement. Before test access, the frozen generic endpoint selects RGF-15k ($-.00218$); the split paired endpoint falls back ($+.0004$). Test F1 is 16.44 versus 8.06 for factual training. Matched post-decision ablations on pre-test records isolate the contribution of pairing and reuse (Appendix G). This supports operational certification in a finite catalog; the IID sampled-sibling planning theorem is tested by the controlled laws above, not by this ReCoRD protocol.

7 Discussion and Conclusion

Counterfactual post-training creates related examples, but an apparent training gain becomes a reliable claim only through an appropriately designed comparison. The relevant response is the factual–CFPT loss difference: shared task difficulty and nuisance response can cancel before a confidence bound is formed. This makes pairing useful beyond the fact that the two models happen to be evaluated on the same dataset. Yet pairing does not create independent anchors, and additional siblings cannot remove the uncertainty that remains between anchors.

Our analysis connects this distinction to a practical question: how much evidence should an auditor collect? Shared-anchor certificates give safety, finite-sample inequalities describe sufficient power, and independent pilot planning accounts for uncertainty in the quantities used to choose the budget. These guarantees operate at different probability levels. A feasible plan can be reliable even when feasibility is rare; convergence to a same-witness population oracle can coexist with expensive pilot learning. The controlled experiments make both effects visible. Pairing and reuse save anchors in matched comparisons, while conservative planning and abstention remain substantial costs.

The resulting procedure is an auditable baseline for frozen models under declared covered shifts, rather than a guarantee about end-to-end CFPT optimization. Its main practical limitation is learning a tight enough negative margin from a small pilot. Known support enclosures and valid labels also restrict direct use with automatically generated text. Promising next steps are sharper sequential planning with total cost control, power guarantees for finite-catalog sampling, and independent semantic audits of generated interventions. The central design principle is to evaluate a model improvement and the evidence needed to establish it together: confidence in the final decision should not conceal uncertainty or cost in the plan that produced it.

AI USE STATEMENT

AI-assisted tools were used for language polishing, grammar checking, and literature retrieval and discovery (e.g., identifying relevant related work). The authors take full responsibility for the content and integrity of the manuscript.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In Hal Daumé, III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 145–155. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ahuja20a.html.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. doi: 10.3150/14-BEJ605. URL https://doi.org/10.3150/14-BEJ605.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://link.springer.com/article/10.1007/s10994-009-5152-4.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020. doi: 10.1109/JSAIT.2020.2991139. URL https://doi.org/10.1109/JSAIT.2020.2991139.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://jmlr.org/papers/v21/20-163.html.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5514–5528, Toronto, Canada, jul 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.302. URL https://aclanthology.org/2023.acl-long.302/.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2964–2984, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.216. URL https://aclanthology.org/2022.findings-emnlp.216/.

John C. Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019. URL https://jmlr.org/papers/v20/17-750.html.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. doi: 10.1214/20-AOS2004. URL https://doi.org/10.1214/20-AOS2004.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David Blei. Data augmentations for improved (large) language model generalization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70638–70653. Curran Associates, Inc., 2023. doi: 10.52202/075280-3096. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/df88b275bef31ac96c85f0c4013734fc-Paper-Conference.pdf.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. doi: 10.1287/mnsc.38.6.884. URL https://doi.org/10.1287/mnsc.38.6.884.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://doi.org/10.1214/20-AOS1991.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/johansson16.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C. Lipton. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02114.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx).

In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13201–13221, Torino, Italia, may 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1156/.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.learningtheory.org/colt2009/papers/012.pdf.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Aditya Krishna Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8–10):1561–1595, 2018. doi: 10.1007/s10994-018-5715-3. URL https://link.springer.com/article/10.1007/s10994-018-5715-3.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 10–18, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/muandet13.html. Part 1.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1670–1686, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.117. URL https://aclanthology.org/2022.acl-long.117/.

Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 708–717. PMLR, 2016. URL https://proceedings.mlr.press/v48/patrini16.html.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, pp. 857–869. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167. URL https://doi.org/10.1111/rssb.12167.

Marek Petrik, Mohammad Ghavamzadeh, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2298–2306, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/9a3d458322d70046f63dfd8b0153ece4-Abstract.html.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115v2.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/russo16.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://doi.org/10.1109/JPROC.2021.3058954.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.10745.

Irina Shevtsova. On the absolute constants in the Berry–Esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*, 2011. doi: 10.48550/arXiv.1111.6554. URL https://arxiv.org/abs/1111.6554.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3437–3452. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/steinke20a.html.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Amy L. Whitehead, Steven A. Julious, Cindy L. Cooper, and Michael J. Campbell. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research*, 25(3): 1057–1073, 2016. doi: 10.1177/0962280215588241. URL https://journals.sagepub.com/doi/10.1177/0962280215588241.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, jun 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Janet Wittes and Erica Brittain. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1–2):65–72, 1990. doi: 10.1002/sim.4780090113. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.4780090113.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/ad71c82b22f4f65b9398f76d8be4c615-Abstract.html.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018. doi: 10.48550/arXiv.1810.12885. URL https://arxiv.org/abs/1810.12885.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55006–55021. Curran Associates, Inc., 2023. doi: 10.52202/075280-2400. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161/.

Appendix

The appendix follows the dependencies of the theory: formal setup, supporting results, complete proofs, audit construction, and pilot planning. Controlled experiments and real-model studies then document the evidence at the sampling and semantic scope of each protocol.

Contents

A Problem Setup **17**

A.1 Invariant-task model . . . 17

A.2 Clustered counterfactual data . . . 18

A.3 CFPT policies and the learning map . . . 18

A.4 Comparative CFPT utility . . . 19

B Results **20**

B.1 The effective sample size of counterfactual data . . . 20

B.2 From representation consistency to covered-shift transfer . . . 21

B.3 Comparative CFPT utility under covered shifts . . . 22

C Proofs of the Supporting Results **24**

C.1 Proof of Theorem B.1 . . . 24

C.2 Proof of Theorem B.2 . . . 28

C.3 Proof of Proposition B.1 . . . 30

C.4 Proof of Proposition B.2 . . . 32

C.5 Proof of Theorem B.3 . . . 33

C.6 Proof of Proposition B.3 . . . 34

C.7 Proof of Proposition B.4 . . . 34

C.8 Proof of Proposition B.5 . . . 36

C.9 Proof of Corollary B.1 . . . 36

C.10 Scope of the comparative result . . . 36

D Shared-Anchor Certificates and Certification Power **37**

D.1 Proof of Proposition 1 . . . 37

D.2 Separate-model comparison . . . 37

D.3 Proof of Proposition D.1 . . . 37

D.4 Proof of Proposition 2 . . . 38

D.5 Proof of Proposition 3 . . . 39

D.6 Explicit finite-sample power widths . . . 39

D.7 Proof of Theorem 1 . . . 39

D.8 Population forecasts and covariance attribution . . . 40

D.9 Split-role certificates and power . . . 41

D.10 Proofs for split-role certificates . . . 43

E Pilot-Based Audit Planning: Construction and Proofs **45**

E.1 Statistical experiment and the information available to the planner . . . 45
E.2 Explicit simultaneous pilot intervals . . . 45
E.3 Power witnesses . . . 46
E.4 Global KL majorization and the finite-sample power witness . . . 46
E.5 Proof of Theorem 2 . . . 47
E.6 Why the oracle limit does not imply total-cost efficiency . . . 48
E.7 Proof of the local information benchmark . . . 48

F Controlled Audit Experiments **49**
F.1 Independent-pilot planning . . . 49
F.2 Costs, failures, and the limits of the observed result . . . 49
F.3 Matched pairing and anchor-reuse factorial . . . 50
F.4 Frozen design, confidence bands, and path independence . . . 52
F.5 Residual estimator ablations . . . 53
F.6 Targeted mean–variance covariance stress . . . 55
F.7 Exact-enumeration efficiency and allocation . . . 56
F.8 Calibration, anchors, and siblings are distinct resources . . . 56
F.9 Finite-calibration policy selection . . . 59
F.10 Analytic mechanism checks . . . 60
F.10.1 A loss-aligned covered-shift witness . . . 60
F.10.2 Local exactness for an arbitrary fixed pair . . . 60
F.10.3 Why paired utility can be tighter than two separate bounds . . . 61
F.10.4 Finite-sample anchor–sibling decomposition . . . 61
F.11 Finite-sample interpretation . . . 61

G Prospective ReCoRD Certification **61**
G.1 Prospective finite-catalog protocol . . . 61
G.2 ReCoRD attribution only . . . 63
G.3 Retrospective finite-view comparison . . . 63

H Language-Model Diagnostics and Scope **63**
H.1 Exploratory training-volume study . . . 63
H.2 Task and common protocol . . . 64
H.3 Operational proxies and estimands . . . 64
H.4 Paired utility audit on locked Qwen predictions . . . 64
H.5 Cluster scaling and fixed synthetic budget . . . 64
H.6 Covered-transfer intervention . . . 65
H.7 Matched-error mechanism study . . . 65
H.8 Evidence boundary . . . 65
H.9 Shift-granularity diagnostic . . . 65
H.10 Failure boundaries: semantic corruption and support escape . . . 66
H.11 HANS: sensitivity beyond average reference gain . . . 66

H.12 HANS: what can the existing information certify? 67
H.13 QA: aligning the audit with generation quality 68
H.14 A wrong-shift-class boundary 68
H.15 HANS: an external high-headroom diagnostic 68

A Problem Setup

A.1 Invariant-task model

CFPT changes nuisance S while preserving the response law determined by task factor C; representation Z and action A determine prediction loss. Here C is a semantic task state and need not be directly observed. The phrase "task-preserving intervention" means that the intervention changes only S while keeping the conditional response kernel fixed. A generated rewrite that changes the underlying task state is outside this model, even if an automatic verifier accepts it. Let C, S, Y, Z and A be standard Borel spaces, let μ be a law on C, let $q(dy \mid c)$ be the invariant response kernel, and let $\mathsf{x} : C \times S \to X$ be a measurable input map. In environment e,

$$C \sim \mu, \qquad Y \mid C = c \sim q(\cdot \mid c), \qquad S \mid C = c \sim \pi_e(\cdot \mid c), \qquad X = \mathsf{x}(C, S).$$

The reference intervention law is $\nu(ds \mid c)$. A model w has kernels $Q_w(dz \mid x)$ and $M_w(da \mid z)$, so $P_w(dz \mid c, s) = Q_w(dz \mid \mathsf{x}(c, s))$. For a measurable bounded loss $0 \le \ell \le B$, define

$$r_w(c, s) = \int q(dy \mid c) \int P_w(dz \mid c, s) \int M_w(da \mid z)\ell(a, y),$$
$$R_e(w) = \int \mu(dc) \int \pi_e(ds \mid c) r_w(c, s), \qquad R_\nu(w) = \int \mu(dc) \int \nu(ds \mid c) r_w(c, s).$$

Transfer concerns $R_e - R_\nu$. In LLMs, C is task intent and S is style, domain, template, language, or format; ν and $\widehat{\nu}$ are the intended and generator-induced edit laws, respectively.

The reference representation mixture and the counterfactual quantities are

$$P_w^\nu(dz \mid c) = \int P_w(dz \mid c, s')\nu(ds' \mid c),$$
$$I_\nu(w) = \int \mu(dc) \int \nu(ds \mid c) D_{\mathrm{KL}}(P_w(\cdot \mid c, s) \| P_w^\nu(\cdot \mid c)),$$
$$L_{\mathrm{CF}}(w) = \int \mu(dc) \int \nu(ds \mid c) \int \nu(ds' \mid c) D_{\mathrm{KL}}(P_w(\cdot \mid c, s) \| P_w(\cdot \mid c, s')),$$
$$J_{\mathrm{rev}}(w) = \int \mu(dc) \int \nu(ds \mid c) D_{\mathrm{KL}}(P_w^\nu(\cdot \mid c) \| P_w(\cdot \mid c, s)),$$
$$V_c(w) = \mathrm{Var}_{S \sim \nu(\cdot \mid c)}[r_w(c, S)], \qquad V_\nu(w) = \int V_c(w)\, \mu(dc).$$

All displayed maps and probability kernels are measurable, and KL is initially interpreted as an extended-real quantity. For the decomposition, Theorem B.2's finite pairwise certificate supplies one-sided KL finiteness and absolute continuity.

Assumption A.1 (Invariant task and valid intervention)**.** All environments share μ and q. Only π_e may vary. The reference and approximate intervention kernels modify S conditional on the same C and leave the response kernel equal to q.

The assumption is a statement about the data-generating law, not a guarantee provided by a text generator. In the source-tag experiments, C is the semantic content of the original premise–hypothesis pair and S is the appended source cue, so the assumption is exact by construction. For free natural-language rewrites, semantic preservation is an empirical condition that must be audited separately; the automatic screen used below does not turn it into a theorem premise.

Assumption A.2 (Square-integrable test coverage)**.** $\pi_e(\cdot \mid c) \ll \nu(\cdot \mid c)$ for μ-almost every c, with a measurable density ratio

$$a_e(c,s) := \frac{d\pi_e(\cdot \mid c)}{d\nu(\cdot \mid c)}(s), \qquad \int \mu(dc) \int (a_e(c,s)-1)^2 \nu(ds \mid c) < \infty.$$

Assumption A.3 (Intervention-law fidelity)**.** The approximate intervention kernel $\widehat{\nu}$ has nuisance-law mismatch

$$\int \mathrm{TV}\big(\widehat{\nu}(\cdot \mid c), \nu(\cdot \mid c)\big)\, \mu(dc) \le \varepsilon_{\mathrm{gen}}.$$

A uniform essential-supremum TV bound by $\varepsilon_{\mathrm{gen}}$ is a sufficient, but not necessary, condition for Assumption A.3.

For the centered transfer result, define

$$\chi^2_c = \int (a_e(c,s)-1)^2 \nu(ds \mid c),$$
$$I_c(w) = \int \nu(ds \mid c) D_{\mathrm{KL}}(P_w(\cdot \mid c,s) \| P^{\nu}_w(\cdot \mid c)),$$
$$\bar{\chi}^2_e = \int \chi^2_c \mu(dc).$$

A.2 Clustered counterfactual data

Let $T_i = (C_i, Y_i)$ be IID from $\mu(dc)q(dy \mid c)$. Given T_i, draw $S_{i1:K}$ IID from $\widehat{\nu}(\cdot \mid C_i)$ and set $D_i = (C_i, Y_i, S_{i1:K})$. Distinct clusters are conditionally independent. Let $\mathcal{D}$ be the resulting standard Borel cluster space. Let the seed space $\mathcal{W}_0$ and model space $\mathcal{W}$ be standard Borel, and let $\mathcal{A} : \mathcal{W}_0 \times \mathcal{D}^n \to \mathcal{W}$ be measurable. Let $W_0 \perp D_{1:n}$ and $W = \mathcal{A}(W_0, D_{1:n})$. The cluster loss and population quantities are

$$\phi_w(D_i) = \frac{1}{K}\sum_{j=1}^{K} \ell_w(C_i, S_{ij}, Y_i), \quad R_{n,K}(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_w(D_i), \quad R_{\widehat{\nu}}(w) = \mathbb{E}_D[\phi_w(D)].$$

For a fresh evaluation collection, $R'_{n,K}(w)$ denotes the corresponding empirical average, independent of the learned model in the fresh statement. For an independent fresh model $w_\perp$, define $T = (C, Y)$,

$$m_w(T) = \mathbb{E}_{S|T}\ell_w(T,S), \quad v_w(T) = \mathrm{Var}_{S|T}[\ell_w(T,S)],$$
$$\sigma^2_T(w) = \mathrm{Var}_T[m_w(T)], \quad \widehat{\sigma}^2_{S,n}(w) = \frac{1}{n}\sum_{i=1}^{n} v_w(T_i).$$

where the conditional law of S is $\widehat{\nu}(\cdot \mid C)$.

The seed W_0 contains the training-randomness sources sampled independently of the data; a deterministic learner has constant W_0. The conditional mutual information measures data dependence with that seed revealed. It does not generally discount information: independence gives

$$I(W; D \mid W_0) = I(W; D) + I(W_0; D \mid W) \ge I(W; D).$$

For independent uniform bits D, W_0 and $W = D \operatorname{xor} W_0$, the unconditional information is zero but the conditional information is $\log 2$. Conditioning can reveal dependence masked by randomization. Expectations in the adaptive statement average over seeds and training clusters. Finite conditional information is an additional assumption; adding noise whose full seed is conditioned on does not itself ensure it. Randomness chosen adaptively from data belongs in the data-dependent output W.

A.3 CFPT policies and the learning map

The comparative results below are conditional on a learned factual–CFPT pair, but the pair is not an unexplained primitive of the experimental pipeline. We make the training interface explicit here. A CFPT policy is a tuple

$$h = (\alpha, \beta, q, K, \mathcal{G}, \mathcal{A}),$$

where α controls the weight of generated supervision, β the weight of a consistency term, q the within-anchor sibling weights, K the number of generated siblings, $\mathcal{G}$ the generator, and $\mathcal{A}$ the training algorithm and its fixed compute budget. Only coordinates that are varied in an experiment are treated as policy coordinates; optimizer settings that are held fixed are part of the common budget.

For a factual training block D_F and a policy-specific generated block $D_{\mathrm{CF},h}$, write

$$\widehat{R}_F^{\mathrm{tr}}(\theta) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \ell_{\mathrm{tr}}(\theta; x_i, y_i), \tag{A.1}$$

$$\widehat{R}_{\mathrm{CF},h}^{\mathrm{tr}}(\theta) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \sum_{j=1}^{K} q_{ij}^h \ell_{\mathrm{tr}}(\theta; x_{ij}^h, \widetilde{y}_{ij}^h), \qquad q_{ij}^h \geq 0, \quad \sum_j q_{ij}^h = 1. \tag{A.2}$$

The policy objective is

$$\widehat{J}_h(\theta) = (1-\alpha)\widehat{R}_F^{\mathrm{tr}}(\theta) + \alpha \widehat{R}_{\mathrm{CF},h}^{\mathrm{tr}}(\theta) + \beta \widehat{L}_{\mathrm{cons},h}^{\mathrm{tr}}(\theta) + \lambda_{\mathrm{reg}}\Omega(\theta). \tag{A.3}$$

The actual trained model is represented by the measurable learning map

$$\widehat{\theta}_h = \mathcal{A}_{\mathcal{B}}(\theta_{\mathrm{init}}, D_F, D_{\mathrm{CF},h}, h, \xi), \tag{A.4}$$

where $\mathcal{B}$ is the fixed optimization budget and ξ collects randomness independent of the training data. The factual baseline is the same map with the counterfactual and consistency weights set to zero, or the corresponding pre-registered baseline policy when a study explicitly fixes a different comparison.

Training loss ℓ_{tr} and evaluation loss ℓ need not be the same. Theorems B.1–B.3 use the bounded evaluation envelope $0 \leq \ell \leq B$; in the Qwen experiments we therefore use bounded action loss or a clipped predictive loss for the certificate and report unclipped NLL only as a secondary descriptive metric. Equation (A.4) is an interface, not an end-to-end optimization theorem: the comparative theorem audits the learned pair after training and does not assert that increasing α or β improves the optimizer's outcome.

A.4 Comparative CFPT utility

The first two results describe sampling and transfer for one learned model. The final main result compares a factual baseline and a counterfactual post-trained model. We condition on the training sigma-field when the pair has already been learned; all quantities below are then evaluated on fresh anchors unless stated otherwise.

Let w_F and w_C denote the factual and CFPT models, respectively. Their conditional reference risks are already defined by $r_w(c,s)$ in Section A. Define the signed paired loss difference

$$g(c,s) := r_{w_C}(c,s) - r_{w_F}(c,s), \qquad -B \leq g(c,s) \leq B. \tag{A.5}$$

For any conditional nuisance law $\pi(\cdot \mid c)$ sharing the task marginal μ, write

$$\Delta_\pi := R_\pi(w_C) - R_\pi(w_F) = \int \mu(dc) \int \pi(ds \mid c) g(c,s). \tag{A.6}$$

Thus $\Delta_\pi < 0$ means that CFPT improves over the factual baseline under π, whereas $\Delta_\pi > 0$ means harm. The reference conditional mean and paired differential sensitivity are

$$\bar{g}(c) := \int g(c,s)\nu(ds \mid c), \tag{A.7}$$

$$V_c^\Delta := \int \big(g(c,s) - \bar{g}(c)\big)^2 \nu(ds \mid c), \qquad V_\nu^\Delta := \int V_c^\Delta \, \mu(dc). \tag{A.8}$$

Unlike $V_\nu(w_F)$ and $V_\nu(w_C)$ separately, V_ν^Δ measures only the nuisance-dependent part of the *comparison* between the two models.

For a deployment law π, assume the conditional density ratio $a_\pi(c,s) = d\pi(\cdot \mid c)/d\nu(\cdot \mid c)$ exists and define

$$\chi_c^2(\pi\|\nu) := \int (a_\pi(c,s) - 1)^2 \nu(ds \mid c), \qquad \bar{\chi}_\pi^2 := \int \chi_c^2(\pi\|\nu)\, \mu(dc). \tag{A.9}$$

The associated covered shift class is

$$\mathcal{E}_\rho(\nu) := \left\{\pi : \pi(\cdot \mid c) \ll \nu(\cdot \mid c) \text{ for } \mu\text{–a.e. } c,\ \bar{\chi}^2_\pi \le \rho,\ \pi_C = \mu\right\}. \tag{A.10}$$

If the task marginal also changes, the resulting task-shift term is outside this conditional nuisance-shift theorem and must be controlled separately.

Assumption A.4 (Fresh comparative evaluation)**.** The pair (w_F, w_C) is fixed conditional on the training sigma-field, and the anchors and sibling interventions used to estimate Δ_ν and V_ν^Δ are independent of that sigma-field. If model selection uses a calibration pool, that pool is not reused as the reported fresh evaluation pool.

Assumption A.5 (Reference-law fidelity for comparative quantities)**.** When the observed generator law is $\widehat{\nu}$ rather than ν, the mean-difference error satisfies

$$\int \mathrm{TV}(\widehat{\nu}(\cdot \mid c), \nu(\cdot \mid c))\, \mu(dc) \le \varepsilon_{\mathrm{gen}},$$

and the reported upper confidence estimate for V_ν^Δ includes a separate valid bound for its $\widehat{\nu}$–to–ν mismatch. Setting $\widehat{\nu} = \nu$ makes this latter term zero.

The reference-law condition is distinct from semantic validity. To make this distinction explicit, let P_h^{val} denote the task-preserving rewrite law for policy h and let P_h^{obs} denote the law actually seen by an audit. An optional semantic audit may certify the contamination model

$$P_h^{\mathrm{obs}} = (1 - \eta_{\mathrm{sem},h}) P_h^{\mathrm{val}} + \eta_{\mathrm{sem},h} Q_h, \qquad 0 \le \eta_{\mathrm{sem},h} \le 1, \tag{A.11}$$

for an arbitrary failure law Q_h. Here $\varepsilon_{\mathrm{nuis},h}$ controls mismatch between the valid nuisance law and the intended reference law, whereas $\eta_{\mathrm{sem},h}$ controls violations of the invariant-task assumption. They are not interchangeable. Since a paired loss difference lies in $[-B, B]$, the contamination contribution to a fixed-pair audit is at most $2B\eta_{\mathrm{sem},h}$. Without an independent semantic audit, this quantity is unknown and cannot be set to zero. This interface controls the audit distribution; it does not claim to undo the effect of corrupted training examples on the learned model itself.

B Results

The results follow the intervention lifecycle: clustered sampling, transfer under a covered nuisance shift, and the competition between robustness and semantic corruption.

B.1 The effective sample size of counterfactual data

Theorem B.1 (Two-level adaptive information and fresh concentration)**.** *Under the cluster protocol and $I(W; D_{1:n} \mid W_0) < \infty$, set $\mathcal{I}_T = I(W; T_{1:n} \mid W_0)$ and $\mathcal{I}_S = I(W; S_{1:n,1:K} \mid W_0, T_{1:n})$. Then $I(W; D_{1:n} \mid W_0) = \mathcal{I}_T + \mathcal{I}_S$ and*

$$\left|\mathbb{E}\left[R_{\widehat{\nu}}(W) - R_{n,K}(W)\right]\right| \le B \min\left\{\sqrt{\frac{\mathcal{I}_T + \mathcal{I}_S}{2n}},\ \sqrt{\frac{\mathcal{I}_T}{2n}} + \sqrt{\frac{\mathcal{I}_S}{2nK}}\right\}. \tag{B.1}$$

The first branch treats complete clusters as the independent units; the second separates anchor and intervention adaptivity. Its intervention term vanishes as K grows only when $\mathcal{I}_S = o(nK)$. For a fresh collection and an independent model $w_\perp$, let $t = \log(4/\delta)$. With probability at least $1 - \delta$,

$$\left|R'_{n,K}(w_\perp) - R_{\widehat{\nu}}(w_\perp)\right| \le \sqrt{\frac{2\sigma_T^2(w_\perp)t}{n}} + \frac{2Bt}{3n} + \sqrt{\frac{2\widehat{\sigma}^2_{S,n}(w_\perp)t}{nK}} + \frac{2Bt}{3nK}. \tag{B.2}$$

The anchor-only construction in Appendix C.1 rules out a universal adaptive theorem whose only scale is $1/\sqrt{nK}$.

The proof is given in Appendix C.1. If $\mathcal{I}_S \le n\iota_S$ for an intervention-information budget ι_S, the second adaptive contribution is at most $B\sqrt{\iota_S/(2K)}$; without such control its numerator may grow with K.

Anchor-only example. To see why repeated edits cannot replace independent anchors, let $C_i \sim \mathrm{Bernoulli}(1/2)$ and take the bounded loss $\ell(C_i, S, Y) = BC_i$, with any conditional intervention law. Every intervention attached to anchor i then has the same loss, so for every K,

$$R_{n,K} = B\frac{1}{n}\sum_{i=1}^{n} C_i, \qquad R_{\widehat{\nu}} = \frac{B}{2}, \qquad \mathrm{Var}(R_{n,K}) = \frac{B^2}{4n}.$$

With probability 2^{1-n}, all anchors are equal and the deviation is $B/2$. Thus increasing K leaves the anchor fluctuation unchanged. The full construction and the resulting impossibility statement are given in Appendix C.1.

The adaptive statement is an information-stability guarantee: it requires finite conditional mutual information between the training clusters and the learned model. Finite output alphabets provide a sufficient condition. Randomization or additive noise alone need not suffice after revealing the full independent seed; the conditional information assumption must be checked.

B.2 From representation consistency to covered-shift transfer

Theorem B.2 (Counterfactual consistency under a covered shift)**.** *Under Assumptions A.1– A.3 and* $L_{\mathrm{CF}}(w) < \infty$*,*

$$L_{\mathrm{CF}}(w) = I_\nu(w) + J_{\mathrm{rev}}(w),$$

where in particular $I_\nu(w), J_{\mathrm{rev}}(w) \le L_{\mathrm{CF}}(w)$*. The head-aware risk sensitivity obeys*

$$|R_e(w) - R_\nu(w)| \le \int \sqrt{\chi_c^2 V_c(w)}\,\mu(dc) \le \sqrt{\bar{\chi}_e^2 V_\nu(w)}.$$

The representation certificate gives the complementary hierarchy

$$|R_e(w) - R_\nu(w)| \le \frac{B}{\sqrt{2}}\int \sqrt{\chi_c^2 I_c(w)}\,\mu(dc) \le B\sqrt{\frac{\bar{\chi}_e^2 I_\nu(w)}{2}} \le B\sqrt{\frac{\bar{\chi}_e^2 L_{\mathrm{CF}}(w)}{2}}.$$

Moreover,

$$|R_{\widehat{\nu}}(w) - R_\nu(w)| \le B\varepsilon_{\mathrm{gen}}.$$

Consequently, with

$$U_{\mathrm{CF}}(w) = R_{\widehat{\nu}}(w) + B\varepsilon_{\mathrm{gen}} + B\sqrt{\frac{\bar{\chi}_e^2 L_{\mathrm{CF}}(w)}{2}},$$

we have $R_e(w) \le U_{\mathrm{CF}}(w)$*. If the density ratio is bounded by* ρ_e*, then* $\bar{\chi}_e^2 \le \rho_e - 1$*, giving a density-ratio bound. In particular,* $\pi_e = \nu$ *or* $I_\nu(w) = 0$ *implies* $R_e(w) = R_\nu(w)$*.*

The proof is given in Appendix C.2.

The risk bound couples test coverage to nuisance variation used by the head. The forward–reverse information identity is known in predictive-uncertainty analysis (Malinin & Gales, 2021). Its role here is to connect measurable counterfactual consistency to this conditional transfer interface.

Proposition B.1 (Head-aware and operational interfaces)**.** *For the Gaussian-smoothed channel* $Z = f_w(\mathsf{x}(C,S)) + \xi$*,* $\xi \sim \mathcal{N}(0, \sigma^2 I)$*,* $L_{\mathrm{CF}}(w) = (2\sigma^2)^{-1}\mathbb{E}\|f_w(\mathsf{x}(C,S)) - f_w(\mathsf{x}(C,S'))\|_2^2$*. For the induced action kernel, data processing and Pinsker give*

$$V_\nu \le B^2 I_\nu^A/2 \le B^2 L_{\mathrm{CF}}^A/2 \le B^2 L_{\mathrm{CF}}/2.$$

Together with Theorem B.2, these are action- and representation-level certificates for the displayed stochastic channels, not a claim about arbitrary deterministic networks.

The proof, including the Lipschitz–Wasserstein mismatch variant, is given in Appendix C.3.

The task-preserving witness fixes $Y = C$ and changes only the test-time association of a nuisance with C; it separates representation leakage from head use and is fully evaluated in Appendix F.10.

Proposition B.2 (Sharpness and necessity of the transfer interface)**.** *The head-aware bound is exactly attainable; the representation certificate has the same local order in the task-preserving witness (and becomes zero when the head ignores the nuisance). If the test law escapes the reference support, there are invariant-task instances with zero reference risk, zero consistency, and zero generator mismatch but arbitrary positive test risk. Hence coverage is necessary for any nontrivial reference-only guarantee.*

The complete proof is given in Appendix C.4.

B.3 Comparative CFPT utility under covered shifts

Theorem B.2 controls the transfer of one model's reference risk. A post-training decision, however, compares a factual model with a counterfactual model. The relevant object is therefore the nuisance sensitivity of their loss difference, not the two sensitivities considered separately.

Theorem B.3 (Comparative CFPT utility under covered nuisance shifts)**.** *Under Assumptions A.1 and A.2, fix a pair (w_F, w_C) and define $g, \Delta_\pi, V_\nu^\Delta$ by (A.5)–(A.8). Then, for every covered deployment law π sharing the task marginal μ,*

$$\Delta_\pi - \Delta_\nu = \int \mu(dc) \int \nu(ds \mid c)\,(a_\pi(c,s) - 1)\big(g(c,s) - \bar{g}(c)\big), \tag{B.3}$$

and consequently

$$|\Delta_\pi - \Delta_\nu| \le \int \sqrt{\chi_c^2(\pi\|\nu) V_c^\Delta}\,\mu(dc) \le \sqrt{\bar{\chi}_\pi^2 V_\nu^\Delta}. \tag{B.4}$$

In particular, for the shift class in (A.10),

$$\Delta_\nu - \sqrt{\rho V_\nu^\Delta} \le \Delta_\pi \le \Delta_\nu + \sqrt{\rho V_\nu^\Delta} \qquad (\pi \in \mathcal{E}_\rho(\nu)). \tag{B.5}$$

$$\sqrt{V_\nu^\Delta} \le \sqrt{V_\nu(w_C)} + \sqrt{V_\nu(w_F)}. \tag{B.6}$$

The paired radius is no larger than the separate-model radius. The inequality is strict unless the centered nuisance responses meet the equality condition of the $L^2(\mu \otimes \nu)$ triangle inequality. Thus

$$\Delta_\nu + \sqrt{\rho V_\nu^\Delta} < 0 \Longrightarrow \Delta_\pi < 0 \quad \text{for all} \quad \pi \in \mathcal{E}_\rho(\nu),$$

which certifies benefit over the whole covered shift class. The condition $\Delta_\nu - \sqrt{\rho V_\nu^\Delta} > 0$ certifies harm, and all other cases are uncertain. The coefficient one in (B.4) is locally sharp: with one task state, $S \in \{-1, +1\}$, uniform ν, $g(S) = m + \alpha S$, and $\pi_\delta(S = s) = (1 + \delta s)/2$, equality holds for $|\delta| \le 1$.

The proof is given in Appendix C.5. The theorem is a comparative extension of the single-model transfer interface: a common nuisance response cancels from g, so a pair can have a smaller utility-transfer radius even when both models are individually sensitive. It is a certificate for a fixed learned pair under fresh evaluation, not a guarantee that the CF objective improves optimization.

Proposition B.3 (Local exactness of the paired robust interval)**.** *For the fixed pair in Theorem B.3, let $h(c,s) = g(c,s) - \bar{g}(c)$, $V = V_\nu^\Delta$, and $M = \operatorname{ess\,sup}_{\mu\otimes\nu} |h(c,s)|$. If $V > 0$ and $0 \le \rho \le V/M^2$, then the endpoints in (B.5) are attainable:*

$$\sup_{\pi \in \mathcal{E}_\rho(\nu)} \Delta_\pi = \Delta_\nu + \sqrt{\rho V_\nu^\Delta}, \qquad \inf_{\pi \in \mathcal{E}_\rho(\nu)} \Delta_\pi = \Delta_\nu - \sqrt{\rho V_\nu^\Delta}. \tag{B.7}$$

Thus the paired interval is an exact robust utility envelope for every fixed model pair in this local shift regime, rather than only a bound attained by a special binary witness. The statement is for the global conditional chi-square class in (A.10); $V = 0$ is the degenerate case in which the interval collapses.

The proof is given in Appendix C.6.

Proposition B.4 (Finite-sample paired utility certificate)**.** *Let $T_i = (C_i, Y_i)$ be the anchor task state and let*

$$Z_{ij} = \ell(w_C; C_i, S_{ij}, Y_i) - \ell(w_F; C_i, S_{ij}, Y_i)$$

be the paired loss difference on an anchor–sibling evaluation block, with $|Z_{ij}| \le B$. Siblings share Y_i and are conditionally independent given T_i. Write

$$m(T_i) = \mathbb{E}[Z_{ij} \mid T_i], \qquad v(T_i) = \operatorname{Var}(Z_{ij} \mid T_i),$$

and define

$$\sigma_A^2 = \operatorname{Var}(m(T_i)), \qquad \sigma_S^2 = \mathbb{E}[v(T_i)].$$

For n independent anchors and K conditionally independent siblings per anchor, let $\widehat{\Delta}$ be the empirical average of the Z_{ij} and let $\Delta_{\hat{\nu}}$ denote its population mean under the observed generator

law. Suppose an independent calibration block supplies valid upper estimates $\widehat{\sigma}_A^2 \geq \sigma_A^2$ *and* $\widehat{\sigma}_S^2 \geq n^{-1}\sum_{i=1}^n v(T_i)$ *on an event of probability at least* $1-\delta/2$*. Also let* $\overline{O}_g$ *be a valid, jointly calibrated upper bound on* $\mathrm{osc}(g)$ *(the always-valid choice is* $\overline{O}_g = 2B$*). The second inequality is a valid upper bound for the realized conditional sibling variance of the fresh anchors; it may be obtained from an independent calibration block plus an anchor-level concentration term. With* $t = \log(8/\delta)$*, define*

$$\Gamma_{n,K} := \underbrace{\frac{2Bt}{3n} + \sqrt{\frac{2\widehat{\sigma}_A^2 t}{n} + \frac{4B^2t^2}{9n^2}}}_{\Gamma_A(t)} + \underbrace{\frac{2Bt}{3nK} + \sqrt{\frac{2\widehat{\sigma}_S^2 t}{nK} + \frac{4B^2t^2}{9n^2K^2}}}_{\Gamma_S(t)}. \tag{B.8}$$

When an independent calibration block supplies a valid upper bound $\widehat{\sigma}_{\mathrm{cl}}^2 \geq \sigma_A^2 + \sigma_S^2/K$*, the IID cluster means also give the combined radius*

$$\Gamma_{n,K}^{\mathrm{cl}}(\delta) := \frac{2Bt}{3n} + \sqrt{\frac{2\widehat{\sigma}_{\mathrm{cl}}^2 t}{n} + \frac{4B^2t^2}{9n^2}}. \tag{B.9}$$

Both radii are valid on their stated joint calibration events. The decomposed version exposes the anchor and sibling terms; the combined version is tighter when the cluster-variance calibration is reliable. With no variance calibration, the cluster-Hoeffding fallback

$$\Gamma_n^{\mathrm{H}}(\delta) := B\sqrt{\frac{2\log(2/\delta)}{n}} \tag{B.10}$$

is also valid because each cluster mean lies in $[-B, B]$*. Then, with probability at least* $1-\delta$*, simultaneously,*

$$|\widehat{\Delta} - \Delta_{\widehat{\nu}}| \leq \Gamma_{n,K}, \qquad \mathrm{Var}(\widehat{\Delta}) = \frac{\sigma_A^2}{n} + \frac{\sigma_S^2}{nK}. \tag{B.11}$$

If, in addition, the generator-fidelity and sensitivity calibration event is included in the same joint calibration budget and provides $\varepsilon_{\mathrm{gen}}$ *and* $\overline{V}_\nu^\Delta \geq V_\nu^\Delta$*, then for every* $\pi \in \mathcal{E}_\rho(\nu)$*,*

$$U_\Delta := \widehat{\Delta} + \Gamma_{n,K} + \overline{O}_g\varepsilon_{\mathrm{gen}} + \sqrt{\rho\,\overline{V}_\nu^\Delta} \tag{B.12}$$

is a valid upper certificate for Δ_π *on the joint calibration and evaluation event. The displayed probability* $1-\delta$ *applies only after all calibration sub-events have been assigned a total failure budget no larger than* $\delta/2$*. In particular,* $U_\Delta < 0$ *certifies benefit,* $\widehat{\Delta} - \Gamma_{n,K} - \overline{O}_g\varepsilon_{\mathrm{gen}} - \sqrt{\rho\,\overline{V}_\nu^\Delta} > 0$ *certifies harm, and all other outcomes are reported as uncertain. If the observed audit law additionally follows the contamination model in (A.11), the valid-task certificate adds the term* $2B\eta_{\mathrm{sem}}$*; this term is separate from the nuisance-law correction and is available only when its bound is independently calibrated.*

The proof is given in Appendix C.7. The proposition makes the theorem operational without treating siblings as independent anchors: the first radius is an $n^{-1/2}$ anchor term and the second is an $(nK)^{-1/2}$ within-anchor term. The certificate is conditional on valid calibration radii and fresh evaluation; it is not a claim that the paired diagnostic dominates direct target-law validation when clean target labels are available.

The proposition also has a fully observable, deliberately conservative version. Since $|Z_{ij}| \leq B$ implies $\sigma_A^2 \leq B^2$ and $v(T) \leq B^2$, one may set both variance upper estimates equal to B^2 and avoid an unknown-variance oracle. The resulting radius is

$$\Gamma_{n,K}^{\mathrm{bd}}(\delta) = \frac{2Bt}{3n} + \sqrt{\frac{2B^2t}{n} + \frac{4B^2t^2}{9n^2}} + \frac{2Bt}{3nK} + \sqrt{\frac{2B^2t}{nK} + \frac{4B^2t^2}{9n^2K^2}}, \qquad t = \log(8/\delta). \tag{B.13}$$

This bound is often wider than a calibrated empirical-variance radius, but it is computable from the evaluation-loss envelope alone. A tighter calibration may replace either B^2 only when its own probability event is included in the joint error budget.

Proposition B.5 (Finite-class paired-sensitive safe selection)**.** *Let* $\mathcal{H}$ *be a finite set of* J *CFPT policies fixed before the audit block is observed, and let* $h = 0$ *denote the factual baseline. Conditional on the*

training data, each $h \in \mathcal{H}$ produces a fixed model pair (w_0, w_h) through (A.4). On an independent paired audit block, compute $\widehat{\Delta}_h$ with the shared task label Y_i. Write $g_h = r_{w_h} - r_{w_0}$ for its paired loss difference. For the contamination interface, define the three corresponding means

$$\Delta_h^{\mathrm{obs}} := \mathbb{E}_{P_h^{\mathrm{obs}}}[g_h], \qquad \Delta_h^{\mathrm{val}} := \mathbb{E}_{P_h^{\mathrm{val}}}[g_h], \qquad \Delta_{\nu,h}^{\mathrm{val}} := \mathbb{E}_{\mu\otimes\nu}[g_h].$$

Write $\Delta_{\pi,h}^{\mathrm{val}}$ for the paired risk difference of this fixed model pair under the task-preserving valid component of the deployment law. Let $\overline{O}_h$ be a valid candidate-specific upper bound on $\mathrm{osc}(g_h)$; the bounded choice $\overline{O}_h = 2B$ is always available. Suppose the candidate-specific audit and calibration events satisfy, jointly over $h \in \mathcal{H}$,

$$\left|\widehat{\Delta}_h - \Delta_h^{\mathrm{obs}}\right| \le \Gamma_h, \qquad V_{\nu,h}^{\Delta} \le \overline{V}_{\nu,h}^{\Delta}, \tag{B.14}$$

and supply valid bounds $\varepsilon_{\mathrm{nuis},h}$ and $\eta_{\mathrm{sem},h}$ for the two terms in (A.11). The total failure probability of all these audit, sensitivity, nuisance, and semantic events is at most δ. Define

$$U_h^{\mathrm{pair}}(\rho,\delta) := \widehat{\Delta}_h + \Gamma_h + \overline{O}_h\varepsilon_{\mathrm{nuis},h} + 2B\eta_{\mathrm{sem},h} + \sqrt{\rho\,\overline{V}_{\nu,h}^{\Delta}}. \tag{B.15}$$

Select

$$\widehat{h} = \begin{cases} \arg\min_{h\in\mathcal{H}} U_h^{\mathrm{pair}}(\rho,\delta), & \min_{h\in\mathcal{H}} U_h^{\mathrm{pair}}(\rho,\delta) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{B.16}$$

Then, with probability at least $1-\delta$, every covered valid deployment law $\pi \in \mathcal{E}_\rho(\nu)$ satisfies

$$R_\pi(w_{\widehat{h}}) - R_\pi(w_0) \le 0.$$

The guarantee is a safety statement only: it does not assert that $\widehat{h} \neq 0$, that a beneficial policy is selected, or that the upper certificates rank all policies by their realized risk.

The proof is given in Appendix C.8. The finite-class formulation makes the data-use boundary explicit: the candidate set, training budgets, shift radius, and selection rule are frozen before the audit, while a final test block is used only for reporting realized utility. The paired sensitivity is now part of the selection decision rather than only a post-hoc diagnostic.

Corollary B.1 (Fully bounded safe-selection fallback)**.** *Under the proposition's protocol, if no candidate-specific sensitivity bound is available, use*

$$U_h^{\mathrm{bd}}(\rho,\delta) := \widehat{\Delta}_h + \Gamma_{n,K}^{\mathrm{bd}}(\delta/J) + 2B\varepsilon_{\mathrm{nuis},h} + 2B\eta_{\mathrm{sem},h} + B\sqrt{\rho}.$$

The same factual fallback rule is safe on the covered valid deployment class. This fallback is fully bounded in sensitivity but still conditional on valid nuisance-law and semantic-contamination bounds. If the latter is unavailable, one may set $\eta_{\mathrm{sem},h} = 1$ to obtain a mathematically valid but generally vacuous bound; it must not be reported as an informative certificate. The term is zero only for an exact task-preserving witness.

The corollary is deliberately conservative. Neither result is an optimization theorem for the underlying transformer, and neither guarantees that a non-baseline policy will be selected.

C Proofs of the Supporting Results

C.1 Proof of Theorem B.1

Proof of Theorem B.1. We first prove two complementary adaptive bounds and then the fresh statement. The first adaptive argument treats one complete anchor–intervention cluster as one independent observation. The refinement conditions on the anchors and isolates intervention-specific adaptivity. Fresh evaluation splits into an across-anchor component and a conditionally independent within-anchor component.

Adaptive bound. Let P_D be the marginal law of a complete cluster $D = (C, Y, S_{1:K})$. The protocol gives

$$P_{D_{1:n}|W_0} = \bigotimes_{i=1}^{n} P_D,$$

because $W_0 \perp D_{1:n}$, the anchors are IID, and the intervention draws belonging to different anchors are conditionally independent. Since each loss belongs to $[0, B]$, its within-cluster average satisfies $0 \le \phi_w(D_i) \le B$.

For each i, let D_i^* have law P_D and be conditionally independent of W given W_0. We first prove the one-coordinate inequality

$$|\mathbb{E}[\phi_W(D_i) - \phi_W(D_i^*)]| \le B\sqrt{\frac{I(W; D_i \mid W_0)}{2}}. \tag{C.1}$$

Fix $W_0 = w_0$ outside a null set. Denote the conditional joint law of (W, D_i) by $P_i^{w_0}$ and the product of its marginals by

$$Q_i^{w_0} = P_{W|W_0=w_0} \otimes P_D.$$

Their relative entropy is the pointwise conditional information

$$D_{\mathrm{KL}}(P_i^{w_0} \| Q_i^{w_0}) = I(W; D_i \mid W_0 = w_0).$$

For any measurable $f \in [0, B]$ and any $\lambda > 0$, the Donsker–Varadhan variational inequality applied to $\lambda(f - \mathbb{E}_{Q_i^{w_0}} f)$ gives

$$\lambda\big(\mathbb{E}_{P_i^{w_0}} f - \mathbb{E}_{Q_i^{w_0}} f\big) \le D_{\mathrm{KL}}(P_i^{w_0} \| Q_i^{w_0}) + \log \mathbb{E}_{Q_i^{w_0}} \exp\Big\{\lambda(f - \mathbb{E}_{Q_i^{w_0}} f)\Big\}.$$

Hoeffding's lemma (Hoeffding, 1963) bounds the logarithmic moment-generating function by $\lambda^2 B^2/8$. Therefore

$$\mathbb{E}_{P_i^{w_0}} f - \mathbb{E}_{Q_i^{w_0}} f \le \frac{D_{\mathrm{KL}}(P_i^{w_0} \| Q_i^{w_0})}{\lambda} + \frac{\lambda B^2}{8}.$$

When the KL divergence is positive, choosing $\lambda = \sqrt{8 D_{\mathrm{KL}}(P_i^{w_0} \| Q_i^{w_0})}/B$ gives

$$\mathbb{E}_{P_i^{w_0}} f - \mathbb{E}_{Q_i^{w_0}} f \le B\sqrt{\frac{D_{\mathrm{KL}}(P_i^{w_0} \| Q_i^{w_0})}{2}}.$$

The same conclusion is immediate when the KL divergence is zero, by taking a limit as $\lambda \downarrow 0$. Applying the argument to $B - f$ gives the reverse inequality and hence the absolute-value form. Taking $f(w, d) = \phi_w(d)$, integrating over W_0, and applying Jensen's inequality to the square root yields (C.1):

$$\begin{aligned}|\mathbb{E}[\phi_W(D_i) - \phi_W(D_i^*)]| &\le B\mathbb{E}_{W_0}\sqrt{\frac{I(W; D_i \mid W_0 = w_0)}{2}} \\ &\le B\sqrt{\frac{I(W; D_i \mid W_0)}{2}}.\end{aligned}$$

Because D_i^* is independent of W conditional on W_0 and has the cluster marginal,

$$\mathbb{E}\left[R_{\widehat{\nu}}(W) - R_{n,K}(W)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\phi_W(D_i^*) - \phi_W(D_i)\right].$$

Equation C.1, the triangle inequality, and Cauchy–Schwarz for the finite sum give

$$|\mathbb{E}[R_{\widehat{\nu}}(W) - R_{n,K}(W)]| \le \frac{B}{n\sqrt{2}}\sum_{i=1}^{n}\sqrt{I(W; D_i \mid W_0)} \le B\sqrt{\frac{\sum_i I(W; D_i \mid W_0)}{2n}}.$$

It remains to relate the coordinate informations to the full sample. Write $D_{<i} = (D_1, \ldots, D_{i-1})$. The conditional interaction identity gives

$$\begin{aligned}I(W; D_i \mid W_0, D_{<i}) - I(W; D_i \mid W_0) = I(D_i; D_{<i} \mid W, W_0) \\ - I(D_i; D_{<i} \mid W_0).\end{aligned}$$

The final term is zero because the clusters are independent conditional on W_0, whereas the first term is nonnegative. Consequently,

$$\sum_{i=1}^{n} I(W; D_i \mid W_0) \le \sum_{i=1}^{n} I(W; D_i \mid W_0, D_{<i}) = I(W; D_{1:n} \mid W_0),$$

where the equality is the conditional chain rule. This proves the first branch of (B.1). Notice that D_i contains all K interventions attached to anchor i; no step treats them as K independent task observations.

Adaptive two-level refinement. For any w and anchor $t = (c, y)$, let

$$m_w(t) = \int \ell_w(t, s)\widehat{\nu}(ds \mid c).$$

The function m_w is measurable and belongs to $[0, B]$. Adding and subtracting the anchor average gives the pointwise identity

$$R_{\widehat{\nu}}(W) - R_{n,K}(W) = R_{\widehat{\nu}}(W) - \frac{1}{n}\sum_i m_W(T_i) + \frac{1}{n}\sum_i \left[m_W(T_i) - \frac{1}{K}\sum_j \ell_W(T_i, S_{ij}) \right].$$

Let T_i^* have the anchor marginal and be conditionally independent of W given W_0. Applying the one-coordinate DV–Hoeffding argument above with $f(w, t) = m_w(t)$, then using the conditional independence of the anchors and the same interaction-identity/chain-rule calculation, yields

$$\left| \mathbb{E}\left[R_{\widehat{\nu}}(W) - \frac{1}{n}\sum_i m_W(T_i) \right] \right| \le B\sqrt{\frac{I(W; T_{1:n} \mid W_0)}{2n}} = B\sqrt{\frac{\mathcal{I}_T}{2n}}.$$

For the second term, condition on $(W_0, T_{1:n})$. Given these variables, the nK interventions are independent, although their conditional laws may vary with their anchors. For each (i, j), let S_{ij}^* have law $\widehat{\nu}(\cdot \mid C_i)$ and be conditionally independent of W. The product-law expectation satisfies

$$\mathbb{E}[\ell_W(T_i, S_{ij}^*) \mid W_0, T_{1:n}, W] = m_W(T_i).$$

The conditional one-coordinate argument therefore gives

$$\left|\mathbb{E}[\ell_W(T_i, S_{ij}^*) - \ell_W(T_i, S_{ij})]\right| \le B\sqrt{\frac{I(W; S_{ij} \mid W_0, T_{1:n})}{2}}.$$

Summing over all coordinates and applying Cauchy–Schwarz gives

$$\left| \mathbb{E}\frac{1}{nK}\sum_{i,j}[\ell_W(T_i, S_{ij}^*) - \ell_W(T_i, S_{ij})] \right| \le B\sqrt{\frac{\sum_{i,j} I(W; S_{ij} \mid W_0, T_{1:n})}{2nK}}.$$

Order the intervention coordinates arbitrarily. Their conditional independence given $(W_0, T_{1:n})$ and the conditional interaction identity imply

$$\sum_{i,j} I(W; S_{ij} \mid W_0, T_{1:n}) \le I(W; S_{1:n,1:K} \mid W_0, T_{1:n}) = \mathcal{I}_S.$$

Thus the intervention term is at most $B\sqrt{\mathcal{I}_S/(2nK)}$. The triangle inequality proves the second branch of (B.1); taking the smaller of the two valid branches proves the displayed minimum. Finally, the conditional chain rule gives

$$I(W; D_{1:n} \mid W_0) = \mathcal{I}_T + \mathcal{I}_S.$$

The denominator nK does not alone force the intervention term to vanish, because $\mathcal{I}_S$ may grow with K; vanishing follows when $\mathcal{I}_S = o(nK)$.

Fresh evaluation. Fix an independent model $w_\perp$ and a fresh collection. Let

$$T_i' = (C_i', Y_i'), \qquad m_i = \mathbb{E}_{S|T_i'}\ell_{w_\perp}(T_i', S), \qquad \epsilon_{ij} = \ell_{w_\perp}(T_i', S_{ij}') - m_i.$$

Adding and subtracting $n^{-1}\sum_i m_i$ gives the exact identity

$$R'_{n,K}(w_\perp) - R_{\widehat{\nu}}(w_\perp) = \underbrace{\frac{1}{n}\sum_i (m_i - \mathbb{E}m_i)}_{\text{anchor variation}} + \underbrace{\frac{1}{nK}\sum_{i=1}^{n}\sum_{j=1}^{K}\epsilon_{ij}}_{\text{within-anchor variation}}. \tag{C.2}$$

The variables m_i are IID because the fresh anchors are IID, satisfy $0 \le m_i \le B$, and have variance $\sigma_T^2(w_\perp)$. Apply the two-sided Bernstein inequality to $X_i = (m_i - \mathbb{E}m_i)/n$. Here

$$|X_i| \le \frac{B}{n}, \qquad \sum_i \mathrm{Var}(X_i) = \frac{\sigma_T^2(w_\perp)}{n}.$$

With $t = \log(4/\delta)$, the resulting event has failure probability at most $\delta/2$ and satisfies

$$\left|\frac{1}{n}\sum_i (m_i - \mathbb{E}m_i)\right| \le \sqrt{\frac{2\sigma_T^2(w_\perp)t}{n}} + \frac{2Bt}{3n}. \tag{C.3}$$

We will combine this anchor-level deviation in (C.3) with the conditional sibling deviation below.

Conditional on $T'_{1:n}$ and $w_\perp$, the variables ϵ_{ij} are independent and centered. They lie in $[-m_i, B - m_i]$, hence $|\epsilon_{ij}| \le B$, and their conditional variances are

$$v_i = \mathrm{Var}_{S|T'_i}[\ell_{w_\perp}(T'_i, S)].$$

For $Y_{ij} = \epsilon_{ij}/(nK)$,

$$|Y_{ij}| \le \frac{B}{nK}, \qquad \sum_{i,j}\mathrm{Var}(Y_{ij} \mid T'_{1:n}, w_\perp) = \frac{1}{(nK)^2}\sum_{i=1}^{n} K v_i = \frac{\widehat{\sigma}^2_{S,n}(w_\perp)}{nK}.$$

Conditional Bernstein therefore gives, with conditional failure probability at most $\delta/2$,

$$\left|\frac{1}{nK}\sum_{i,j}\epsilon_{ij}\right| \le \sqrt{\frac{2\widehat{\sigma}^2_{S,n}(w_\perp)t}{nK}} + \frac{2Bt}{3nK}. \tag{C.4}$$

Since the conditional failure bound holds almost surely in the anchors, its unconditional failure probability is also at most $\delta/2$. A union bound, equations C.2–C.4, and the triangle inequality prove (B.2).

For completeness, the same decomposition also yields the simpler bounded-range version. Hoeffding's inequality applied to the IID variables $m_i \in [0, B]$ and, conditionally, to the residuals $\epsilon_{ij} \in [-m_i, B - m_i]$, whose interval length is B, gives

$$|R'_{n,K}(w_\perp) - R_{\widehat{\nu}}(w_\perp)| \le B\sqrt{\frac{\log(4/\delta)}{2n}} + B\sqrt{\frac{\log(4/\delta)}{2nK}}$$

with probability at least $1 - \delta$. The variance-sensitive theorem follows from Bernstein rather than this corollary. Independence of $w_\perp$ from the fresh collection is essential in both conditional arguments.

Necessity of the anchor term. Choose $C_i \sim \mathrm{Bernoulli}(1/2)$ and any conditional intervention kernel, and let $\ell(C_i, S, Y) = BC_i$. Then

$$\phi_w(D_i) = \frac{1}{K}\sum_{j=1}^{K} BC_i = BC_i,$$

so the instance satisfies the bounded-loss cluster protocol for every K and

$$R_{n,K} = \frac{B}{n}\sum_i C_i, \qquad R_{\widehat{\nu}} = \frac{B}{2}, \qquad \mathrm{Var}(R_{n,K}) = \frac{B^2}{4n}.$$

The events $\{C_1 = \cdots = C_n = 0\}$ and $\{C_1 = \cdots = C_n = 1\}$ are disjoint, each has probability 2^{-n}, and on either event $|R_{n,K} - R_{\widehat{\nu}}| = B/2$. Thus this deviation occurs with probability 2^{1-n}, independently of K. For fixed n, any constant c independent of K, and any $\delta < 2^{1-n}$, choose $K > 4c^2/n$. Then $cB/\sqrt{nK} < B/2$, while the probability of exceeding that threshold is at least $2^{1-n} > \delta$. Hence no uniform adaptive high-probability theorem can have $1/\sqrt{nK}$ as its only sampling scale. This does not contradict the fresh residual term: that term is conditional on fresh anchors and accompanies a separate $n^{-1/2}$ anchor term. □

C.2 Proof of Theorem B.2

Proof of Theorem B.2. Fix w and suppress it from the notation when no ambiguity arises. Because $L_{\mathrm{CF}}(w) < \infty$, the conditional pairwise KL integral is finite for μ-almost every c; fix such a c below.

Reference mixture and pairwise KL. For fixed c, write

$$P_s = P_w(\cdot \mid c, s), \qquad \bar{P} = P_w^{\nu}(\cdot \mid c) = \int P_s \nu(ds \mid c).$$

Because P_w and ν are measurable probability kernels on standard Borel spaces, integration of one kernel against the other makes $c \mapsto \bar{P}$ a measurable probability kernel. Standard Borel structure provides the measurable-kernel and Radon–Nikodym machinery below, but it does not by itself imply $P_s \ll \bar{P}$ for almost every s. The latter domination follows here from the finite pairwise-KL assumption.

Indeed, finiteness of

$$\int\int D_{\mathrm{KL}}(P_s \| P_{s'}) \nu(ds \mid c) \nu(ds' \mid c)$$

implies by Fubini that, for $\nu(\cdot \mid c)$-almost every s, both $P_s \ll P_{s'}$ and $P_{s'} \ll P_s$ hold for $\nu(\cdot \mid c)$-almost every s'. For such an s, if $\bar{P}(A) = 0$, then $P_{s'}(A) = 0$ for almost every s', and choosing an s' for which $P_s \ll P_{s'}$ gives $P_s(A) = 0$. Conversely, if $P_s(A) = 0$, then $P_{s'}(A) = 0$ for almost every s' by $P_{s'} \ll P_s$, and hence $\bar{P}(A) = 0$. Thus, for almost every s,

$$P_s \ll \bar{P} \quad \text{and} \quad \bar{P} \ll P_s.$$

This is the point at which the finite pairwise-KL assumption is used; a mixture alone would not suffice (for example, $P_s = \delta_s$ mixed over a continuous uniform law).

Before proving the exact identity, note the useful domination

$$D_{\mathrm{KL}}(P_s \| \bar{P}) \le \int D_{\mathrm{KL}}(P_s \| P_{s'}) \nu(ds' \mid c). \tag{C.5}$$

For fixed P_s, relative entropy is convex in its second argument. Jensen's inequality and $\bar{P} = \int P_{s'} \nu(ds' \mid c)$ give (C.5); no symmetry of KL is used.

Convexity in the first argument gives the reverse-direction domination

$$D_{\mathrm{KL}}(\bar{P} \| P_s) \le \int D_{\mathrm{KL}}(P_{s'} \| P_s) \nu(ds' \mid c). \tag{C.6}$$

Integrating (C.5) over s and (C.6) over s, and then integrating over c, shows directly that

$$I_\nu(w) \le L_{\mathrm{CF}}(w), \qquad J_{\mathrm{rev}}(w) \le L_{\mathrm{CF}}(w).$$

In particular, both one-sided quantities are finite. Hence $\bar{P} \ll P_s$ for $\nu(\cdot \mid c)$-almost every s, while the finite-pairwise-KL argument above gives $P_s \ll \bar{P}$ for almost every s.

For the exact decomposition, the Radon–Nikodym theorem for dominated kernels on standard Borel spaces supplies a jointly measurable version $p_s = dP_s/d\bar{P}$. The preceding domination relations imply $p_s > 0$ $\bar{P}$-almost everywhere for almost every s, and the mixture identity gives

$$\int p_s(z) \nu(ds \mid c) = 1 \quad \text{for } \bar{P}\text{-almost every } z.$$

Using $\bar{P}$ as the common reference measure and applying Fubini under the finite pairwise-KL condition,

$$\begin{aligned}
&\int\int D_{\mathrm{KL}}(P_s \| P_{s'}) \nu(ds \mid c) \nu(ds' \mid c) \\
&= \int\int\int p_s \log \frac{p_s}{p_{s'}} \, d\bar{P} \, \nu(ds \mid c) \nu(ds' \mid c) \\
&= \int\int p_s \log p_s \, d\bar{P} \, \nu(ds \mid c) - \int\int\int p_s \log p_{s'} \, d\bar{P} \, \nu(ds \mid c) \nu(ds' \mid c) \\
&= \int D_{\mathrm{KL}}(P_s \| \bar{P}) \nu(ds \mid c) + \int D_{\mathrm{KL}}(\bar{P} \| P_s) \nu(ds \mid c).
\end{aligned}$$

The last equality uses $\int p_s \nu(ds \mid c) = 1$ in the second term. The preceding finite dominations justify all terms. Integrating over $c \sim \mu$ proves

$$L_{\mathrm{CF}}(w) = I_\nu(w) + J_{\mathrm{rev}}(w).$$

This establishes the identity and the two one-sided inequalities stated in the theorem without any separate finiteness assumption on I_ν or J_{rev}.

Covered-shift bound. Define the conditional loss envelope

$$h_{w,c}(z) = \int q(dy \mid c) \int M_w(da \mid z)\ell(a, y).$$

It is measurable and belongs to $[0, B]$, because it is an iterated expectation of a loss in $[0, B]$. By the definitions of P_w and r_w,

$$r_w(c, s) = \int h_{w,c}(z) P_s(dz).$$

Define

$$\bar{r}_w(c) = \int h_{w,c}(z) \bar{P}(dz).$$

The mixture identity and Fubini's theorem give

$$\bar{r}_w(c) = \int r_w(c, s)\nu(ds \mid c), \qquad R_\nu(w) = \int \bar{r}_w(c)\mu(dc).$$

For every relevant (c, s), the bounded-function total-variation inequality and Pinsker's inequality imply

$$\begin{aligned} |r_w(c, s) - \bar{r}_w(c)| &\le B \ \mathrm{TV}(P_s, \bar{P}) \\ &\le B\sqrt{\frac{D_{\mathrm{KL}}(P_s \| \bar{P})}{2}}. \end{aligned} \tag{C.7}$$

If a conditional KL is infinite, the inequality remains true in the extended sense; $I_\nu(w) \le L_{\mathrm{CF}}(w) < \infty$ excludes an infinite population contribution below.

Assumption A.2 supplies $a_e(c, s) = d\pi_e(\cdot \mid c)/d\nu(\cdot \mid c)(s)$, with $\int a_e(c, s)\nu(ds \mid c) = 1$. Therefore

$$\begin{aligned} R_e(w) - R_\nu(w) &= \int \mu(dc) \int \nu(ds \mid c)(a_e(c, s) - 1) r_w(c, s) \\ &= \int \mu(dc) \int \nu(ds \mid c)(a_e(c, s) - 1)[r_w(c, s) - \bar{r}_w(c)]. \end{aligned}$$

Conditional Cauchy–Schwarz applied directly to this centered identity gives

$$\begin{aligned} |R_e(w) - R_\nu(w)| &\le \int \sqrt{\chi^2_c V_c(w)}\, \mu(dc) \\ &\le \sqrt{\left(\int \chi^2_c d\mu\right)\left(\int V_c(w) d\mu\right)} = \sqrt{\bar{\chi}^2_e V_\nu(w)}. \end{aligned}$$

Let $k_w(c, s) = D_{\mathrm{KL}}(P_s \| \bar{P})$. Using (C.7) and applying Cauchy–Schwarz conditionally on c,

$$\begin{aligned} |R_e(w) - R_\nu(w)| &\le \frac{B}{\sqrt{2}} \int \mu(dc) \int \nu(ds \mid c)|a_e(c, s) - 1| \sqrt{k_w(c, s)} \\ &\le \frac{B}{\sqrt{2}} \int \sqrt{\left(\int (a_e - 1)^2 d\nu\right)\left(\int k_w d\nu\right)}\, \mu(dc) \\ &= \frac{B}{\sqrt{2}} \int \sqrt{\chi^2_c I_c(w)}\, \mu(dc). \end{aligned}$$

A second Cauchy–Schwarz inequality, now over $c \sim \mu$, gives

$$\int \sqrt{\chi^2_c I_c(w)}\, \mu(dc) \le \sqrt{\left(\int \chi^2_c d\mu\right)\left(\int I_c(w) d\mu\right)} = \sqrt{\bar{\chi}^2_e I_\nu(w)}.$$

Combining this with $I_\nu(w) \le L_{\mathrm{CF}}(w)$ proves every transfer inequality in the theorem. No independence between C and S is used.

Generator mismatch and conclusion. For each c, the measurable function $s \mapsto r_w(c,s)$ belongs to $[0,B]$. The bounded-function total-variation inequality gives

$$\left|\int r_w(c,s)\widehat{\nu}(ds \mid c) - \int r_w(c,s)\nu(ds \mid c)\right| \le B\ \mathrm{TV}(\widehat{\nu}(\cdot \mid c), \nu(\cdot \mid c)).$$

Both integrals use the same response kernel by Assumption A.1. The triangle inequality, integration over c, and Assumption A.3 prove

$$|R_{\widehat{\nu}}(w) - R_\nu(w)| \le B\varepsilon_{\mathrm{gen}}.$$

Taking the one-sided directions of this inequality and of the covered-shift bound gives

$$\begin{aligned} R_e(w) &\le R_\nu(w) + B\sqrt{\frac{\bar{\chi}_e^2 L_{\mathrm{CF}}(w)}{2}} \\ &\le R_{\widehat{\nu}}(w) + B\varepsilon_{\mathrm{gen}} + B\sqrt{\frac{\bar{\chi}_e^2 L_{\mathrm{CF}}(w)}{2}} = U_{\mathrm{CF}}(w). \end{aligned}$$

If $0 \le a_e(c,s) \le \rho_e$, then $a_e^2 \le \rho_e a_e$ and

$$\chi_c^2 = \int (a_e - 1)^2 d\nu = \int a_e^2 d\nu - 2\int a_e d\nu + 1 = \int a_e^2 d\nu - 1 \le \rho_e - 1.$$

Integrating over c proves $\bar{\chi}_e^2 \le \rho_e - 1$. If $\pi_e = \nu$, then $a_e = 1$ almost surely and the centered risk difference is zero. If $I_\nu(w) = 0$, then the global centered bound is zero and again $R_e(w) = R_\nu(w)$. This proves the equality cases and completes the theorem. □

C.3 Proof of Proposition B.1

Proof of Proposition B.1. **Gaussian-smoothed representation.** Conditional on $(C,S) = (c,s)$,

$$P_w(\cdot \mid c,s) = N(f_w(\mathsf{x}(c,s)), \sigma^2 I_d).$$

For Gaussian laws with means m, m' and common positive-definite covariance Σ,

$$D_{\mathrm{KL}}(N(m,\Sigma)\|N(m',\Sigma)) = \frac{1}{2}(m-m')^\top \Sigma^{-1}(m-m').$$

Taking $\Sigma = \sigma^2 I_d$ gives

$$D_{\mathrm{KL}}(P_w(\cdot \mid c,s)\|P_w(\cdot \mid c,s')) = \frac{\|f_w(\mathsf{x}(c,s)) - f_w(\mathsf{x}(c,s'))\|_2^2}{2\sigma^2}.$$

Integrating this identity over $c \sim \mu$ and two conditionally independent draws $s, s' \sim \nu(\cdot \mid c)$ gives the displayed expression for $L_{\mathrm{CF}}(w)$. The Gaussian noise is part of the assumed stochastic channel; the calculation does not assert equivalence with an arbitrary deterministic network.

A closed-form transfer calculation. Let $C = c_0$, $S \sim \mathrm{Unif}[-1,1]$, $X = S$, $\widehat{\nu} = \nu$, and $Z = \omega S + \xi$, where $\xi \sim N(0,\sigma^2)$ is independent. The Gaussian formula gives $L_{\mathrm{CF}}(\omega) = \omega^2/(3\sigma^2)$ because $\mathbb{E}(S-S')^2 = 2\mathrm{Var}(S) = 2/3$ for independent uniform draws on $[-1,1]$. For completeness, its risk integral can also be evaluated explicitly. With $a = \omega/\sigma$ and $I(a) = \frac{1}{2}\int_{-1}^{1} s\Phi(-as)\,ds$, integration by parts gives, for $a \ne 0$,

$$\begin{aligned} 2I(a) &= \left[\frac{s^2}{2}\Phi(-as)\right]_{-1}^{1} + \frac{a}{2}\int_{-1}^{1} s^2\varphi(as)\,ds \\ &= \frac{1}{2} - \Phi(a) + \frac{2\Phi(a) - 1 - 2a\varphi(a)}{2a^2}. \end{aligned}$$

Rearranging yields the closed form, and dominated convergence gives $I(0) = 0$. For the head $A = \mathbf{1}\{Z \geq 0\}$, label $Y = 1$, and zero–one loss, take the test density ratio $1 + \beta s$, $|\beta| \leq 1$. Then $\chi^2 = \beta^2/3$, and centering under the uniform reference law gives $R_e - R_\nu = \beta \mathbb{E}_\nu[S\Phi(-aS)] = \beta I(a)$. For $(\omega, \sigma, \beta) = (0.6, 0.5, 0.5)$, this gives $L_{\mathrm{CF}} = 0.48$, $\chi^2 = 1/12$, $R_\nu = 0.5$, $R_e \approx 0.4301$, and the certificate $\sqrt{\chi^2 L_{\mathrm{CF}}/2} \approx 0.1414$, compared with an actual gap of 0.0699.

Predictive variant. Define the predictive action kernel

$$P_w^A(da \mid c, s) = \int M_w(da \mid z) P_w(dz \mid c, s)$$

and its reference mixture $P_w^{A,\nu}(da \mid c) = \int P_w^A(da \mid c, s)\nu(ds \mid c)$. Define $I_\nu^A, L_{\mathrm{CF}}^A, J_{\mathrm{rev}}^A$ by replacing P_w with P_w^A in the definitions from Section A; no common dominating density is needed because relative entropy is defined directly between probability measures on the standard Borel action space.

The action law is the image of the representation law through the same Markov kernel M_w. The data-processing inequality applied to each nuisance pair and to each representation/reference-mixture pair gives

$$L_{\mathrm{CF}}^A(w) \leq L_{\mathrm{CF}}(w), \qquad I_\nu^A(w) \leq I_\nu(w).$$

In particular, $L_{\mathrm{CF}}^A < \infty$. Repeating the finite-pairwise-KL decomposition from the proof of Theorem B.2 gives

$$L_{\mathrm{CF}}^A = I_\nu^A + J_{\mathrm{rev}}^A.$$

The action-level conditional loss envelope is

$$\widetilde{h}_{w,c}(a) = \int q(dy \mid c)\ell(a, y), \qquad 0 \leq \widetilde{h}_{w,c}(a) \leq B.$$

Writing $P_s^A = P_w^A(\cdot \mid c, s)$ and $\bar{P}^A = P_w^{A,\nu}(\cdot \mid c)$, define

$$I_c^A(w) = \int D_{\mathrm{KL}}(P_s^A \| \bar{P}^A)\nu(ds \mid c).$$

Pinsker's inequality yields

$$|r_w(c, s) - \bar{r}_w(c)| \leq B\sqrt{\frac{D_{\mathrm{KL}}(P_s^A \| \bar{P}^A)}{2}}.$$

Squaring and integrating over $s \sim \nu(\cdot \mid c)$ gives

$$V_c(w) \leq \frac{B^2}{2} I_c^A(w), \qquad V_\nu(w) \leq \frac{B^2}{2} I_\nu^A(w).$$

Combining this inequality with the centered risk-variance bound in Theorem B.2, the action-level decomposition, and data processing proves

$$|R_e - R_\nu| \leq \sqrt{\bar{\chi}_e^2 V_\nu} \leq B\sqrt{\frac{\bar{\chi}_e^2 I_\nu^A}{2}} \leq B\sqrt{\frac{\bar{\chi}_e^2 L_{\mathrm{CF}}^A}{2}} \leq B\sqrt{\frac{\bar{\chi}_e^2 L_{\mathrm{CF}}}{2}}.$$

Thus action consistency is a head-aware certificate, while representation consistency remains a valid, potentially looser, upstream certificate.

Wasserstein variant. Let $(\mathcal{S}, d)$ be Polish with its Borel σ-field. Suppose that, for μ-almost every c, both conditional nuisance laws belong to $\mathcal{P}_1(\mathcal{S})$ and $s \mapsto r_w(c, s)$ is $L(c)$-Lipschitz. Assume the right-hand side in the displayed expectation below is integrable. Kantorovich–Rubinstein duality, applied for each such c, gives

$$\left| \int r_w(c, s)\widehat{\nu}(ds \mid c) - \int r_w(c, s)\nu(ds \mid c) \right| \leq L(c) W_1(\widehat{\nu}(\cdot \mid c), \nu(\cdot \mid c)).$$

The kernels make the conditional integrals measurable; assuming $c \mapsto L(c)W_1(\widehat{\nu}_c, \nu_c)$ is integrable, integrate over $c \sim \mu$ to obtain

$$|R_{\widehat{\nu}}(w) - R_\nu(w)| \leq \mathbb{E}_C[L(C)W_1(\widehat{\nu}(\cdot \mid C), \nu(\cdot \mid C))].$$

This is an alternative generator-mismatch interface; it does not remove the separate coverage requirement for transfer to the test environment. □

C.4 Proof of Proposition B.2

Proof of Proposition B.2. **Exact attainment and shift alignment.** We first specify a task-preserving witness. Let C, S be independent uniform signs under the reference law, let $Y = C$, and use the stochastic representation $Z = (C, \gamma S) + \xi$, with $\xi \sim \mathcal{N}(0, \sigma^2 I_2)$ independent and $\sigma > 0$. The head predicts $\operatorname{sign}(Z_1 + \lambda Z_2)$, where $\lambda > 0$, under zero–one loss. A covered test law has $\pi_\beta(s \mid c) = (1 + \beta cs)/2$, $|\beta| \le 1$; the label law is unchanged. Write

$$e_\pm = \Phi\left(-\frac{1 \pm \lambda\gamma}{\sigma\sqrt{1+\lambda^2}}\right).$$

The KL divergence between the two equal-covariance representation laws is $2\gamma^2/\sigma^2$ when the nuisance signs differ and zero otherwise, so the ordered independent-pair average is $L_{\mathrm{CF}} = \gamma^2/\sigma^2$. For each fixed c, uniform reference sampling gives the two conditional risks e_+ and e_- with equal probability. Their conditional mean is $(e_+ + e_-)/2$ and their conditional variance is $(e_+ - e_-)^2/4$. Hence $V_c = V_\nu = (e_+ - e_-)^2/4$ for every c. The test density ratio is $1 + \beta cs$, so $\chi^2_c = \bar{\chi}^2_e = \beta^2$ and

$$|R_e - R_\nu| = \frac{|\beta|}{2}|e_+ - e_-| = \int \sqrt{\chi^2_c V_c}\,\mu(dc) = \sqrt{\bar{\chi}^2_e V_\nu}.$$

When $\beta \neq 0$ and $\lambda\gamma \neq 0$, strict monotonicity of Φ makes this a nonzero equality. Thus the coefficient one in the head-aware bound cannot be reduced uniformly over the theorem's model class.

More generally, write $\bar{r}(c) = \int r_w(c,s)\nu(ds \mid c)$, $h(c,s) = r_w(c,s) - \bar{r}(c)$, and $u(c,s) = a_e(c,s) - 1$. Conditional centering and normalization give

$$R_e - R_\nu = \mathbb{E}_{\mu\nu}[uh], \qquad \mathbb{E}_{\mu\nu}u^2 = \bar{\chi}^2_e, \qquad \mathbb{E}_{\mu\nu}h^2 = V_\nu.$$

For positive variances, the ratio of the gap to the global head-aware bound is therefore the absolute normalized inner product of u and h. Equality holds precisely when $u = th$ almost surely for a nonzero constant t, by the equality condition in Cauchy–Schwarz. This alignment is realizable: assume $B > 0$, set $\widehat{\nu} = \nu$, keep the model and invariant task fixed, and define

$$\pi_t(ds \mid c) = [1 + th(c,s)]\nu(ds \mid c), \qquad 0 \le t \le 1/B.$$

Measurability follows from that of r_w and the kernels. Since $|h| \le B$ and $\int h\, d\nu(\cdot \mid c) = 0$, this is a nonnegative, normalized, covered kernel. Moreover,

$$\bar{\chi}^2_t = t^2 V_\nu, \qquad R_{\pi_t} - R_\nu = tV_\nu = \sqrt{\bar{\chi}^2_t V_\nu}.$$

The construction retains finite L_{CF} whenever the starting model has finite L_{CF}, because the reference law and model are unchanged. If $V_\nu = 0$ or $t = 0$, both sides vanish. The construction uses the model's conditional risk to choose the shift; it establishes attainability, not typical tightness under a model-independent environment.

Local sharpness. Use the task-preserving representation witness just constructed and set

$$a = \frac{1}{\sigma\sqrt{1+\lambda^2}}, \qquad d = \frac{\lambda}{\sigma\sqrt{1+\lambda^2}}.$$

Multiplying the classification margin by C shows that its conditional error depends on $T = CS$ only. Thus, for $t \in \{-1, +1\}$,

$$e_+(\gamma) = \Phi(-a - d\gamma), \qquad e_-(\gamma) = \Phi(-a + d\gamma),$$

and the reference/test coupling in the witness gives

$$R_e - R_\nu = \frac{\beta}{2}\big(e_+(\gamma) - e_-(\gamma)\big), \qquad \chi^2(\pi_\beta \| \nu) = \beta^2, \qquad L_{\mathrm{CF}}(\gamma) = \frac{\gamma^2}{\sigma^2}.$$

The function Φ is differentiable with derivative φ, so

$$\frac{d}{d\gamma}\big(e_+(\gamma) - e_-(\gamma)\big)\Big|_{\gamma=0} = -d\varphi(-a) - d\varphi(-a) = -2d\varphi(a).$$

Because $e_+(0) = e_-(0)$, it follows that

$$\lim_{\gamma\to 0,\ \gamma\neq 0} \frac{|R_e - R_\nu|}{\sqrt{\chi^2(\pi_\beta\|\nu)L_{\mathrm{CF}}(\gamma)}} = \frac{|\beta| d\varphi(a)}{|\beta|/\sigma} = \frac{\lambda}{\sqrt{1+\lambda^2}}\varphi(a) = c_{\sigma,\lambda} > 0.$$

The positive limit implies that, for some $\gamma_0 > 0$ and every $0 < |\gamma| < \gamma_0$, the ratio is at least $c_{\sigma,\lambda}/2$. Hence the transfer gap cannot have a uniformly smaller local order than the square-root product in this class. If $\lambda = 0$, then $d = 0$ and $e_+(\gamma) = e_-(\gamma)$ for every γ, so the realized gap is exactly zero.

Support escape. Let

$$P_{\mathrm{ref}}(dc, ds) = \mu(dc)\nu(ds \mid c), \qquad P_e(dc, ds) = \mu(dc)\pi_e(ds \mid c).$$

If $P_e \not\ll P_{\mathrm{ref}}$, choose a measurable set E such that $P_{\mathrm{ref}}(E) = 0$ and $P_e(E) = p > 0$. Consider the allowed instance with $X = (C, S)$, $\mathsf{x}(c, s) = (c, s)$, constant response $Y = 0$, binary action space $\{0, 1\}$, and loss $\ell(a, 0) = Ba$. Let the shared deterministic representation be

$$Z = \mathbf{1}_E(C, S).$$

Disintegration of $P_{\mathrm{ref}}(E) = 0$ implies $\nu(E_c \mid c) = 0$ for μ-almost every c, where $E_c = \{s : (c, s) \in E\}$. Therefore, under the reference intervention, $Z = 0$ almost surely for almost every anchor c, and any two conditionally sampled reference representations are both the point mass at zero.

Define w_0 to output $A = 0$ for every Z, and let w_1 output $A = Z$. Both models use the same representation kernel. It follows directly from the loss and the reference support that

$$R_\nu(w_0) = R_\nu(w_1) = 0, \qquad I_\nu(w_0) = I_\nu(w_1) = L_{\mathrm{CF}}(w_0) = L_{\mathrm{CF}}(w_1) = 0.$$

Choose $\widehat{\nu} = \nu$, so that $\varepsilon_{\mathrm{gen}} = 0$ and the invariant-task interface is satisfied. Under the test law, however,

$$R_e(w_0) = 0, \qquad R_e(w_1) = B\int \mathbf{1}_E(c, s)P_e(dc, ds) = Bp.$$

Thus the two models have identical reference risk, representation consistency, pairwise counterfactual penalty, and generator mismatch, while their test risks differ by $Bp > 0$. Any reference-only transfer penalty that vanishes at these four zero quantities would assign no correction to either model and therefore cannot uniformly control the escaped test mass. This proves the coverage boundary and completes the proposition. □

C.5 Proof of Theorem B.3

Proof of Theorem B.3. Fix the learned pair (w_F, w_C) and condition on the training sigma-field. All expectations below are therefore over fresh (C, S) draws. Put $h(c, s) = g(c, s) - \bar{g}(c)$ and $a(c, s) = a_\pi(c, s)$. By definition of $\bar{g}$,

$$\int h(c, s)\,\nu(ds \mid c) = 0 \quad \text{for } \mu\text{-almost every } c.$$

Since $\pi(\cdot \mid c) \ll \nu(\cdot \mid c)$, the change-of-measure identity gives

$$\begin{aligned}\Delta_\pi - \Delta_\nu &= \int \mu(dc) \int \nu(ds \mid c)(a(c, s) - 1)g(c, s) \\ &= \int \mu(dc) \int \nu(ds \mid c)(a(c, s) - 1)h(c, s).\end{aligned}$$

The second equality uses $\int (a - 1)\, d\nu = \int d\pi - \int d\nu = 0$ conditionally on c. This proves (B.3).

For fixed c, conditional Cauchy–Schwarz yields

$$\left|\int (a - 1)h\, d\nu(\cdot \mid c)\right| \le \left(\int (a - 1)^2 d\nu(\cdot \mid c)\right)^{1/2} \left(\int h^2 d\nu(\cdot \mid c)\right)^{1/2} = \sqrt{\chi^2_c V^\Delta_c}.$$

Integrating proves the first inequality in (B.4). Applying Cauchy–Schwarz once more to the two functions of c gives

$$\int \sqrt{\chi^2_c V^\Delta_c}\, d\mu \le \left(\int \chi^2_c d\mu\right)^{1/2} \left(\int V^\Delta_c d\mu\right)^{1/2} = \sqrt{\bar{\chi}^2_\pi V^\Delta_\nu},$$

which completes the bound. The definition of $\mathcal{E}_\rho(\nu)$ immediately gives (B.5); the strict endpoint conditions give the benefit and harm implications.

It remains to compare the paired and separate radii. Define

$$h_C(c,s) = r_{w_C}(c,s) - \int r_{w_C}(c,s')\nu(ds' \mid c), \qquad h_F(c,s) = r_{w_F}(c,s) - \int r_{w_F}(c,s')\nu(ds' \mid c).$$

Then $g - \bar{g} = h_C - h_F$, while

$$\|h_C\|_{L^2(\mu\otimes\nu)} = \sqrt{V_\nu(w_C)}, \qquad \|h_F\|_{L^2(\mu\otimes\nu)} = \sqrt{V_\nu(w_F)}.$$

The L^2 triangle inequality gives

$$\sqrt{V_\nu^\Delta} = \|h_C - h_F\|_2 \le \|h_C\|_2 + \|h_F\|_2 = \sqrt{V_\nu(w_C)} + \sqrt{V_\nu(w_F)},$$

which is (B.6). Equality holds precisely when the two centered response functions satisfy the equality condition for the triangle inequality; whenever that condition fails, the inequality is strict.

For sharpness, let C be constant, let $S \in \{-1, +1\}$ have the uniform reference law, and set

$$g(S) = m + \alpha S, \qquad \pi_\delta(S = s) = \frac{1 + \delta s}{2}, \quad |\delta| \le 1.$$

Then $\bar{g} = m$, $V_\nu^\Delta = \alpha^2$, and $a_\delta(s) = 1 + \delta s$. Direct calculation gives

$$\Delta_{\pi_\delta} - \Delta_\nu = \mathbb{E}_\nu[(a_\delta - 1)(g - \bar{g})] = \alpha\delta, \qquad \sqrt{\bar{\chi}^2_{\pi_\delta} V_\nu^\Delta} = |\alpha\delta|.$$

Thus the coefficient one cannot be decreased. This witness is realizable by two bounded losses whenever $|m| + |\alpha| \le B$: take $r_{w_C} = (B + g)/2$ and $r_{w_F} = (B - g)/2$. The sharpness construction is therefore compatible with the bounded-loss setting. □

C.6 Proof of Proposition B.3

Proof of Proposition B.3. Let $h(c,s) = g(c,s) - \bar{g}(c)$. For $V > 0$ define

$$a_\pm(c,s) := 1 \pm \sqrt{\frac{\rho}{V}} h(c,s).$$

The conditional centering identity gives

$$\int a_\pm(c,s)\, \nu(ds \mid c) = 1 \quad \text{for } \mu\text{-almost every } c.$$

The condition $\rho \le V/M^2$ implies $a_\pm \ge 0$ almost everywhere, so $a_\pm$ defines a covered deployment law with task marginal μ. Moreover,

$$\bar{\chi}^2_{\pi_\pm} = \mathbb{E}_{\mu\otimes\nu}[(a_\pm - 1)^2] = \frac{\rho}{V}\mathbb{E}_{\mu\otimes\nu}[h^2] = \rho.$$

The transport identity from Theorem B.3 then gives

$$\Delta_{\pi_\pm} - \Delta_\nu = \mathbb{E}_{\mu\otimes\nu}[(a_\pm - 1)h] = \pm\sqrt{\rho V}.$$

These two laws attain the upper and lower endpoints of the interval already proved in Theorem B.3, establishing (B.7). If $V = 0$, then $h = 0$ almost surely and every covered law has $\Delta_\pi = \Delta_\nu$. □

C.7 Proof of Proposition B.4

Proof of Proposition B.4. We first state explicitly the conditioning and finite-sample calibration event used by the proposition. Let $\mathcal{F}_{\mathrm{tr}}$ be the training sigma-field and let $\mathcal{F}_{\mathrm{cal}}$ contain the independent calibration block used to construct the variance and oscillation bounds. We condition on $\mathcal{F}_{\mathrm{tr}} \vee \mathcal{F}_{\mathrm{cal}}$ when applying concentration to the fresh evaluation block; the calibration event below is then included in the final joint probability statement. Let

$$\mathcal{A} = \left\{ \widehat{\sigma}_A^2 \ge \sigma_A^2, \quad \widehat{\sigma}_S^2 \ge \frac{1}{n}\sum_{i=1}^n v(T_i) \right\}.$$

The second inequality bounds the realized conditional sibling variance of the fresh evaluation anchors. It may be obtained from an independent calibration pool together with a concentration bound for the anchor-level variables $v(T_i)$; it is stronger and more precise than silently replacing this realized quantity by its expectation. By assumption, $\Pr(\mathcal{A}) \geq 1 - \delta/2$.

Let $\bar{Z}_i = K^{-1} \sum_{j=1}^{K} Z_{ij}$ and $m_i = \mathbb{E}[Z_{ij} \mid T_i]$. The exact decomposition is

$$\widehat{\Delta} - \Delta_{\widehat{\nu}} = \underbrace{\frac{1}{n} \sum_{i=1}^{n} (m_i - \mathbb{E} m_i)}_{A_n} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} (\bar{Z}_i - m_i)}_{S_{n,K}}. \tag{C.8}$$

The A_n summands are independent, centered, have variance σ_A^2, and have absolute value at most $2B$. We use the conservative range-form Bernstein inversion for this centered paired difference; explicitly, the resulting radius is

$$\Pr\left\{ |A_n| > \frac{2Bt}{3n} + \sqrt{\frac{2\sigma_A^2 t}{n} + \frac{4B^2 t^2}{9n^2}} \right\} \leq 2e^{-t}.$$

Conditionally on $T_{1:n}$ and $\mathcal{F}_{\mathrm{tr}} \vee \mathcal{F}_{\mathrm{cal}}$, the variables $Z_{ij} - m_i$ are independent and centered, with absolute value at most $2B$. Their conditional total variance after averaging is

$$\mathrm{Var}(S_{n,K} \mid T_{1:n}, \mathcal{F}_{\mathrm{tr}} \vee \mathcal{F}_{\mathrm{cal}}) = \frac{1}{n^2 K} \sum_{i=1}^{n} v(T_i) \leq \frac{\widehat{\sigma}_S^2}{nK} \quad \text{on } \mathcal{A},$$

and the summands are bounded in absolute value by $2B/(nK)$. Conditional Bernstein therefore gives, on $\mathcal{A}$,

$$\Pr\left\{ |S_{n,K}| > \frac{2Bt}{3nK} + \sqrt{\frac{2\widehat{\sigma}_S^2 t}{nK} + \frac{4B^2 t^2}{9n^2 K^2}} \,\middle|\, T_{1:n}, \mathcal{F}_{\mathrm{tr}} \vee \mathcal{F}_{\mathrm{cal}} \right\} \leq 2e^{-t}.$$

Taking $t = \log(8/\delta)$, each two-sided Bernstein event has failure probability at most $\delta/4$. Together with $\Pr(\mathcal{A}^c) \leq \delta/2$, a union bound gives total failure probability at most δ. This proves the first assertion in (B.11) and the radius in (B.8).

For the variance identity, independence across anchors gives zero covariance between distinct cluster averages. Within a cluster, conditional independence gives

$$\mathrm{Var}(\bar{Z}_i) = \mathrm{Var}(m_i) + \mathbb{E}\left[\frac{v(T_i)}{K} \right] = \sigma_A^2 + \frac{\sigma_S^2}{K}.$$

Dividing by n proves the second assertion of (B.11).

For the combined radius, the cluster means $\bar{Z}_i$ are IID across anchors, have mean $\Delta_{\widehat{\nu}}$, variance $\sigma_A^2 + \sigma_S^2/K$, and satisfy $|\bar{Z}_i| \leq B$. Applying the same conservative centered-pair inversion directly to $\bar{Z}_i$ gives

$$\Pr\left\{ \left| \widehat{\Delta} - \Delta_{\widehat{\nu}} \right| > \frac{2Bt}{3n} + \sqrt{\frac{2\widehat{\sigma}_{\mathrm{cl}}^2 t}{n} + \frac{4B^2 t^2}{9n^2}} \right\} \leq 2e^{-t}$$

on histories satisfying the independent calibration event $\widehat{\sigma}_{\mathrm{cl}}^2 \geq \sigma_A^2 + \sigma_S^2/K$; the probability display is conditional on that fixed training/calibration history, and its failure probability is added to the calibration failure probability. This proves (B.9). Finally, Hoeffding's inequality applied to the same bounded IID cluster means gives

$$\Pr\left\{ \left| \widehat{\Delta} - \Delta_{\widehat{\nu}} \right| > B\sqrt{\frac{2\log(2/\delta)}{n}} \right\} \leq \delta,$$

which proves (B.10). These alternatives require their own calibration events when their variance estimates are data-derived; they may be compared only on a joint event that includes all selected bounds.

Finally, boundedness of g gives

$$|\Delta_\nu - \Delta_{\widehat{\nu}}| \le \int \operatorname{osc}(g(c,\cdot))\,\mathrm{TV}(\nu(\cdot \mid c), \widehat{\nu}(\cdot \mid c))\,\mu(dc) \le \overline{O}_g \varepsilon_{\mathrm{gen}},$$

where $\operatorname{osc}(g) = \sup g - \inf g \le 2B$. Theorem B.3 and $\overline{V}_\nu^\Delta \ge V_\nu^\Delta$ then imply

$$\Delta_\pi \le \Delta_\nu + \sqrt{\rho V_\nu^\Delta} \le \widehat{\Delta} + \Gamma_{n,K} + \overline{O}_g \varepsilon_{\mathrm{gen}} + \sqrt{\rho\,\overline{V}_\nu^\Delta} = U_\Delta.$$

This proves the upper certificate. Applying the same inequalities to the lower endpoint gives the harm certificate. If neither strict sign condition holds, the valid interval contains zero and the only justified output is "uncertain." □

C.8 Proof of Proposition B.5

Proof of Proposition B.5. For each candidate, let

$$\mathcal{A}_h := \{|\widehat{\Delta}_h - \Delta_h^{\mathrm{obs}}| \le \Gamma_h\}$$

be the audit event, and let $\mathcal{C}$ be the intersection over all $h \in \mathcal{H}$ of the sensitivity, nuisance, and semantic-contamination calibration events appearing in the proposition. The assumed joint error budget is precisely

$$\mathcal{E} := \Big(\bigcap_{h\in\mathcal{H}} \mathcal{A}_h\Big) \cap \mathcal{C}, \qquad \Pr(\mathcal{E}) \ge 1 - \delta.$$

This notation makes explicit that the audit block is drawn from the observed law P_h^{obs}, not directly from the valid task-preserving law.

On $\mathcal{E}$, the three changes of measure are handled separately. First, the audit concentration event gives

$$\Delta_h^{\mathrm{obs}} \le \widehat{\Delta}_h + \Gamma_h.$$

Second, the mixture relation in (A.11) and $|g_h| \le B$ give

$$|\Delta_h^{\mathrm{obs}} - \Delta_h^{\mathrm{val}}| \le 2B\eta_{\mathrm{sem},h}.$$

Third, the valid nuisance-law bound gives

$$|\Delta_h^{\mathrm{val}} - \Delta_{\nu,h}^{\mathrm{val}}| \le \overline{O}_h \varepsilon_{\mathrm{nuis},h}.$$

Combining these displays yields, for every candidate,

$$\Delta_{\nu,h}^{\mathrm{val}} \le \widehat{\Delta}_h + \Gamma_h + \overline{O}_h \varepsilon_{\mathrm{nuis},h} + 2B\eta_{\mathrm{sem},h}.$$

Theorem B.3 and $V_{\nu,h}^\Delta \le \overline{V}_{\nu,h}^\Delta$ then imply, simultaneously for every $\pi \in \mathcal{E}_\rho(\nu)$,

$$\Delta_{\pi,h}^{\mathrm{val}} \le U_h^{\mathrm{pair}}(\rho,\delta).$$

If all candidate certificates are nonnegative, the selection rule chooses the factual baseline, whose comparative difference is identically zero. If the minimum certificate is negative, the selected candidate has a negative upper bound, and the preceding inequality proves safety for every covered valid deployment law. Thus the selected policy is safe on the whole covered class on the same simultaneous event, which proves the proposition. □

C.9 Proof of Corollary B.1

Proof of Corollary B.1. For every candidate, boundedness gives $V_{\nu,h}^\Delta \le B^2$ and $\operatorname{osc}(g_h) \le 2B$. Substitute these two bounds into (B.15), use the fully observable radius in (B.13), and apply the same finite-class union bound and factual fallback argument as in the proof of Proposition B.5. □

C.10 Scope of the comparative result

The proof concerns a fixed pair of learned predictors evaluated on fresh, paired anchors. It does not assert that a particular CFPT optimizer reduces V_ν^Δ, nor does it identify semantic preservation for an unverified text generator. These conditions concern evaluation of trained models, independently of how their parameters were obtained.

D SHARED-ANCHOR CERTIFICATES AND CERTIFICATION POWER

D.1 PROOF OF PROPOSITION 1

Condition on the frozen candidate models, task law, and design. For IID siblings, $M \in [-B, B]$ and $Q \in [0, 2B^2]$; $\mathbb{E}[Q \mid C] = \mathrm{Var}(Z \mid C)$ by independence of each sibling pair. In the enumerated case $Q \in [0, B^2]$ by the bounded variance inequality and it equals the conditional variance. Thus $V \le B^2$ in both cases, justifying the deterministic cap on the upper endpoint. The mean event is $\Delta_j \le \bar{M}_j + r_j$ and the sensitivity event is $V_j \le V_{U,j}$. Each fails with probability at most e. For IID anchors apply the rescaled one-sided empirical Bernstein inequality (Maurer & Pontil, 2009) to M. For bounded IID Q/c_Q, the Bernoulli convex-MGF bound and Chernoff optimization give $\Pr\{D_{\mathrm{KL}}(\bar{Q}/c_Q \| V/c_Q) \ge \log(1/e)/N, \bar{Q} < V\} \le e$. Inverting yields the upper operator H_{c_Q}.

For simple random sampling without replacement, the sum's exponential moment is at most that under independent sampling from the same finite catalog (Bardenet & Maillard, 2015). Apply this comparison separately to the fixed catalog of M and the fixed catalog of Q. Hoeffding's inequality gives the stated mean radius; the same bounded Bernoulli-MGF Chernoff argument gives the identical KL upper operator. Neither step treats distinct sampling roles as independent. Finite-population corrections could tighten these bounds but are not used here.

Union over the two marginal failures for each of J candidates gives probability at most $2Je = \delta$. On the complement, monotonicity of $\sqrt{\rho V}$ yields $\Psi_j \le U_{\mathrm{sh},j}$ for every j. Hence any data-dependent choice among candidates with negative endpoints is safe on that event. An unsafe choice requires an event failure, proving family-wise false-acceptance control. This proof is conditional on a fixed candidate family; it does not authorize adding outcome-chosen models, shift classes, or methods without accounting for selection.

D.2 SEPARATE-MODEL COMPARISON

Proposition D.1 (Separate models on shared anchors)**.** *For each frozen model $X \in \{F, C_1, \ldots, C_J\}$ with loss in $[0, B]$, compute its anchor mean M_X and sibling variance statistic Q_X on the same N anchors. Set $e = \delta/[2(J+1)]$, and use the preceding mean radius with span B and sensitivity cap $B^2/4$ (observation cap $c_X = B^2/2$ for IID siblings, $B^2/4$ for enumeration). Then*

$$U_{\mathrm{sep},j} = \bar{M}_{C_j} - \bar{M}_F + r_{C_j} + r_F + \sqrt{\rho}(\sqrt{V_{U,C_j}} + \sqrt{V_{U,F}}) \tag{D.1}$$

bounds every candidate's covered robust increase with probability at least $1 - \delta$, for IID anchors or the stated finite-catalog design. Using this same e for the paired certificate gives a matched per-event ledger. Splitting the anchors across roles also remains valid.

There are $J + 1$ one-sided mean events (upper for CF, lower for factual) and $J + 1$ sensitivity events. The triangle inequality and a union bound prove the claim, without requiring independent model responses or estimation roles. The guarantee is method-wise, not permission to select among all four implementations after outcomes. The resulting 2×2 design separates representation (paired/separate) from data use (shared/split). Its pairing contrast includes sensitivity estimation as well as mean covariance; the leading covariance-only limit is stated below.

D.3 PROOF OF PROPOSITION D.1

For J CF models and the common factual model, construct the upper mean events for the CF models and the lower mean event for factual. Construct one upper sensitivity event for each of the $J + 1$ models. Each event has failure at most $e = \delta/[2(J+1)]$, by the same bounded KL and mean arguments as Proposition 1, with span B instead of $2B$. Their union has failure at most δ. On the intersection,

$$\Psi_j \le \Delta_j + \sqrt{\rho}(\sqrt{V_{C_j}} + \sqrt{V_F}) \le U_{\mathrm{sep},j}.$$

The first inequality follows either by bounding each model's centered response separately or by the L^2 triangle inequality. A shared factual model needs no additional repeated failure allocation. Paired certificates at this same e cost only $2Je \le \delta$; this deliberate conservative matching removes a ledger difference from the factorial comparison. Means and sensitivities can use the same anchors.

For without-replacement sampling the bounded-MGF comparison applies separately to each fixed finite-catalog response. The proof does not transport an IID empirical variance bound to a without-replacement setting.

In the primary controlled candidate, nuisance coefficients are equal and opposite. Disjoint-sibling statistics satisfy $Q_C = Q_F = Q_\Delta/4$ sample by sample. Since $H_{c/4}(x/4; N, e) = H_c(x; N, e)/4$, the physical caps preserve the same identity and the two sensitivity radii sum to the paired one. The mean-block means agree exactly; EB linear terms agree because $B + B = 2B$. Only $s_C + s_F$ versus s_Δ remains. This exact identity holds in both columns of the IID factorial, not merely in the population or asymptotically. It does not hold for arbitrary real-model responses, as the ReCoRD analysis illustrates.

The four factorial guarantees are method-wise. Declaring the best of four after inspection as a new safe selection would require accounting for that choice. Here they are comparative experimental endpoints, not an adaptive four-method selection procedure.

D.4 Proof of Proposition 2

For every admissible conditional density r, $\mathbb{E}[r - 1 \mid C] = 0$ and $\mathbb{E}(r-1)^2 \le \rho$. Completing the square pointwise gives

$$(r-1)(Z - \mu(C)) \le \eta(r-1)^2 + (Z - \mu(C))^2/(4\eta).$$

Consequently $\Psi \le \Delta + \eta\rho + V/(4\eta)$ for every $\eta > 0$. For two IID siblings x, y, $\mathbb{E}[(x-y)^2/2 \mid C] = v(C)$, proving $\mathbb{E}Y_\eta = \Delta + V/(4\eta)$. This accounts for unknown inner moments without enumerating nuisance states. For even K, Y_η averages $(x+y)/2 + (x-y)^2/(8\eta)$ over disjoint sibling pairs. This convex quadratic on $[-B, B]^2$ is at most its largest corner value, $\max\{B, B^2/(2\eta)\}$, and is at least $-B$. Applying the one-sided empirical Bernstein bound of Maurer & Pontil (2009, Theorem 4), rescaled to span R_η, with failure $\delta/(JH)$ and taking a union bound proves Proposition 2. Dependence between different η or candidates is harmless. Independence is across anchors, not across rows in the complete expanded dataset.

A finite-budget comparison. For the generic endpoint run safety at $\delta/2$, set $t = \log(4JH/\delta)$ and $L' = \log(4JH/\delta)$, and let $q_\eta = \eta\rho + V/(4\eta) - \sqrt{\rho V}$. A sufficient power condition is

$$\sigma_\eta\left(\sqrt{2L'/N} + \sqrt{2t/N}\right) + R_\eta\left\{\frac{2L'}{3N} + \frac{2\sqrt{tL'}}{\sqrt{N(N-1)}} + \frac{7t}{3(N-1)}\right\} < \gamma - s(\rho) - q_\eta \tag{D.2}$$

for any fixed $\eta \in \mathcal{H}$ when $\Psi \le -\gamma$ and no mismatch correction is needed. Together with Proposition D.3, this compares finite sufficient budgets, including calibration and bounded range, rather than just leading variances. KL calibration can further use Ω_K in its power analysis (Appendix D.4); the implemented cap-based confidence operator remains unchanged.

For the power bound, use safety failure $\delta/2$. Fix a grid element for each candidate independently of outcomes. Bernstein's upper tail gives $\bar{Y} - \mathbb{E}Y \le \sigma_\eta\sqrt{2L'/N} + 2R_\eta L'/(3N)$. Sample-standard-deviation concentration gives $\sqrt{s_\eta^2} \le \sigma_\eta + R_\eta\sqrt{2L'/(N-1)}$ (Maurer & Pontil, 2009, Theorem 10). A union bound over these two events for the J fixed choices costs at most $2Je^{-L'} \le \delta/2$. Substitution in the endpoint gives (D.2); safety and power then hold jointly with probability at least $1 - \delta$. The frozen grid contributes q_η, which is distinct from population slack s and sampling error. A finite grid need not certify every arbitrarily small negative margin even with infinite observations. In particular, at $V = 0$ its smallest element contributes $\rho\eta_{\min}$.

An explicit generic sufficient anchor count. With $T = \log(4JH/\delta)$ and $\kappa_\eta = \gamma - s(\rho) - q_\eta > 0$, put $c_0 = 16/3 + 2\sqrt{2}$. Since $N - 1 \ge N/2$ for $N \ge 2$, (D.2) follows from

$$N > \max\left\{2, \frac{32T\sigma_\eta^2}{\kappa_\eta^2}, \frac{2c_0TR_\eta}{\kappa_\eta}\right\}.$$

Minimizing this sufficient count over a predeclared K, η grid, and comparing it with the $m + n$ conditions of Proposition D.3, makes calibration cost explicit. Charge KN rather than N when compute is constrained. These constants are conservative sufficient conditions; observed power is measured separately and they do not order every estimator.

At $V = 0$, $Q = 0$ almost surely and $\sigma_\eta = \sigma_A$. For an adequately predeclared grid, the generic remainder balances $\rho\eta$ against $B^2T/(N\eta)$, giving order $B\sqrt{\rho T/N}$, in addition to $\sigma_A\sqrt{T/N} + BT/N$ and grid slack. Thus the root-rate small-sensitivity behavior of explicit KL is not exclusive to that procedure. A fixed finite grid still has the approximation floor stated above.

D.5 Proof of Proposition 3

Write $X = Z - \mu(C)$. For an IID pair X, X', expansion gives $\mathbb{E}(X - X')^4 = 2\kappa_4 + 6v^2$, whence $\mathrm{Var}((X-X')^2/2 \mid C) = (\kappa_4+v^2)/2$. Averaging $K/2$ independent pairs and applying total variance yields Ω_K. Expanding the covariance of the pair mean and squared difference gives $\kappa_3/2$; averaging pairs yields $\mathrm{Cov}(M, Q \mid C) = \kappa_3/K$. Total covariance gives (6). Finally apply total variance to M and bilinearity to Y_η. No conditional symmetry or homoskedasticity is assumed in these identities. In particular, a negative mean–variance covariance must not be discarded when evaluating generic DRO.

For calibration observations $Q \in [0, c_Q]$, Bernstein yields, except with probability e',

$$\bar{Q} \le V + \sqrt{2\Omega_K L'/m} + 2c_Q L'/(3m).$$

Combining this with the deterministic KL-root inequality proved in Appendix D.10 gives an alternative to the uniform A_Q of Proposition D.3:

$$A_Q^{\mathrm{het}} = \sqrt{V + \sqrt{2\Omega_K L'/m} + 2c_Q L'/(3m)} - \sqrt{V} + \sqrt{c_Q L/m}.$$

One may pre-select the sharper of the two theoretical expressions using population parameters in a power calculation, with the same single tail budget; this is not an oracle input to an observed certificate. At $V = 0$, $Q = 0$ almost surely and the exact endpoint is preferable. The KL cap term can dominate even if the observed variance of Q is small. Thus heterogeneity affects the two implementations differently, without yielding a general ranking. The leading mean-only separate/paired variance ratio for equal anchor scales and correlation r is $2/(1 - r)$; finite budgets also pay calibration, nuisance-estimation, confidence, and range terms.

D.6 Explicit finite-sample power widths

For a mean variable with variance σ^2, span R, and failure e, write $t = \log(2/e)$, $u = \log(1/\beta)$, and define

$$\begin{aligned}\mathcal{B}_R(\sigma, N; e, \beta) &= \sigma(\sqrt{2u/N} + \sqrt{2t/N}) \\ &\quad + R\{2u/(3N) + 2\sqrt{tu}/\sqrt{N(N-1)} + 7t/[3(N-1)]\}, \qquad \text{(D.3)}\\ \mathcal{A}_c(V, \Omega, N; e, \beta) &= \min\Big\{\sqrt{c/N}(\sqrt{\log(1/e)} + \sqrt{u}), \\ &\quad \sqrt{V + \sqrt{2\Omega u/N} + 2cu/(3N)} - \sqrt{V} + \sqrt{c\log(1/e)/N}\Big\}. \qquad \text{(D.4)}\end{aligned}$$

The minimum chooses between two deterministic population-parameter power bounds; it does not select between data-dependent confidence events. At $V = 0$, replace $\mathcal{A}_c$ by $\sqrt{c(1 - e^{-\log(1/e)/N})}$, since $Q = 0$ almost surely.

D.7 Proof of Theorem 1

For a bounded response X with variance σ^2 and span R, one-sided Bernstein gives $\bar{X} - \mathbb{E}X \le \sigma\sqrt{2u/N} + 2Ru/(3N)$ except with probability β. Sample-standard-deviation concentration gives $s_X \le \sigma + R\sqrt{2u/(N-1)}$ except with probability β (Maurer & Pontil, 2009). Substitution into the empirical Bernstein endpoint yields (D.3). This is a power upper bound on the random endpoint, distinct from the confidence event controlling its lower tail.

For $Q \in [0, c]$ with mean V and variance Ω, Bernstein gives $\bar{Q} \le V + \sqrt{2\Omega u/N} + 2cu/(3N)$ with failure β. The deterministic KL-root inequality $\sqrt{H_c(\bar{Q}; N, e)} \le \sqrt{\bar{Q}} + \sqrt{c \log(1/e)/N}$ yields the second branch of (D.4). The uniform branch is the bounded KL-root power bound proved below, also with failure β. Which theoretical branch is smaller is a deterministic function of population moments, so choosing it requires no union over two data-dependent procedures. It is not a claim that both events hold at cost β. At $V = 0$, bounded nonnegative Q is identically zero, and $H_c(0; N, e) = c(1 - e^{-\log(1/e)/N})$ exactly.

Combining the two mean-power events and one sensitivity-power event gives $U_P \le -\gamma + s_P + W_P$ except with probability 3β per paired candidate. For separate models there are $J + 1$ marginal triples, with a lower fluctuation for the factual mean equivalently handled as an upper fluctuation of its negative. Their total cost is $3(J + 1)\beta$. For generic DRO, fix one η per candidate using only population parameters, independently of the sample. Its expected response plus $\eta\rho$ equals $\Psi + s_P + q_\eta$, and its bounded anchor variance is σ_η^2. Two power events give (9); optimizing the observed valid endpoint over its frozen grid can only lower it. The choice of β in the theorem is a common conservative ledger for the separate method-wise power statements, not joint coverage of arbitrary method search. Safety costs δ independently of the stated power failure budget; joint safety and power have failure at most their sum.

For split roles, use n in the mean term and m in the sensitivity term. The same marginal-event reasoning applies. Shared roles use N in both, not $2N$ independent observations. The nuisance-variance components obey

$$\tau_\Delta^2 = \sigma_A^2 + V/K, \qquad \Omega = \mathrm{Var}_C\, v + (\mathbb{E}\kappa_4 + \mathbb{E}v^2)/K, \qquad \zeta = \mathrm{Cov}_C(\mu, v) + \mathbb{E}\kappa_3/K.$$

Thus K cannot remove σ_A^2 or $\mathrm{Var}_C\, v$. The bounded range and logarithms remain even if these variances are small. Population slack s_P, the larger separate slack s_S, and fixed-dual slack q_η consume margin before any statistical width is added.

D.8 Population forecasts and covariance attribution

Oracle efficiency, not a minimax claim. For interior V, put $\alpha = \sqrt{\rho}/(2\sqrt{V})$ and $\sigma_I^2 = \tau_\Delta^2 + \alpha^2\Omega + 2\alpha\zeta$. The paired moment endpoint has the local expansion $\sqrt{N}(U_P - \Delta - \sqrt{\rho V}) \Rightarrow \mathcal{N}(a_P, \sigma_I^2)$, where

$$a_P = \sqrt{2t}\,\tau_\Delta + \sqrt{\rho(c_Q - V)\log(1/e)/2}.$$

With $z_p = \Phi^{-1}(p)$ the standard-normal quantile, an oracle local power benchmark is $N_p \simeq (a_P + z_p\sigma_I)^2/(\gamma - s_P)^2$; finite-budget forecasts below retain the actual KL inversion and range term. For generic fixed η, replace a_P by $\sqrt{2\log(2/e_G)}\,\sigma_\eta$ and use margin $\gamma - s_P - q_\eta$. In the balanced controlled law, the separate endpoint has the same sensitivity offset $d = \sqrt{\rho(c_Q - V)\log(1/e)/2}$ and influence variance as the paired endpoint. At equal population margin, put $D_p = d + z_p\sigma_I$; then

$$\frac{N_{S,p}}{N_{P,p}} \simeq \left\{ \frac{\sqrt{2t}(\tau_C + \tau_F) + D_p}{\sqrt{2t}\,\tau_\Delta + D_p} \right\}^2, \qquad \tau_\Delta^2 = \tau_C^2 + \tau_F^2 - 2\,\mathrm{Cov}(M_C, M_F). \tag{D.5}$$

This includes the paired fluctuation of the separate endpoint's mean, while showing how nuisance cost D_p attenuates covariance savings. Bounded-range terms matter at finite budgets. Appendix D.8 gives the assumptions, separate influence function, proofs, and oracle-versus-sufficient comparisons.

Local oracle benchmark. Assume a fixed IID bounded law with $0 < V < B^2$, nondegenerate relevant variances, fixed J, e, ρ, and a location-shift family in Δ that preserves the centered joint law of (M, Q). For this interior regime, Taylor expansion of binary KL inversion gives

$$H_c(\bar{Q}; N, e) = \bar{Q} + \sqrt{2\bar{Q}(c - \bar{Q})\log(1/e)/N} + O_p(N^{-1}).$$

The cap is asymptotically inactive. The delta method and sample-variance consistency yield

$$\sqrt{N}(U_P - \Delta - \sqrt{\rho V}) = a_P + N^{-1/2}\sum_i \{M_i - \Delta + \alpha(Q_i - V)\} + o_p(1).$$

The IID CLT gives the preceding normal limit and the local power benchmark along $N \to \infty$, $\gamma - s_P \to 0$ with $\sqrt{N}(\gamma - s_P)$ finite. It is neither an information lower bound nor a minimax-optimality claim. For fixed generic η, apply the CLT directly to Y_η; its range correction is $O(N^{-1})$. A grid-selected η does not automatically have this single-η limit when several elements compete.

For the separate expansion also assume $0 < V_X < B^2/4$. For separate shared moments, put $\alpha_X = \sqrt{\rho}/(2\sqrt{V_X})$. Its influence variable is

$$I_S = M_C - M_F - \Delta + \alpha_C(Q_C - V_C) + \alpha_F(Q_F - V_F),$$

with variance

$$\begin{aligned}\sigma_{I_S}^2 = \tau_\Delta^2 + \alpha_C^2\Omega_C + \alpha_F^2\Omega_F + 2\alpha_C\alpha_F \operatorname{Cov}(Q_C, Q_F) \\ + 2\alpha_C \operatorname{Cov}(M_C - M_F, Q_C) + 2\alpha_F \operatorname{Cov}(M_C - M_F, Q_F).\end{aligned}$$

Its confidence offset is $\sqrt{2t}(\tau_C + \tau_F) + \sum_X \sqrt{\rho(c_X - V_X)\log(1/e)/2}$. Hence $N_{S,p} \simeq (a_S + z_p\sigma_{I_S})^2/(\gamma - s_S)^2$. In the balanced primary law, homogeneity gives identical sensitivity offsets and influence variables for paired and separate methods, yielding (D.5). When nuisance uncertainty is negligible, $D_p = z_p\tau_\Delta$, recovering the mean-only ratio $[\{\sqrt{2t}(\tau_C+\tau_F)+z_p\tau_\Delta\}/\{(\sqrt{2t}+z_p)\tau_\Delta\}]^2$. This approximation is justified, for example, by $\rho = 0$ or an oracle known-zero-sensitivity comparison. Merely observing $V = 0$ does not justify dropping its KL calibration cost: that boundary retains a nonzero $N^{-1/2}$ radius, and the interior delta-method expansion does not apply.

Our finite-N oracle forecast substitutes exact design moments in the actual KL and empirical-Bernstein operators, retains the linear range term, and adds z_p times the delta-method sampling standard deviation. For shared roles this uses the full mean–variance covariance; for split roles, the mean and sensitivity estimates lie on disjoint blocks. The latter's variance is $\tau_\Delta^2/n + \operatorname{Var}(I_Q)/m$, with no cross term. Derivatives of the KL radius are calculated numerically from the fixed operator, not fitted to empirical thresholds. This is a post-simulation population-moment calculation made without fitting the confirmation outcomes, not a prospective data-based power guarantee. Rigorous sufficient budgets are separately obtained by inverting Theorem 1.

D.9 Split-role certificates and power

Independent mean and calibration blocks provide an alternative to shared-anchor estimation. Propositions B.4–B.5 specify the needed events but do not themselves construct their calibration bounds. We now give a bounded procedure, including an explicit failure ledger. This construction uses no oracle sensitivity or population variance.

Protocol and independent units. Conditional on training, freeze J model pairs, the loss envelope $[0, B]$, ν, ρ, δ, and all selection rules. Draw m calibration anchors and n independent audit anchors from the same task law. Their labels must be valid for the loss being certified. In the sampled-sibling design, each anchor has $K \geq 2$ conditionally IID draws from ν. Alternatively, enumerate a known finite nuisance support and integrate its known probabilities exactly. Enumeration is not K independent task observations. All statements below condition on the trained models, and require independence from their training and selection history. Reusing an inspected test block is not prospective confirmation.

For a paired response $Z \in [-B, B]$, define the calibration cluster mean M_i and nuisance-variance observation Q_i by

$$M_i = K^{-1}\sum_j Z_{ij}, \qquad Q_i = \frac{1}{\lfloor K/2 \rfloor}\sum_{j=1}^{\lfloor K/2 \rfloor}\frac{(Z_{i,2j-1} - Z_{i,2j})^2}{2}.$$

Thus $\mathbb{E}Q_i = V_\nu^\Delta$ and $0 \leq Q_i \leq 2B^2$. With exact enumeration, instead use $M_i = \mathbb{E}_\nu[Z \mid T_i]$ and $Q_i = \operatorname{Var}_\nu(Z \mid T_i) \leq B^2$. In either design, independent calibration-anchor pairs give

$$D_i = (M_{2i} - M_{2i-1})^2/2, \qquad \mathbb{E}D_i = \tau_K^2 := \operatorname{Var}(M_i), \quad 0 \leq D_i \leq 2B^2.$$

In the sampled design $\tau_K^2 = \sigma_A^2 + \sigma_S^2/K$; with enumeration $\tau_K^2 = \sigma_A^2$. The nonnegative plug-in estimate $[s_M^2 - \bar{Q}/K]_+$ is descriptive, not an upper confidence bound.

The identity $\mathbb{E}Q_i = V_\nu^\Delta$ uses deterministic task labels and deterministic evaluated losses, as in the controlled and benchmark-conditional experiments. With random labels at fixed C, it instead estimates $V_{\mathrm{obs}} = \mathbb{E}_{C,Y}\,\mathrm{Var}_{S|C}(Z \mid C, Y) \geq V_\nu^\Delta$ by Jensen. Safety remains conservative; power must then include the additional non-vanishing radius $\sqrt{\rho}(\sqrt{V_{\mathrm{obs}}} - \sqrt{V_\nu^\Delta})$. This restriction does not change the population transfer theorem.

A bounded calibration operator. For independent $X_i \in [0, c]$, let $a = \bar{X}/c$ and define

$$H_c(X; e) = c\sup\{u \in [a, 1] : D_{\mathrm{KL}}(a\|u) \leq \log(1/e)/|X|\}. \tag{D.6}$$

Here $D_{\mathrm{KL}}(a\|u)$ is the Bernoulli KL defined in the main text, with continuous endpoint conventions. Bounded observations need not be Bernoulli. This is a confidence inversion, not the chi-square shift class. Use $\bar{\tau}^2 = \min(B^2, H_{2B^2}(D; e))$ and $\bar{V} = \min(B^2, H_{c_Q}(Q; e))$, where $c_Q = 2B^2$ for sampled siblings and B^2 for enumeration. For single-model losses, replace the paired range $2B$ by B, with variance cap $B^2/4$.

Proposition D.2 (Observable joint certificate)**.** *Under the preceding sampling protocol, reserve external failure budget $\delta_{\mathrm{ext}} < \delta$ for any supplied mismatch and contamination bounds, and set $e = (\delta - \delta_{\mathrm{ext}})/(12J)$. For each policy, allocate one event to each of the variance bound, sensitivity bound, two-sided Bernstein mean bound, and two-sided Hoeffding mean bound, for each of the paired, CF, and factual responses. Writing $t = \log(2/e)$, define*

$$r_h = \min\left\{\frac{2Bt}{3n} + \sqrt{\frac{2\bar{\tau}_h^2 t}{n} + \frac{4B^2t^2}{9n^2}},\quad B\sqrt{\frac{2t}{n}}\right\}, \tag{D.7}$$

$$\widetilde{V}_h = \min\{\bar{V}_h, (\sqrt{\bar{V}_{h,\mathrm{CF}}} + \sqrt{\bar{V}_{h,\mathrm{F}}})^2\}, \tag{D.8}$$

$$C_h = 2B(\bar{\varepsilon}_h + \bar{\eta}_h), \qquad U_h = \widehat{\Delta}_h + r_h + C_h + \sqrt{\rho\widetilde{V}_h}, \quad L_h = \widehat{\Delta}_h - r_h - C_h - \sqrt{\rho\widetilde{V}_h}. \tag{D.9}$$

Then, jointly for every policy and covered law, $L_h \leq \Delta_{\pi,h} \leq U_h$ with probability at least $1 - \delta$. Selecting the smallest negative U_h, or factual training if there is none, is safe on this event. The same joint event covers the separate-model certificate obtained by adding its two mean radii and nuisance radii. A bounded alternative replaces r_h by its Hoeffding branch and $\widetilde{V}_h$ by B^2. If the audit law differs from the calibration reference law, including through contamination, use only the Hoeffding mean branch unless an independent audit-law variance bound is supplied. A reference-law variance is not silently transported to the audit law.

Irreducible envelope slack. Define the exact utility target and its sensitivity-envelope slack by

$$\Psi_h(\rho) = \sup_{\pi \in \mathcal{E}_\rho(\nu)} \Delta_{\pi,h}, \qquad s_h(\rho) = \Delta_{\nu,h} + \sqrt{\rho V_h} - \Psi_h(\rho) \geq 0.$$

Proposition B.3 gives $s_h = 0$ in its local regime. Outside that regime, a negative Ψ_h alone need not make this particular certificate powerful, regardless of audit size. Non-local envelope slack is not statistical error.

Proposition D.3 (KL-calibrated certification power)**.** *Assume exact reference sampling, deterministic valid labels, and $\Psi_h(\rho) \leq -\gamma$. Run Proposition D.2 with safety budget $\delta/2$. For zero external failure budget, set $e = \delta/(24J)$, $e' = \delta/(4J)$, $L = \log(1/e)$, $L' = \log(1/e')$, and $t = \log(2/e)$. Define*

$$A_c(r) = \sqrt{c/r}(\sqrt{L} + \sqrt{L'}), \qquad A_D = A_{2B^2}(\lfloor m/2 \rfloor), \quad A_Q = A_{c_Q}(m), \tag{D.10}$$

$$\tau_K^2 = \sigma_A^2 + \sigma_S^2/K, \qquad \kappa = \gamma - s_h(\rho) - 2C_h > 0. \tag{D.11}$$

For enumeration, replace τ_K^2 by σ_A^2. Then h has a negative upper certificate with probability at least $1 - \delta$ whenever

$$2\sqrt{2t/n}\,(\tau_K + A_D) + \frac{8Bt}{3n} + \sqrt{\rho}A_Q < \kappa. \tag{D.12}$$

In particular, for any pre-chosen $a \in (0, 1)$, sufficient sizes are

$$m \geq \frac{\rho c_Q(\sqrt{L} + \sqrt{L'})^2}{a^2\kappa^2}, \quad m \geq 2, \tag{D.13}$$

$$n > \max\left\{\frac{32t(\sigma_A^2 + \sigma_S^2/K + 2\tau_K A_D + A_D^2)}{(1-a)^2\kappa^2},\quad \frac{16Bt}{3(1-a)\kappa}\right\}. \tag{D.14}$$

With $V = 0$, replace A_Q by the exact deterministic value $\sqrt{c_Q(1 - e^{-L/m})}$ and replace $(\sqrt{L}+\sqrt{L'})^2$ in (D.13) by L. Thus zero sensitivity has an $O(m^{-1/2})$ nuisance-radius uncertainty, not $O(m^{-1/4})$. For a valid nonzero external failure ledger, use its reduced e.

What calibration actually costs. The essential inequality, valid uniformly down to $V = 0$, is

$$\sqrt{H_c(X; e)} \le \sqrt{\mathbb{E}X} + A_c(|X|)$$

except on an event of probability e'. In particular, $H_c - \mathbb{E}X \le 2\sqrt{\mathbb{E}X}A_c + A_c^2$. The variance dependence is explicit: for sensitivity this is $2\sqrt{V}A_Q + A_Q^2$, while its contribution after taking a square root is at most A_Q. Substituting a generic Hoeffding width before taking that square root would lose a factor and suggest $m^{-1/4}$; it does not describe the implemented KL calibration. Up to logarithms, the statistical terms in (D.12) are

$$\tau_K \sqrt{\frac{\log(J/\delta)}{n}} + \frac{B \log(J/\delta)}{\sqrt{mn}} + \frac{B \log(J/\delta)}{n} + B\sqrt{\frac{\rho \log(J/\delta)}{m}}. \tag{D.15}$$

Here V also enters the available margin: when $\gamma = -\Psi_h$, $\kappa = -\Delta_{\nu,h} - \sqrt{\rho V_h} - 2C_h$. A positive κ is essential. Increasing m, n cannot remove s_h, a genuine mismatch/label correction, or the population nuisance radius. For stochastic labels the additional radius $\sqrt{\rho}(\sqrt{V_{\text{obs}}} - \sqrt{V})$ must also be deducted. These are sufficient finite-sample sizes, not minimax-optimal constants.

Corollary D.1 (When pairing saves independent information)**.** *Suppose the paired and separate population robustness envelopes coincide and are negative, and both use valid independent calibration. For a shared anchor define the CF and factual cluster means M_C, M_F. Their mean-estimation variances satisfy*

$$\tau_\Delta^2 = \tau_C^2 + \tau_F^2 - 2\,\mathrm{Cov}(M_C, M_F).$$

For $\tau_\Delta > 0$, at fixed positive margin, when calibration and Bernstein linear terms are negligible, the leading sufficient audit budget for separate mean bounds relative to a paired mean bound scales, apart from confidence logarithms, as

$$\frac{(\tau_C + \tau_F)^2}{\tau_\Delta^2}. \tag{D.16}$$

This statement concerns these certificate constructions, not an information lower bound for every method called "separate." In particular, a generic robust estimator using paired observations can share or exceed the benefit. The finite-budget ratio must count calibration as well as audit anchors.

Increasing K helps the mean term until $\sigma_S^2/K \ll \sigma_A^2$; it does not eliminate the anchor floor or substitute for calibration anchors. For a fixed anchor budget $N = m + n$, ignoring the smaller mixed and linear terms, minimizing $\alpha/\sqrt{n} + \beta/\sqrt{m}$ suggests $m/n = (\beta/\alpha)^{2/3}$, where $\alpha = 2\sqrt{2t}\tau_K$ and $\beta = \sqrt{\rho c_Q}(\sqrt{L} + \sqrt{L'})$. This is a planning approximation, not a data-dependent rule for retuning a completed audit. Use (D.12) when these neglected terms are appreciable, and charge $K(m + n)$ when model evaluations are the budget.

Proposition D.4 (An anchor necessity bound)**.** *For every $0 < \gamma \le B/2$, there are anchor-only paired problems with zero nuisance variation for which any test having false acceptance at most $1/4$ at zero gain and acceptance at least $3/4$ at gain $-\gamma$ requires $n \ge B^2/(2\gamma^2)$ independent observed anchors. Siblings do not alter this requirement. If calibration labels are also available to the test, n here denotes the total independent anchors in both blocks.*

Proofs and the exact event accounting are in Appendix D.10. Reference-only selection uses the same anchors but has no covered-shift guarantee; a validation interval for the reference mean is likewise not a robust certificate.

D.10 Proofs for split-role certificates

Bounded calibration lemma. For $X/c \in [0, 1]$ with mean p, convexity gives $\mathbb{E}e^{\lambda X/c} \le 1 - p + pe^{\lambda}$ for every real λ. Independence and Chernoff optimization therefore imply, for $a < p$, $\Pr(\bar{X}/c \le a) \le \exp[-mD_{\mathrm{KL}}(a\|p)]$. Inverting this monotone lower-tail bound yields $\Pr(\mathbb{E}X > H_c(X; e)) \le e$. This proves (D.6) without assuming binary observations. The argument also applies to independent non-identically distributed bounded observations with a common target given by the mean of their expectations, by concavity of the log Bernoulli moment generating function.

Proof of Proposition D.2. Two conditionally independent siblings satisfy $\mathbb{E}[(Z - Z')^2/2 \mid T] = \mathrm{Var}(Z \mid T)$. Averaging disjoint pairs preserves this expectation; concentration is applied to the resulting Q_i across anchors. For independent cluster means, $\mathbb{E}[(M - M')^2/2] = \mathrm{Var}(M)$, so the D_i are independent bounded observations. Their ranges and the variance caps follow from the range of Z. The bounded calibration lemma establishes both upper bounds.

Conditional on the independent calibration block and its upper-variance event, two-sided Bernstein for independent cluster means in $[-B, B]$ gives

$$\Pr(|\widehat{\Delta} - \Delta_{\mathrm{obs}}| > r \mid \text{calibration}) \le 2\exp\left[-\frac{nr^2}{2\bar{\tau}^2 + (4B/3)r}\right].$$

Solving its quadratic exponent gives the first branch of r_h. Hoeffding gives its second branch at failure probability e. The minimum is valid on the intersection, which is why both events are budgeted. Apply the identical argument to each single-model response with its own range B. Theorem B.3 permits the triangle improvement to $\widetilde{V}_h$ on the already allocated events. The mean transport/contamination corrections yield $|\Delta_{\mathrm{obs}} - \Delta_\nu^{\mathrm{val}}| \le C_h$. A union bound over $12J$ events and the external events costs at most $12Je + \delta_{\mathrm{ext}} = \delta$. The comparative transfer theorem now gives the two endpoints for all covered laws, simultaneously over candidates. The safe-selection implication follows by taking the smallest negative endpoint; factual fallback has risk difference zero identically. With a mismatched audit law, the Hoeffding branch is distribution-free, whereas the calibrated reference variance need not transfer. Random labels yield a conservative sensitivity by conditional Jensen: averaging over labels before taking the squared centered response cannot increase its variance. □

Square-root KL width lemma. Write $p = \mathbb{E}X/c$, $a = \bar{X}/c$, and $r = |X|$. Binary KL dominates squared Hellinger distance, hence $D_{\mathrm{KL}}(a\|u) \ge (\sqrt{u} - \sqrt{a})^2$. For completeness, Jensen gives $D_{\mathrm{KL}}(P\|Q) \ge -2\log\sum_x \sqrt{P_x Q_x} \ge 2(1 - \sum_x \sqrt{P_x Q_x})$; the latter includes both nonnegative coordinate differences. Inverting gives $\sqrt{H_c} \le \sqrt{\bar{X}} + \sqrt{cL/r}$. To control the observed mean without losing the square-root rate, convexity and $1 + x \le e^x$ yield $\mathbb{E}e^{\lambda X/c} \le \exp[p(e^\lambda - 1)]$ for $\lambda \ge 0$. Consequently, for $q \ge p$,

$$\Pr(a \ge q) \le \exp[-r\{q\log(q/p) - q + p\}] \le \exp[-r(\sqrt{q} - \sqrt{p})^2].$$

The last inequality follows by setting $x = q/p \ge 1$ and observing $x\log x - 2x + 2\sqrt{x} \ge 0$ (its derivative is $\log x - 1 + x^{-1/2} \ge 0$). The endpoint $p = 0$ is deterministic. Thus, with probability at least $1 - e'$, $\sqrt{\bar{X}} \le \sqrt{\mathbb{E}X} + \sqrt{cL'/r}$, proving the stated uniform square-root width. When $\mathbb{E}X = 0$, nonnegativity implies $X_i = 0$ almost surely and $D_{\mathrm{KL}}(0\|u) = -\log(1-u)$ gives $H_c = c(1 - e^{-L/r})$ exactly.

Proof of Proposition D.3. Intersect the safety event (failure $\delta/2$) with the two upper-width events per policy ($2Je' = \delta/2$). Calibration caps and the triangle improvement only decrease upper bounds, so $\sqrt{\bar{\tau}^2} \le \tau_K + A_D$ and $\sqrt{\widetilde{V}} \le \sqrt{V} + A_Q$. For the Bernstein branch, $\sqrt{x + y} \le \sqrt{x} + \sqrt{y}$ gives

$$r_h \le \sqrt{2t/n}(\tau_K + A_D) + 4Bt/(3n).$$

Safety implies $\widehat{\Delta} \le \Delta_\nu + r_h + C_h$; substituting in the certificate and using $\Delta_\nu + \sqrt{\rho V} = \Psi_h + s_h$ gives

$$U_h \le -\gamma + s_h + 2C_h + 2\sqrt{2t/n}(\tau_K + A_D) + 8Bt/(3n) + \sqrt{\rho}A_Q.$$

Equation (D.12) makes this negative. Allocating $a\kappa$ to the sensitivity width gives (D.13); dividing the remaining $(1 - a)\kappa$ equally between the square-root and linear mean terms gives (D.14). The exact zero-mean KL inversion replaces the sensitivity width event at $V = 0$. If $a^2\kappa^2 < \rho c_Q$, the exact zero-V calibration condition can alternatively be written

$$m \ge \frac{L}{-\log(1 - a^2\kappa^2/(\rho c_Q))}.$$

If the fraction is at least one, the sensitivity-width condition is automatic; if $\rho = 0$ it is absent. Taking a slightly larger integer n ensures the strict negativity required for acceptance. No claim is made that the selected accepted candidate minimizes true robust risk. □

Proof of Corollary D.1. The variance identity follows by expanding $\mathrm{Var}(M_C - M_F)$. Covariance reduction in differences also motivates common random numbers (Glasserman & Yao, 1992); here the paired response must additionally support a covered-shift certificate. A paired mean radius has leading term $\tau_\Delta\sqrt{2t/n}$; adding two separate radii gives $(\tau_C + \tau_F)\sqrt{2t/n}$ when their confidence logarithms agree. Squaring to attain a fixed mean-error margin gives (D.16). Actual logarithms can differ because a common factual bound is reusable; these change constants, not the covariance identity. The claim explicitly removes population-envelope differences and requires negligible calibration error. It therefore does not predict the finite-budget ratio without counting m. A paired DRO estimator uses the same covariance cancellation and is not subject to a separate-radii penalty. □

Proof of Proposition D.4. Let $T \sim \mathrm{Bernoulli}(p)$ and $Z = B(2T - 1)$, identically for all siblings. This is realized by two losses in $[0, B]$: the candidate loss is BT, the factual loss is $B(1 - T)$. Compare $p_0 = 1/2$ with $p_1 = 1/2 - \gamma/(2B)$. The means are zero and $-\gamma$, respectively, and $V = 0$. The elementary inequality $D_{\mathrm{KL}}(p_1\|p_0) \le \chi^2(p_1\|p_0) = \gamma^2/B^2$ and Pinsker imply that the total variation between n-anchor experiments is at most $\gamma\sqrt{n/(2B^2)}$. An acceptance-probability difference of at least $1/2$ requires total variation at least $1/2$, giving the stated bound. Repeating a deterministic sibling adds no statistical information. □

E Pilot-Based Audit Planning: Construction and Proofs

E.1 Statistical experiment and the information available to the planner

Let an anchor produce a vector of bounded factual and CF losses at K IID nuisance siblings. Different anchors are IID; sibling observations within an anchor need not be marginally independent. Models, candidates, loss, nuisance law and ρ are fixed before data collection. The pilot contains two independent anchor blocks of sizes r_0, r_1; all final observations are independent of both. The larger sibling count in the frozen menu may be collected in the pilot, with prefix sub-vectors used for smaller K. These sub-vectors are dependent across menu entries, which causes no problem for a simultaneous confidence event.

Before the pilot, freeze a finite menu $\mathcal{H}$ of audit sizes, sibling counts, estimators and confidence allocations, and valid known support enclosures for every scalar linear combination used below. Support enclosures must not be estimated by observed minima and maxima. In the general bounded-loss problem the global loss range supplies these enclosures; the controlled experiment supplies tighter known-law supports. We do not claim that real-model outputs have those tighter supports. No probabilities on the support, population means, covariance matrices, or final outcomes are inputs to the data-driven planner. The final size is fixed before final outcomes; this is distinct from an anytime-valid confidence-sequence design (Howard et al., 2021).

The design block chooses a centering constant for each scalar variable, and a positive tangency point for each sensitivity component. Conditional on the design block all these are fixed. The estimation block constructs simultaneous moment intervals across *all* entries of $\mathcal{H}$, including all K, budget-dependent tangent responses, confidence allocations, and fixed-dual choices. In particular, pilot choice of a dual parameter does not constitute uncorrected search on final outcomes.

E.2 Explicit simultaneous pilot intervals

For a scalar $X \in [l, u]$, let $c_0 \in [l, u]$ be its design-block center, $R = u - l$, and $D = \max\{|l - c_0|, |u - c_0|\}$. On r_1 estimation anchors, form intervals for $\mathbb{E}X$ and $\mathbb{E}(X - c_0)^2$, and an upper bound for $\mathbb{E}|X - c_0|^3$. Apply the one-sided empirical Bernstein radius with respective spans R, D^2, D^3. There are five events per scalar: two for the mean, two for the centered second moment, one for the absolute third moment. If the search uses F scalar variables, use $e_{\mathrm{pilot}} = \delta_{\mathrm{plan}}/(5F)$ and a union bound. Conditional on the design pilot, simultaneous coverage is at least $1 - \delta_{\mathrm{plan}}$; averaging preserves this probability.

Write the mean interval as $[\mu^-, \mu^+]$, the second-moment interval as $[s_2^-, s_2^+]$, and the third-moment upper bound as s_3^+, clipping to their known support bounds. Define

$$d_- = \operatorname{dist}(c_0, [\mu^-, \mu^+]), \qquad d_+ = \max\{|\mu^- - c_0|, |\mu^+ - c_0|\},$$
$$(\sigma^-)^2 = \max\{0, s_2^- - d_+^2\}, \qquad (\sigma^+)^2 = \min\{R^2/4, s_2^+ - d_-^2\},$$
$$T^+ = \min\{R(\sigma^+)^2,\ [(s_3^+)^{1/3} + d_+]^3\}.$$

On the coverage event these bound the variance and third central absolute moment. The variance identity proves the first two bounds. Minkowski's inequality proves the cubic bound, and $|X - \mathbb{E}X|^3 \le R(X - \mathbb{E}X)^2$ proves its alternative. Numerically we clip upper variances at zero if intervals conflict outside the coverage event; no guarantee is claimed for such conflicts. Variance lower bounds of zero are retained. The method never substitutes an observed zero for a known population zero.

For each K we retain estimates of the means and covariance matrix of $(M_\Delta, Q_\Delta, M_C, M_F, Q_C, Q_F)$. The rigorous witness uses direct intervals for its linear response Y. Thus the identity $\operatorname{Var}(M + aQ) = \operatorname{Var} M + a^2 \operatorname{Var} Q + 2a \operatorname{Cov}(M, Q)$ is respected without separately discarding the covariance or adding independent confidence errors to its terms. Likewise, $\operatorname{Var}(M_C - M_F)$ includes factual–CF covariance. Point covariance estimates are diagnostics; the confidence set for Y provides the actual guarantee.

E.3 Power witnesses

Two valid power witnesses. One witness maximizes the sufficient inequalities of Theorem 1 over $\mathcal{C}_r$. A sharper alternative uses a global tangent, rather than an asymptotic linearization, of the concave function $f_N(q) = \sqrt{\rho \min\{V_{\max}, H_c(q; N, e_Q)\}}$. For paired KL this gives $U_P \le \bar{Y} + b_N +$ the EB mean radius, with $Y = M + f_N'(q_0)Q$ and $b_N = f_N(q_0) - f_N'(q_0)q_0$. Separate KL uses two tangents; fixed-dual DRO uses $Y = M + Q/(4\eta)$ and $b_N = \rho\eta$ directly. Write each majorant as

$$\bar{Y} + b_N + \sum_{j=1}^{d} \left\{ \sqrt{2t_j/N}\, s_j + \frac{7R_j t_j}{3(N-1)} \right\}, \qquad t_j = \log(2/e_j).$$

Let superscripts $+/-$ denote pilot upper/lower bounds and let $T_Y^+ = \sup_{\mathcal{C}_r} \mathbb{E}|Y - \mathbb{E}Y|^3$. Set $\varepsilon_N = .56T_Y^+/[(\sigma_Y^-)^3\sqrt{N}]$ and $q_N = 1 - (1-p)/J + d\beta_s + \varepsilon_N$. If $\sigma_Y^- > 0$, $q_N < 1$, and

$$\mu_Y^+ + b_N + \sum_j \left\{ \sqrt{2t_j/N} \left(\sigma_j^+ + R_j \sqrt{\frac{2\log(1/\beta_s)}{N-1}} \right) + \frac{7R_j t_j}{3(N-1)} \right\} + \Phi^{-1}(q_N)\sigma_Y^+/\sqrt{N} < 0, \tag{E.1}$$

then Berry–Esseen and sample-standard-deviation concentration imply the required conditional power. A zero lower variance disables this witness; it does not justify a Gaussian approximation. Theorem 1 remains available. Neither witness assumes mean/variance independence.

Worst-case moment witness. Let the pilot supply upper mean and variance bounds and $V \in [V^-, V^+]$, $\Omega \le \Omega^+$ for each sensitivity variable. The function $\mathcal{B}_R$ increases with its standard-deviation argument. In the adaptive branch of $\mathcal{A}_c$ the increment $\sqrt{V + d} - \sqrt{V}$ decreases with V and increases with d. Consequently a valid paired sufficient condition is

$$\Delta^+ + \sqrt{\rho V^+} + \mathcal{B}_{2B}(\sigma_M^+, N; e_M, \beta) + \sqrt{\rho}\, \mathcal{A}_c(V^-, \Omega^+, N; e_Q, \beta) < 0.$$

Use both single-model terms for the separate condition. For fixed-dual DRO the condition is $\mu_Y^+ + \rho\eta + \mathcal{B}_{R_\eta}(\sigma_Y^+, N; e, \beta) < 0$. The default $\beta = (1-p)/[6(J+1)]$ is conservative for every method. The population robust-envelope slack is not estimated away by this construction: only a negative upper envelope can make the inequality feasible. No $V = 0$ exception is triggered by finite observations alone.

E.4 Global KL majorization and the finite-sample power witness

The set $\{(x, y) : D_{\mathrm{KL}}(x\|y) \le L/N,\ y \ge x\}$ is convex, so its upper boundary $y(x)$ is concave. The physical variance clipping and nonnegative square root preserve concavity and monotonicity. Thus

$f_N(q) = \sqrt{\rho \min\{V_{\max}, H_c(q; N, e_Q)\}}$ is concave. For an interior differentiability point q_0,

$$f_N(q) \le f_N(q_0) + f'_N(q_0)(q - q_0), \qquad 0 \le q \le c.$$

At a kink any supporting supergradient works. The implementation uses a flat supporting line when clipping is active. Away from the cap, put $x = q_0/c$, $y = H_c(q_0; N, e_Q)/c$. Implicit differentiation gives

$$\frac{dy}{dx} = \frac{[\mathrm{logit}(y) - \mathrm{logit}(x)]y(1-y)}{y - x}, \qquad f'_N(q_0) = \frac{\sqrt{\rho}}{2\sqrt{H_c(q_0; N, e_Q)}} \frac{dy}{dx}.$$

Tangency points are floored at a fixed positive constant; they are not reselected on final data. Numerical tests check global domination across the full admissible q range, both span choices and multiple budgets.

The paired endpoint's upper majorant is $\bar{Y} + b_N + \sqrt{2t/N} s_M + 7R_M t/[3(N-1)]$ with $Y = M + f'_N(q_0)Q$. The separate endpoint uses $Y = M_\Delta + a_C Q_C + a_F Q_F$ and two marginal sample-SD terms. Generic fixed-dual DRO already has this form with one SD term and no tangency error. Conditional on the pilot all coefficients are fixed. Sample-SD concentration (Maurer & Pontil, 2009) gives, with failure β_s per SD term,

$$s_j \le \sigma_j + R_j \sqrt{2 \log(1/\beta_s)/(N-1)}.$$

Berry–Esseen gives

$$\Pr\left\{\bar{Y} - \mu_Y \le z\sigma_Y/\sqrt{N}\right\} \ge \Phi(z) - \frac{.56\mathbb{E}|Y - \mu_Y|^3}{\sigma_Y^3 \sqrt{N}}.$$

The constant .56 is conservatively larger than the available IID constant (Shevtsova, 2011). On the pilot coverage event, substitution of the upper numerator and lower variance bounds is valid. Choose $z = \Phi^{-1}(q_N) \ge 0$. A union bound over the SD failures then yields power at least $1 - (1-p)/J$ whenever (E.1) is negative. A further union bound over J candidates gives family power at least p. Neither independence of the SD events, nor independence between them and the mean event, is assumed.

E.5 Proof of Theorem 2

Safety. Condition on any pilot realization. Its selected menu entry, including budget, K, ledger and dual parameter, is now fixed. The fresh final block has the assumed IID law. Each selected endpoint retains its own valid simultaneous candidate ledger, so the probability of falsely certifying any candidate is at most δ. This holds even if the pilot confidence set failed. On the infeasible branch no certificate is issued. Averaging over pilot realizations preserves safety.

Planning reliability and final power. Let E_r denote simultaneous pilot coverage. It has probability at least $1 - \delta_{\mathrm{plan}}$. On E_r, every menu entry that passes either power witness is genuinely sufficient; selecting among such entries does not require another union bound over final sampling events. Only one selected entry is executed on fresh observations. Theorem 1 or the preceding Berry–Esseen argument supplies its conditional power. Therefore

$$\Pr(\text{certify all}) \ge p \Pr(E_r \cap \{\text{feasible}\}).$$

This explicitly separates safety failure δ, planning failure δ_{plan}, and final failure $1 - p$. The experiment's frequent infeasibility on law B is compatible with the theorem and prevents an unconditional target-power claim there.

Oracle comparison and excess planning cost. Let $F_h(\vartheta)$ be the population sufficient-witness value for menu entry h, where ϑ includes all moments and the limiting design tangencies. For a BE entry also require $q_h(\vartheta) < 1$ and a strictly positive variance denominator. Define N_{wit} as the minimum budget of a feasible population entry under exactly these witnesses, support bounds, ledgers, menu and tie-breakers. This oracle is not the normal forecast N_{Gauss} or the minimum actual power budget N_{power}.

As $r_0, r_1 \to \infty$, the bounded-variable pilot intervals shrink and their centers converge. For a fixed finite menu and positive variance denominators, witness continuity gives uniform convergence of

their estimated values. If no limiting decision is on a boundary, the signs and BE eligibility are eventually correct with probability tending to one. The selected feasible budget thus equals N_{wit} with probability tending to one, proving the ratio statement. No claim is made about an ever-expanding menu without further uniformity conditions.

Locally, if an interpolated population witness has an isolated root N_* with $|\partial_N F| \geq c_* > 0$, a uniform perturbation at most ϵ_r shifts that root by at most $O(\epsilon_r/c_*)$. On a neighborhood with nonvanishing variances the confidence intervals give $\epsilon_r = O_p(\sqrt{\log(F/\delta_{\mathrm{plan}})/r_1} + r_0^{-1/2})$. A discrete grid adds its rounding error. The constant may be very large near a zero robust margin or variance denominator. This is a same-witness local stability result, not practical near-oracle efficiency on the reported small pilots.

E.6 Why the oracle limit does not imply total-cost efficiency

The limit in Theorem 2 holds for the final audit budget. At a fixed law N_{wit} is finite; sending $r \to \infty$ therefore makes $(r + \widehat{N})/N_{\mathrm{wit}}$ diverge. This is not an accounting detail: a regular confidence-interval planner faces a direct obstruction in a small-margin regime.

Consider a local subclass with known nuisance and variance, unknown negative envelope margin $\gamma \downarrow 0$, and oracle budget $N_{\mathrm{wit}}(\gamma) \sim C\sigma^2/\gamma^2$ for fixed $C > 0$. Suppose the estimation pilot satisfies the local CLT $\sqrt{r}(\widehat{\gamma} - \gamma)/\sigma \Rightarrow Z \sim \mathcal{N}(0,1)$, and uses the ordinary lower margin $\gamma_L = \widehat{\gamma} - c\sigma/\sqrt{r} + o_p(\sigma/\sqrt{r})$ with fixed $c > 0$. Ignore grid rounding and set $\widehat{N} = C\sigma^2/\gamma_L^2$ when $\gamma_L > 0$, abstaining otherwise. Then, whenever $\sigma/(\gamma\sqrt{r})$ remains bounded,

$$\frac{\gamma_L}{\gamma} = 1 + \frac{\sigma}{\gamma\sqrt{r}}(Z_r - c) + o_p\left(\frac{\sigma}{\gamma\sqrt{r}}\right), \qquad Z_r \Rightarrow \mathcal{N}(0,1).$$

If this ratio of scales has a positive limiting value, the budget ratio has a nondegenerate limit and a positive abstention probability; it cannot converge in probability to one. If the scale diverges, feasibility tends to $1 - \Phi(c) < 1$, which likewise precludes that convergence. Hence final-block near-oracle convergence in this regular construction requires $\sigma/(\gamma\sqrt{r}) \to 0$, implying

$$\frac{r}{N_{\mathrm{wit}}(\gamma)} \sim \frac{r\gamma^2}{C\sigma^2} \longrightarrow \infty.$$

Thus the familiar pilot-CI plug-in route cannot simultaneously obtain this strong final-block oracle limit and negligible pilot cost in the local unknown-margin subclass. This is a statement about the specified regular construction, not an impossibility theorem for every adaptive auditor. It does not exclude finite constant-factor planning, informative external margin knowledge, or procedures that validly reuse pilot information. It explains why we report total costs and do not advertise the oracle convergence statement as practical near-optimality.

E.7 Proof of the local information benchmark

Consider the family whose only unknown parameter is $A_i \sim \mathrm{Bernoulli}(\theta)$. Conditional distributions of common factual noise and all nuisance siblings given A_i are the same under θ_0 and θ_1. Enlarge the observed transcript to include A_i itself; data processing cannot make this enlarged experiment less informative. Let T count every anchor observed, including both pilot blocks and any final block. It is a bounded stopping time in the implemented finite-menu procedure; an integrable version requires the usual likelihood integrability conditions. The stopped likelihood chain rule gives

$$D_{\mathrm{KL}}(P_{\theta_1}^{\mathrm{transcript}} \| P_{\theta_0}^{\mathrm{transcript}}) \leq \mathbb{E}_{\theta_1} T D_{\mathrm{KL}}(\theta_1 \| \theta_0).$$

If D is the event of issuing a negative certificate, safety at the null gives $P_{\theta_0}(D) \leq \delta$ and unconditional alternative power gives $P_{\theta_1}(D) \geq q > \delta$. Binary data processing and monotonicity of binary KL give $D_{\mathrm{KL}}(q\|\delta) \leq \mathbb{E}T D_{\mathrm{KL}}(\theta_1\|\theta_0)$, proving (10). This is the usual change-of-measure argument (Kaufmann et al., 2016); no global minimax assertion is needed. Pilot observations are not treated as free side information.

For a differentiable robust target with nonzero derivative, an interior null θ_0, and alternatives approaching that null with $\Psi(\theta_1) = -\gamma$, the bound is asymptotic to

$$\frac{2\theta_0(1-\theta_0)[\Psi'(\theta_0)]^2 D_{\mathrm{KL}}(q\|\delta)}{\gamma^2}.$$

It is local to this one-dimensional subfamily. It does not lower-bound the nuisance calibration information of every possible unknown nuisance law. A large ratio to this benchmark can reflect both confidence construction and the greater generality of the auditor; it cannot all be attributed to algorithmic inefficiency under identical knowledge.

The pilot itself has an information requirement. A related lower bound applies to any uniformly reliable planner in this subfamily, not only the confidence-interval construction. Assume the final certificate is conditionally safe for every pilot history, and $p > \delta$. At the null, no returned plan can actually have conditional certification power at least p: that would contradict safety. Thus planning reliability forces $P_{\theta_0}(\text{feasible}) \le \delta_{\text{plan}}$. If the alternative requires $P_{\theta_1}(\text{feasible}) \ge 1 - \eta > \delta_{\text{plan}}$, then the feasible-plan indicator is itself a pilot-only test. Binary data processing gives, for a fixed pilot budget r,

$$r \;\ge\; \frac{D_{\mathrm{KL}}(1-\eta\|\delta_{\text{plan}})}{D_{\mathrm{KL}}(\theta_1\|\theta_0)}. \tag{E.2}$$

The same stopped-transcript argument handles an integrable adaptive pilot count in expectation. In the regular small-margin subfamily this is an $\Omega(\gamma^{-2})$ pilot requirement at fixed reliability and feasibility levels. Hence a vanishingly small pilot relative to that local information scale cannot both return plans reliably and guarantee their power uniformly. This necessary condition does not explain every constant in our conservative implementation, nor prove that its budgets are optimal.

F Controlled Audit Experiments

F.1 Independent-pilot planning

Let $A \sim \text{Bernoulli}(\theta)$, $W \sim \text{Unif}\{-1,1\}$, and $S \sim \text{Unif}\{-\sqrt{3/2}, 0, \sqrt{3/2}\}$ independently. Define

$$f = .45 + wW, \qquad c = f - d + a(A - .5) + b\sqrt{1 + h(2A-1)}S, \qquad d = \gamma + \sqrt{\rho}\, b.$$

The three frozen parameter tuples (a, w, b, h, γ) are $(.22, .15, .035, 0, .035)$, $(.30, .12, .065, .5, .025)$, and $(.16, .18, .095, -.4, .020)$. All losses are in $[0,1]$. At $\theta = .5$ the paired and separate population envelopes equal $-\gamma$. The chi-square optimizer is nonnegative on these laws and the intervening local-null path, so this is also the robust target. Generally

$$\Psi(\theta) = -d + a(\theta - .5) + \sqrt{\rho}\, b\sqrt{1 + h(2\theta - 1)}.$$

The local null solves $\Psi(\theta_0) = 0$; the alternative is .5.

The frozen menu has $K = 2, 8$, duals $.1, .25, .5$, mean-ledger fractions $.35, .65$, and 32 rounded budgets $256(1.25)^i$, $i = 0, \ldots, 31$. The objective minimizes fresh anchors, then evaluations, then a fixed lexical tie-breaker. Pilot cost is fixed and charged separately. For paired endpoints $e_M = \delta a_M / J$ and $e_Q = \delta(1 - a_M)/J$; separate endpoints divide these by $J + 1$ instead of J. Generic pilot-selected duals use $e = \delta/J$. The pilot uses the same budget for every law; it is never enlarged after an infeasible result. The saved budget lock includes all moment diagnostics, chosen allocations, K, estimator, and which witness passed.

All 144 pilot plans were locked before the final sampling command. Final means and covariance matrices are obtained from exact multinomial counts of the finite anchor–sibling response support. Thus each of the 256 replications at a locked plan represents exactly N fresh IID anchors, even though no redundant individual rows need be stored. Different pilot/final/benchmark RNG streams are frozen and disjoint. An independent 1024-replication empirical benchmark retains every budget and configuration. Binomial intervals are Bonferroni simultaneous across all its configurations and budgets within each law. Inverting lower and upper power bands gives a confidence set for the first menu-grid crossing, without assuming power monotonicity. The sets at $(80\%, 90\%)$ are: law A $[1907, 1907]$, $[1907, 1907]$; law B $[2980, 2980]$, $[2980, 3725]$; law C $[2384, 2980]$, $[2980, 2980]$. A collapsed interval identifies a grid point, not a continuous optimum.

F.2 Costs, failures, and the limits of the observed result

The same-witness population oracle needs 2980 anchors on A and 4657 on B/C at both target powers. These are smaller than the corresponding Theorem 1 sufficient budgets, but remain larger than the

Gaussian forecasts. The pilot medians in Table 1 are much larger; no empirical near-oracle planning result is obtained. All 109 feasible plans use the BE witness. There are 35 infeasible pilots, with most on law B.

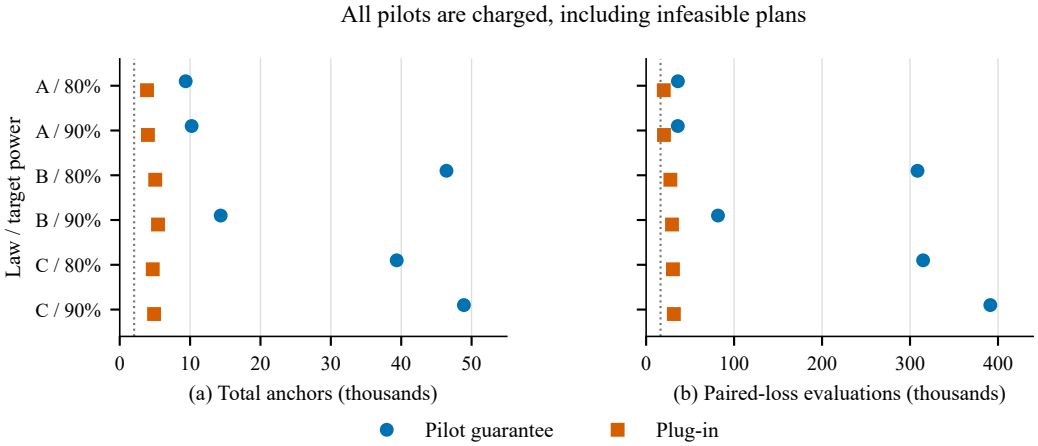


Figure 3: All pilot anchors and evaluations are charged, including infeasible branches. Points are means over the 24 pilot realizations; dotted lines mark the pilot cost. Aborted plans collect no final block. A lower cost caused by more abstentions is not an efficiency gain at fixed power. A paired-loss evaluation means two model-loss values.

Table 2: Local benchmark at nominal unconditional power $q = (1 - \delta_{\text{plan}})p$. Ratios use Monte Carlo mean total costs; they are descriptive, not confidence bounds on expected-cost ratios. In particular law B does not achieve the nominal objective after abstentions, so its ratios cannot be advertised as attained efficiency.

Law / p	Local lower bound	Mean total anchors	Ratio	Paired evaluations
A / .8	42.3	9,373	221.5	36,128
A / .9	51.4	10,218	198.8	35,999
B / .8	168.2	46,430	276.1	308,513
B / .9	204.3	14,344	70.2	81,668
C / .8	62.3	39,359	631.6	314,870
C / .9	75.7	48,906	646.1	391,251

The broad gap is not a claim of a matching lower bound: the benchmark knows a one-parameter conditional family, whereas the valid planner protects unknown bounded moments. Remaining research must address the cost of learning a sufficiently tight negative margin from a small pilot. The present experiment supplies a mathematically valid planning baseline and a direct negative test of its practical near-oracle appeal. The existing real-model ReCoRD decision remains a prospective *certification* example, not prospective *planning* validation.

F.3 Matched pairing and anchor-reuse factorial

The manifest freezes $\rho = .05$, method-wise $\delta = .025$, $J = 3$, and all grids before simulation. Independent signs A, W, H_0 and an independent uniform nuisance S on L equally spaced, centered, unit-variance states produce

$$u = (1 + hH_0)/\sqrt{1 + h^2}, \qquad d_k = .025 + \sqrt{\rho}kb,$$
$$f = .5 + aA - \tfrac{1}{2}buS,$$
$$c_k = .5 - d_k + a(rA + \sqrt{1 - r^2}W) + (k - \tfrac{1}{2})buS, \qquad k \in \{.75, 1, 1.25\}.$$

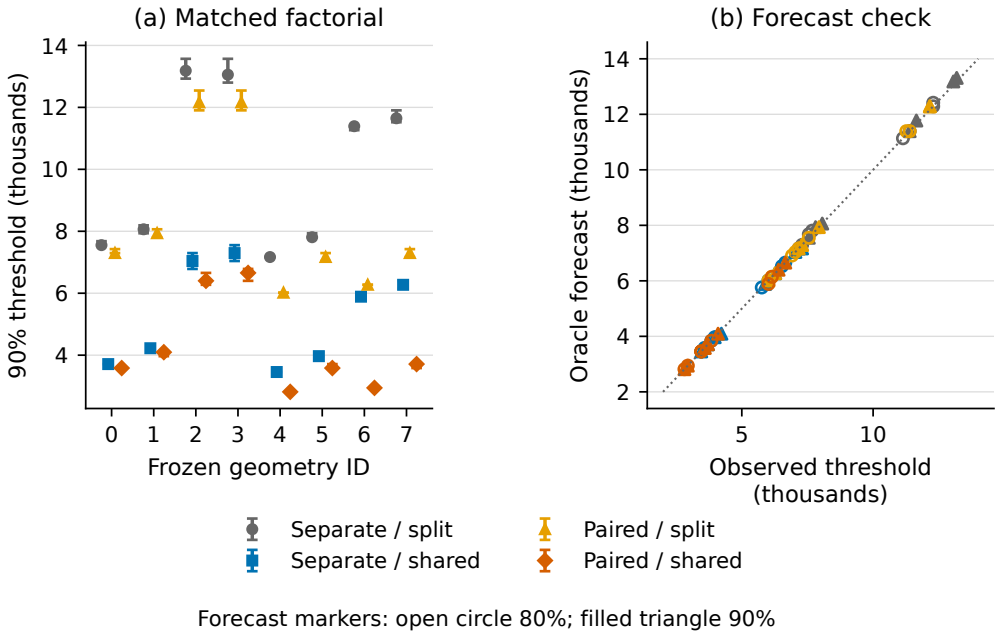


Figure 4: Left: complete matched factorial at 90% power, with 95% threshold confidence sets simultaneous over 13 methods and all 125 frozen budgets within each geometry, not across geometries. IDs enumerate correlation, anchor scale, and nuisance regime. Right: all 64 parameter-only oracle forecasts (eight cells, four methods, two powers) versus confirmed thresholds. No coefficient is fitted to outcomes; the oracle calculation is not a finite-sample guarantee.

Let $c_h = 2h/(1+h^2)$ and $\kappa_{4,S} = \mathbb{E}S^4$. For each paired candidate the population components are explicitly

$$\sigma_A^2 = 2a^2(1-r), \quad V = k^2b^2, \quad \Omega_K = (kb)^4\left[c_h^2 + \frac{(1+c_h^2)(\kappa_{4,S}+1)}{K}\right], \quad \mathrm{Cov}(M,Q) = 0.$$

Symmetry removes conditional skew and independent H_0 removes mean–variance covariance here; the theorem and implementation do not assume these cancellations in general. Support cardinality itself does not enter the sampled generic certificate, although the distribution's moments do. All losses are in $[0,1]$. Both population envelopes equal $-.025$: the paired nuisance standard deviation is kb, equal to the sum of single-model standard deviations. The optimizing density is nonnegative throughout this grid, so this is also the true robust margin.

A matched factorial with resolved thresholds. The matched factorial uses eight cells specified by design parameters: $r \in \{-.5, .98\}$, $a \in \{.04, .12\}$, and nuisance regimes $(b,h) = (.04, 0), (.12, .8)$. All three candidates, $\rho = .05$, $\delta = .025$, support size 64, $K = 2$, and robust margin $-.025$ are unchanged. Paired and separate population envelopes coincide. The four methods use identical EB-mean/KL-sensitivity operators and $e = \delta/[2(J+1)]$; split roles use $N/4$ for sensitivity and $3N/4$ for means, while shared roles reuse all N. Thus an optimizer cannot silently change the allocation within a factorial contrast. For the middle candidate, $Q_C = Q_F = Q_\Delta/4$: KL sensitivity radii and linear range terms match exactly, so pairing changes only the mean standard-deviation term.

The frozen 128-spaced grid runs from 512 to 16,384 anchors. Planning (256 replications) locks transition regions; independent confirmation (1600 replications) retains every frozen point. Exact binomial bands are simultaneous over all methods and budgets within a geometry. Inverting them gives confidence sets for the *first grid crossing*, without assuming monotone power. Collapsed sets identify a grid index, not a continuous optimum. Both 80%/90% thresholds and evaluation costs are released.

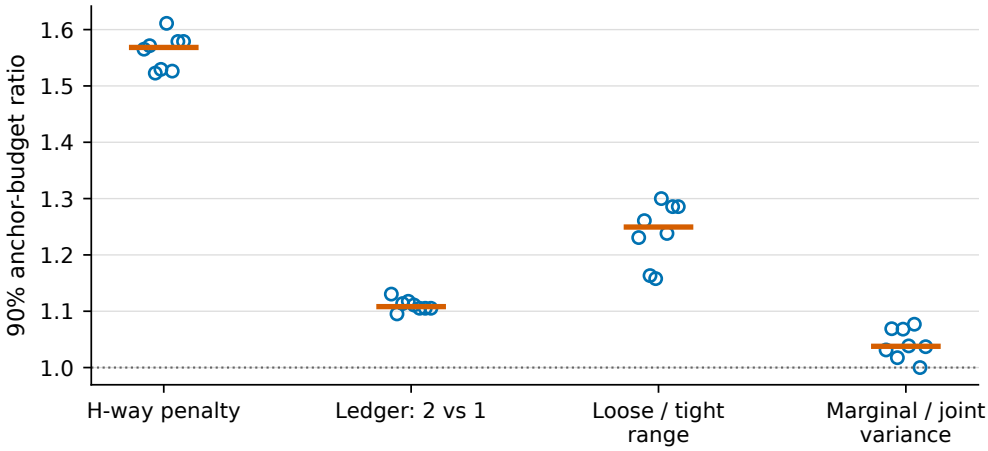


Figure 5: One-component fixed-dual contrasts at 90% power. Each open point is a design cell; short horizontal bars are medians. The same η, observations, and all other endpoint terms are retained within each contrast. Ratios describe point thresholds; full simultaneous uncertainty is released.

Pairing reduces budgets along *both* paths, but by different factors. A symmetric allocation of total log saving gives pairing factors 1.02–1.91 and reuse factors 1.81–2.10 at 90%. The log interaction ranges from $-.096$ to approximately zero; it must not be assumed absent. Appendix D.8 gives both path ratios, interaction intervals, and the exact additive log decomposition. Parameter-only finite-N oracle forecasts agree within 3.04% of all 64 confirmed grid thresholds. Rigorous sufficient budgets are larger and reported separately. This supports predictive accounting within the frozen loss laws, not universal calibration of the Gaussian approximation.

The residual is not automatically a joint-estimation gain. With finer budgets, shared-moment and grid-DRO reverse ordering by nuisance regime: point thresholds favor moment in four cells and grid-DRO in four at 90%; some contrast intervals overlap. To identify the source of this difference, we freeze $\eta = .25$ and change one component at a time. The *same fixed dual* with the H-way ledger requires 1.52–1.61 times the anchors of its single-dual ledger. A loose rather than tight valid response span costs 1.16–1.30; separate rather than joint EB variance aggregation costs only 1.00–1.08 when ledger and range are matched. The fully specified ablations, including moment and grid-DRO endpoints, fixed-dual DRO, and matched-confidence DRO, are in the appendix. The Hoeffding version remains censored above 16,384 in every cell. Changing fixed-dual choice also changes population slack q_η; that change is never labeled a pure multiplicity or joint-estimation effect.

F.4 Frozen design, confidence bands, and path independence

The eight geometries are: $r \in \{-.5, .98\}$, $a \in \{.04, .12\}$, and the nuisance regimes $(b, h) = (.04, 0), (.12, .8)$, with support size 64 and $K = 2$. These cells are selected by design parameters, independently of confirmation outcomes. All three candidates, $\rho = .05$, $\delta = .025$, and robust margin .025 are unchanged. The primary factorial uses identical empirical-Bernstein mean/KL-sensitivity operators and $e = \delta/[2(J + 1)]$. Split roles use $m = N/4, n = 3N/4$; shared roles use all N for both. Fixing this allocation avoids conflating factorial effects with different allocation optimizers. This experiment does not claim that $1/4$ is always an optimal split.

A pilot at $N = 8192$ and six implementation tests, including direct raw-loss versus sufficient-statistic agreement for all four methods, precede the refinement. The frozen grid is $N = 512, 640, \ldots, 16384$. Planning (256 replications) locks transition regions before independent confirmation (1600 replications). Confirmation evaluates the entire frozen grid, including points outside those regions, to avoid

assuming monotone power or missing an earlier crossing. It never tunes a method, grid, or law. For each geometry, exact binomial intervals with Bonferroni allocation over all 13 methods and all 125 grid points form a simultaneous 95% band. The lower endpoint of a threshold confidence set is the first grid point whose power upper bound reaches the target; the upper endpoint is the first whose lower bound does so. These contain the first crossing without a monotonicity assumption. A missing upper crossing is right-censored. A collapsed interval identifies a grid index, not an exact continuous threshold; the 128-anchor resolution still applies. Budget ratios inherit conservative intervals from this simultaneous band. Bands are per geometry, not simultaneous across all eight geometries.

Let $N_{SS}, N_{SH}, N_{PS}, N_{PH}$ be separate-split, separate-shared, paired-split, and paired-shared thresholds. Report both pairing contrasts $N_{SS}/N_{PS}, N_{SH}/N_{PH}$ and both reuse contrasts $N_{SS}/N_{SH}, N_{PS}/N_{PH}$. The log interaction is

$$\mathcal{I} = \log \frac{N_{SS} N_{PH}}{N_{PS} N_{SH}}.$$

A symmetric, path-independent allocation of total log saving uses

$$A_P = \tfrac{1}{2} \log \frac{N_{SS} N_{SH}}{N_{PS} N_{PH}}, \qquad A_R = \tfrac{1}{2} \log \frac{N_{SS} N_{PS}}{N_{SH} N_{PH}}, \quad A_P + A_R = \log \frac{N_{SS}}{N_{PH}}.$$

This is descriptive factorial accounting, not a causal claim that changes to confidence construction are biological or training interventions.

The 64 parameter-only forecasts (8 cells, 4 factorial methods, 2 powers) differ from measured grid thresholds by less than 3.04%; their median ratio is 1.000. Rigorous sufficient shared budgets range from 5968 to 15792 anchors, or 2.10–2.60 times measured thresholds. These conservative guarantees are distinct from the sharper oracle approximation. The maximum observed boundary-null family error frequency is 0.0000; zero counts are not proof of zero risk. All 13 methods and all frozen budgets are retained.

F.5 Residual estimator ablations

All fixed-dual comparisons use $\eta = .25$, declared before outcomes; there is no tuning block or tuning cost. The original 31-point grid remains unchanged. Comparing a grid optimum with this fixed point changes both selection and multiplicity; that comparison alone is *not* a pure ledger effect. We additionally evaluate the *same fixed point* with $e = \delta/(JH)$ versus δ/J to isolate the H penalty, and with $e = \delta/(2J)$ to match the moment method's event allocation. With fixed η and matched e, compare empirical Bernstein to Hoeffding on the same bounded Y_η to isolate variance adaptation. Compare the valid loose span $R_{\mathrm{sum}} = 2B + 2B^2/(4\eta)$ to the tighter $R_\eta = B + \max\{B, B^2/(2\eta)\}$, changing no other component.

For an exact variance-aggregation control, sum two marginal EB endpoints for M and $Q/(4\eta)$, each at the same $e = \delta/(2J)$, without clipping. The mean, dual term, and linear range penalty are then identical to a joint EB endpoint with span R_{sum}; their difference is exactly

$$\sqrt{2\log(2/e)/N} \left\{ s_M + \frac{s_Q}{4\eta} - s_{M+Q/(4\eta)} \right\} \geq 0.$$

This controlled contrast isolates variance aggregation, including covariance, by the sample-standard-deviation triangle inequality. It establishes no unique DRO geometry. Replacing only the marginal Q confidence construction by KL (with its physical cap) quantifies a distinct sensitivity-bound effect. The physical cap does not confound that comparison here: throughout the frozen laws and grid, the worst-case EB sensitivity upper endpoint is below 0.335 and the KL upper endpoint below 0.367, both below the cap one. These deterministic bounds use the maximum possible sample Q, not observed outcomes. Linearizing the current moment endpoint at the fixed dual also introduces q_η; it is not a pure joint-estimation comparison.

H penalty: anchor-cost ratio 1.523–1.611 (median 1.568). Ledger 2 vs 1: anchor-cost ratio 1.095–1.130 (median 1.108). Loose vs tight range: anchor-cost ratio 1.158–1.300 (median 1.249). Marginal vs joint variance: anchor-cost ratio 1.000–1.077 (median 1.038). The fixed-dual Hoeffding version reaches neither target within the frozen budget in any cell; it remains right-censored rather than triggering grid expansion. KL versus EB sensitivity construction changes rank by nuisance regime. None of these contrasts establishes a universal estimator ordering.

Table 3: Dense-grid thresholds and simultaneous confidence sets. IDs enumerate correlation, anchor scale, then nuisance regime, in ascending order. All use $K = 2$; paired loss evaluations are twice the anchor count.

ID	Separate/split	Separate/shared	Paired/split	Paired/shared
		80% power		
0	7296 [7296, 7424]	3584 [3584, 3584]	7168 [7040, 7168]	3456 [3456, 3456]
1	7680 [7680, 7808]	3968 [3840, 4096]	7552 [7424, 7680]	3840 [3840, 3968]
2	12288 [12160, 12672]	6528 [6400, 6656]	11392 [11136, 11648]	6016 [5760, 6144]
3	12288 [11904, 12544]	6656 [6528, 6912]	11264 [11136, 11648]	6144 [5888, 6272]
4	7168 [7040, 7168]	3456 [3456, 3456]	6016 [6016, 6016]	2816 [2816, 2816]
5	7552 [7552, 7680]	3840 [3840, 3968]	6912 [6784, 7040]	3456 [3456, 3456]
6	11136 [11136, 11264]	5760 [5760, 5760]	6144 [6144, 6144]	2944 [2944, 2944]
7	11392 [11264, 11520]	6016 [6016, 6144]	7040 [6912, 7040]	3456 [3456, 3584]
		90% power		
0	7552 [7552, 7680]	3712 [3712, 3712]	7296 [7296, 7424]	3584 [3584, 3584]
1	8064 [7936, 8192]	4224 [4096, 4224]	7936 [7808, 8064]	4096 [3968, 4096]
2	13184 [12928, 13568]	7040 [6784, 7296]	12160 [11904, 12544]	6400 [6272, 6656]
3	13056 [12800, 13568]	7296 [7040, 7552]	12160 [11904, 12544]	6656 [6400, 6784]
4	7168 [7168, 7168]	3456 [3456, 3456]	6016 [6016, 6016]	2816 [2816, 2816]
5	7808 [7808, 7936]	3968 [3968, 4096]	7168 [7040, 7296]	3584 [3584, 3712]
6	11392 [11264, 11392]	5888 [5888, 5888]	6272 [6272, 6272]	2944 [2944, 2944]
7	11648 [11520, 11904]	6272 [6144, 6400]	7296 [7168, 7424]	3712 [3584, 3712]

Table 4: Both paths of the factorial decomposition at 90% power; ratios and confidence intervals for both power targets are retained in machine-readable results.

ID	Pair/split	Pair/shared	Reuse/separate	Reuse/paired	Log interaction
0	1.035	1.036	2.034	2.036	-0.001
1	1.016	1.031	1.909	1.938	-0.015
2	1.084	1.100	1.873	1.900	-0.014
3	1.074	1.096	1.789	1.827	-0.021
4	1.191	1.227	2.074	2.136	-0.030
5	1.089	1.107	1.968	2.000	-0.016
6	1.816	2.000	1.935	2.130	-0.096
7	1.596	1.690	1.857	1.966	-0.057

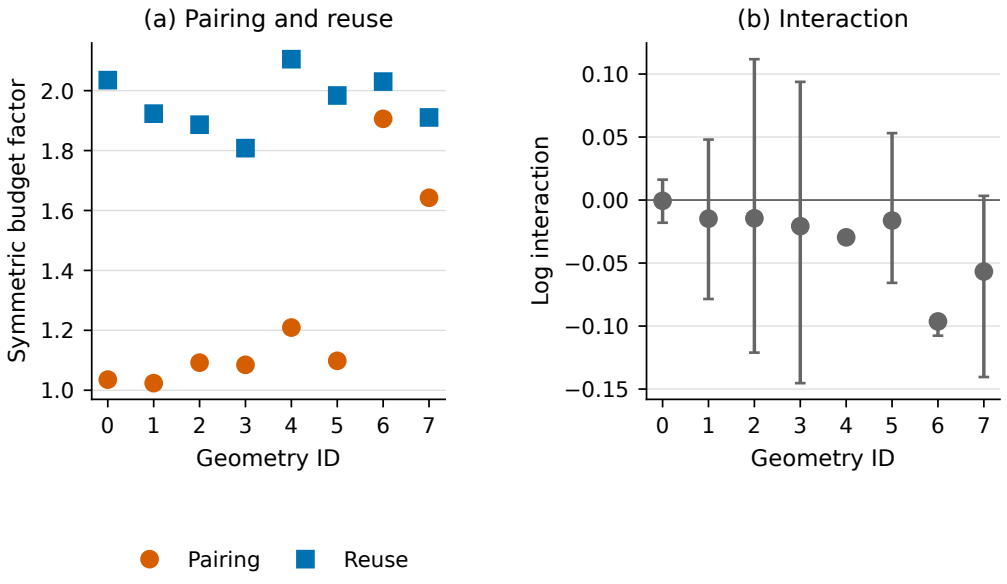
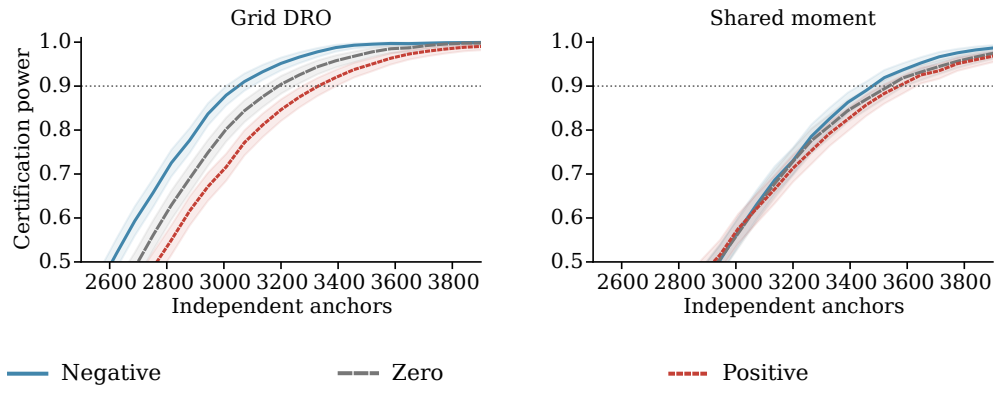


Figure 6: Symmetric pairing/reuse factors and the log interaction. Confidence sets describe discrete grid thresholds, not continuous optima.



Power axis zoom: 0.5–1.0; original bands retained.

Figure 7: Covariance stress with independent power confirmation. Bands are simultaneous across both methods, all signs, and all budgets. The generic response preserves the covariance term; moment confidence bounds treat marginal uncertainties separately. This does not imply covariance has no effect on the sampling fluctuation of the moment endpoint.

F.6 Targeted mean–variance covariance stress

Let H, W be independent uniform signs and S an independent standardized Bernoulli variable: it equals $\sqrt{.7/.3}$ with probability .3 and $-\sqrt{.3/.7}$ otherwise. Set $a = .12, b = .18, h = .8, K = 8$, $\rho = .05$, $\gamma = .025$, and

$$Z = d + tH + \sqrt{a^2 - t^2}W + \sqrt{b^2(1 + hH)}S, \qquad d = -\gamma - \sqrt{\rho}b.$$

Let $\xi = \mathbb{E}\kappa_3/K$, which is nonzero, and for $z \in \{-1, 0, 1\}$ choose $t = .08z - \xi/(b^2 h)$. Thus $\mathrm{Cov}(M, Q) = .08zb^2h$ while $\mathrm{Var}\,\mu = a^2$, $V = b^2$, Ω_K, and true robust margin are fixed. Taking factual loss .5 and candidate loss $.5 + Z$ gives bounded losses throughout. The optimizing conditional density is positive, so the true robust value is exactly $-.025$, not just its envelope. We check the variance identity at fixed $\eta_* = b/(2\sqrt{\rho})$ using 200,000 independent anchors per setting. Empirical variances are .016326, .019012, .021518, versus theoretical .016363, .018939, .021514 for negative, zero, and positive covariance, respectively; discrepancies are below .4%. The finite-budget experiment below instead uses the frozen 31-point dual grid.

Covariance and certification power. The three skewed laws have negative, zero, and positive $\mathrm{Cov}(M, Q)$, equal anchor variance, and equal robust margin. Power is evaluated at 64-anchor spacing with 400 planning and 3000 independent confirmation replications. Generic-DRO 90% thresholds are 3072, 3200, and 3392, with simultaneous sets $[3008, 3136]$, $[3200, 3264]$, and $[3328, 3456]$. The positive/negative difference is 10.4%; all three sets are disjoint. Shared-moment sets overlap, so no corresponding three-way ordering is claimed.

Full threshold intervals. The three laws share anchor variance and robust margin, with nonzero skewness. The budget grid is $N = 1536, 1600, \ldots, 4096$, 400 planning replications, and 3000 independent confirmation replications. The three laws match the stated marginal means and variances but need not match every higher joint moment; the power ordering is evidence consistent with the variance prediction, not a distribution-free causal isolation of covariance. The transition neighborhoods are locked before confirmation; all frozen points are retained. Exact binomial 95% bands are simultaneous across all six method/sign curves and all 41 budgets.

For generic DRO the 90% negative and positive covariance thresholds differ by 320 anchors (10.4% of the negative threshold). All three 90% confidence sets are disjoint. Shared-moment sets overlap substantially: no analogous three-way ordering is asserted for that method. The reported separation concerns the 64-anchor grid.

F.7 Exact-enumeration efficiency and allocation

The exact-enumeration comparison freezes $\rho = .05$, $\delta = .025$, three candidates, all budget grids, and all evaluation rules before simulation. A subsequent uniformly refined anchor-budget grid changes no candidate, radius, class, or method. Independent anchors A and nuisance S are uniform signs. The factual loss is $f = .55 + .30A + .03S$; candidate losses are $c_h = .55 - d_h + .31A - .03S$, $d_h \in \{.030, .025, .020\}$. All losses lie in $[0, 1]$. Enumerating both styles gives $g_h = -d_h + .01A - .06S$, with paired anchor variance .0001; the separate anchor variances are .09 and .0961. Crucially, the paired and separate *population envelopes coincide*:

$$E_h^{\text{pair}} = E_h^{\text{sep}} = -d_h + .06\sqrt{.05} \in \{-.01658, -.01158, -.00658\}.$$

Both approaches are therefore asymptotically capable of certification. Equal envelopes isolate finite-sample cost from population slack. Section F.9 separately studies policy selection when that slack differs.

Three procedures share the same finite observations and conditional chi-square class: our paired KL certificate, separate-model KL certification, and a generic paired DRO baseline. The latter tunes a nonnegative-density DRO dual variable on half the calibration block, calibrates the variance of its dual loss on the other half, and bounds its expectation on the final audit block. It directly estimates the robust functional without separately upper-bounding sensitivity (Appendix F.7). Each method controls its three-candidate family at .025; these are equal method-wise budgets, not simultaneous guarantees across all three procedures. No population variance or sensitivity is passed to an auditor.

The paired KL certificate uses 10.3–17.2 times fewer anchors than separate-model certification on the refined grid. Generic paired DRO needs .48–.67 times as many as paired KL. Thus the central empirical claim is the information benefit of pairing, not optimality of our particular moment certificate. The leading covariance ratio in Corollary D.1 equals 3721 here, but is not a finite-budget prediction: calibration and bounded-range terms dominate long before that regime. Every reported upper bound covered its known robust target in the simulations; this is a coverage check, not proof of zero failure. All three candidates are beneficial in this isolation study; harmful-policy selection remains tested separately in the finite-calibration decision study.

F.8 Calibration, anchors, and siblings are distinct resources

A second frozen factorial study independently varies $m \in \{1024, 4096, 16384, 65536\}$, $n \in \{256, 1024, 4096, 16384\}$, and $K \in \{2, 8, 32\}$, with 300 replications per cell. Here siblings are genuinely conditionally IID. Let $f = .50 + .15A + .06S$ and $g = -.05 + aA - .12S$. The sibling-dominated and anchor-dominated settings use respectively $a = .002$ and .12; both have $\sigma_S^2 = .0144$ and the same negative population envelope $-.02317$.

At $m = 4096, n = 1024$, increasing K from 2 to 32 raises power from .067 to .993 in the sibling-dominated case, but only from .010 to .020 in the anchor-dominated case. Allocation matters even at fixed total anchors: in the sibling-dominated setting with $K = 32$, $(m, n) = (4096, 1024)$ has power .993, whereas $(1024, 4096)$ has power zero. The cheapest grid point with Wilson lower power above .9 uses 5120 anchors and 163,840 paired view evaluations in the first setting; minimizing view evaluations instead chooses 8192 anchors, $K = 2$, and 16,384 evaluations. Thus more siblings can save anchors while spending more compute. Full grids and descriptive budget frontiers, including all zero-power cells, are released; no allocation is revised after testing.

Exact observations, without oracle calibration. For the primary experiment, each sampled anchor yields both nuisance views. Their weighted mean is an affine function of the anchor sign, and their conditional variance is the squared style coefficient, constant across anchors in this population. Counting the four sign combinations of successive calibration anchors is therefore an exact sufficient-statistic implementation of the finite loss-block procedure. It supplies the empirical Q and D means to the same KL inversion, not their population expectations in place of confidence bounds. Binomial counts likewise give the actual sampled audit mean. Numerical checks compare this implementation with explicit loss arrays and the original scalar KL operator, including the exact all-zero endpoint. Known population envelopes are used only for post-decision scoring.

For each method the candidate family is simultaneous at .025. Paired KL allocates $e = \delta/(4J)$ to its variance calibration, sensitivity calibration, and two audit-mean events. Separate KL allocates

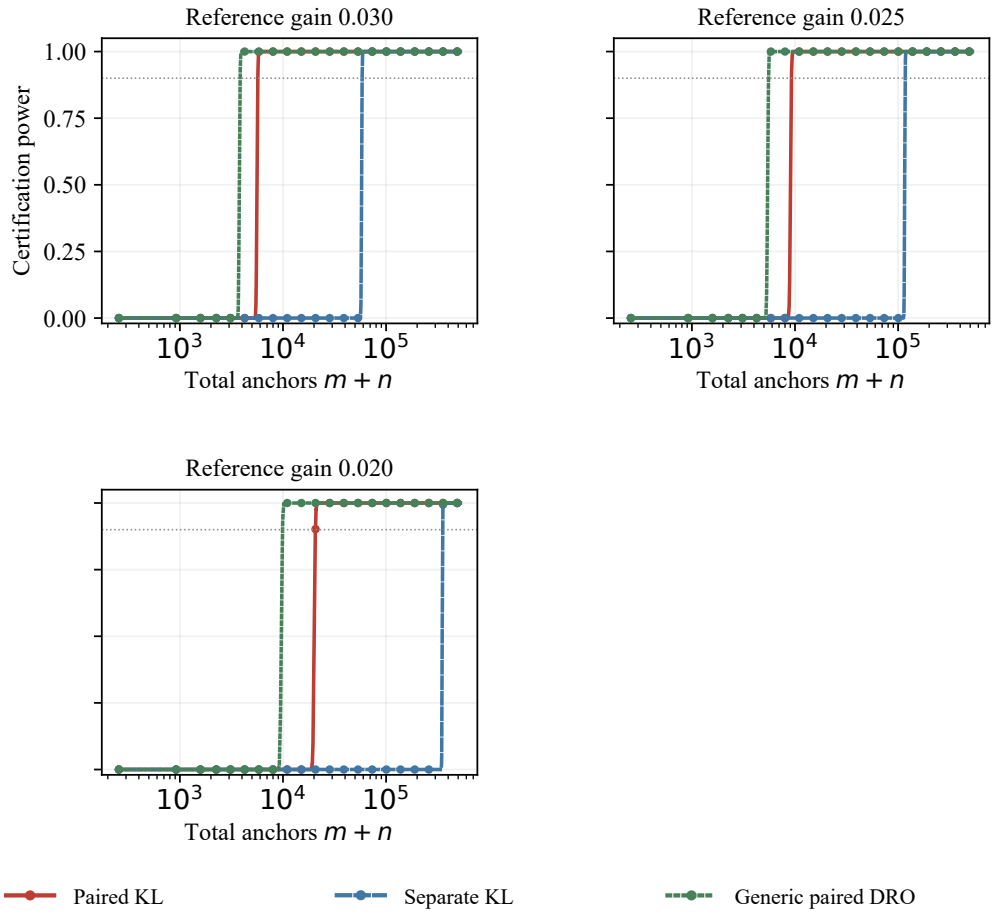


Figure 8: Certification power versus total independent anchors, $m = n$; 1,000 replications per refined grid point. Shading is a 95% Wilson Monte Carlo interval. Population paired and separate envelopes are equal and negative for every candidate. Both paired estimators certify earlier; generic paired DRO is more efficient than the explicit sensitivity certificate.

$e = \delta/[4(J+1)]$, reusing the one factual-model bound across candidates. This is stronger than unnecessarily replicating the factual failure cost. Generic paired DRO uses $e = \delta/(4J)$; three events suffice, so one event allowance is unused. The comparison uses the paired certificate without the optional triangle improvement to avoid silently borrowing unbudgeted marginal calibration events. All corrections are zero by the synthetic data design.

Generic paired conditional DRO baseline. For each fixed dual value $\eta > 0$, define the anchor response

$$\phi_\eta(g(T,\cdot)) = \sup_{r \geq 0,\ \mathbb{E}_\nu r = 1} \{\mathbb{E}_\nu[rg(T,S)] - \eta \mathbb{E}_\nu[(r-1)^2]\}.$$

Weak duality for the same average conditional chi-square class gives $\Psi(\rho) \leq \eta\rho + \mathbb{E}_T \phi_\eta$. Since the reference density is feasible and its penalty is zero, $\mathbb{E}_\nu g \leq \phi_\eta \leq \max_S g$, hence $\phi_\eta \in [-1, 1]$. With two equally weighted nuisance states, let $M = (g_+ + g_-)/2$ and $b = |g_+ - g_-|/2$. Solving the bounded scalar quadratic gives

$$\phi_\eta = M + \begin{cases} b^2/(4\eta), & b \leq 2\eta, \\ b - \eta, & b > 2\eta. \end{cases}$$

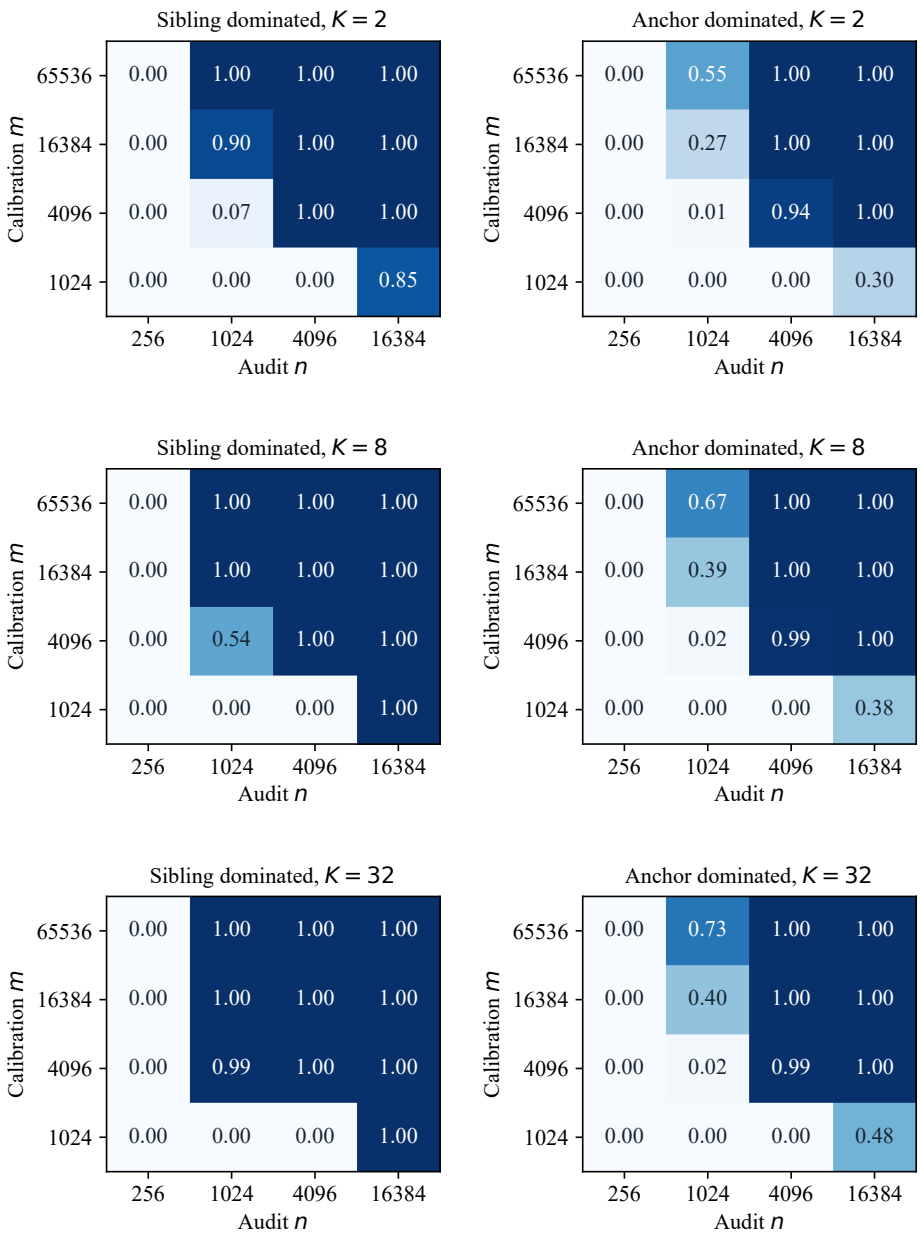


Figure 9: Paired certification power with independently varied m, n, K. Rows use $K = 2, 8, 32$; columns show the sibling-dominated and anchor-dominated regimes. Within each panel, the horizontal axis is audit anchors n and the vertical axis is calibration anchors m. Numbers are empirical powers, not upper confidence bounds. Calibration shortage, audit shortage, and the anchor floor have distinct signatures.

The first $m/2$ calibration anchors choose η to minimize $\eta\rho + \overline{\phi}_\eta$ over 80 geometric grid values from .001 to 2. The remaining $m/2$ anchors calibrate the variance of that chosen bounded response with independent disjoint anchor pairs. The final n anchors estimate its expectation. Conditional on dual tuning, the chosen response is fixed and independent of both later blocks; no 80-way testing penalty is needed. The generic upper bound is its Bernstein/Hoeffding mean upper endpoint plus $\eta\rho$. No test outcomes choose η.

This estimator targets the robust functional directly and enjoys paired cancellation without estimating V separately. Its finite dual grid can add upward approximation slack; the released table reports that slack relative to the known synthetic optimum. The comparison therefore does not handicap the baseline by assigning it the separate-model envelope. The power comparison below includes this generic estimator.

Power and thresholds. The initial primary $m = n$ grid has 23 sizes from 128 to 262,144 per block, with 1000 Monte Carlo replications each. Its 80% and 90% thresholds coincide because the grid is coarse. We therefore froze a *budget-only resolution extension* after seeing those curves: 2001 equally log-spaced sizes across the entire original range, rounded to multiples of four and deduplicated (1679 distinct sizes), 1000 replications each, and seed 2309202633. The separate manifest records this chronology. Candidate models, ρ, shift class, loss, confidence budgets, and methods are unchanged. This refinement improves numerical resolution; it is not an independent confirmatory model experiment.

Power is the fraction of negative upper endpoints for each candidate. The threshold is the first grid size attaining the target and maintaining it at all larger tested sizes. Both empirical-power and pointwise 95% Wilson-lower thresholds are released. The latter differ by less than 0.5% in anchor counts on the refined grid. Wilson intervals quantify Monte Carlo precision, not simultaneous confidence for searching threshold crossings. All per-size powers remain available; no trials or failed candidates are removed. The coarse results are preserved: their 80%/90% total-anchor thresholds were $(6144, 65536, 4096)$, $(12288, 131072, 6144)$, and $(24576, 393216, 12288)$ for paired, separate, and generic paired methods.

Independent resource grid. In the allocation study there are $2 \times 4 \times 4 \times 3 = 96$ frozen cells and 300 replications each. Within a cell, every calibration anchor and final-audit anchor is a separate independent task draw. Common random blocks across cells improve comparison precision but are not counted as additional replications. Sibling sums and disjoint-pair variance observations are drawn by exact multinomial sufficient statistics of K IID signs. Calibration and audit remain disjoint even when different cells reuse simulation randomness. Population $\tau_K^2 = a^2 + .0144/K$ makes the floor explicit.

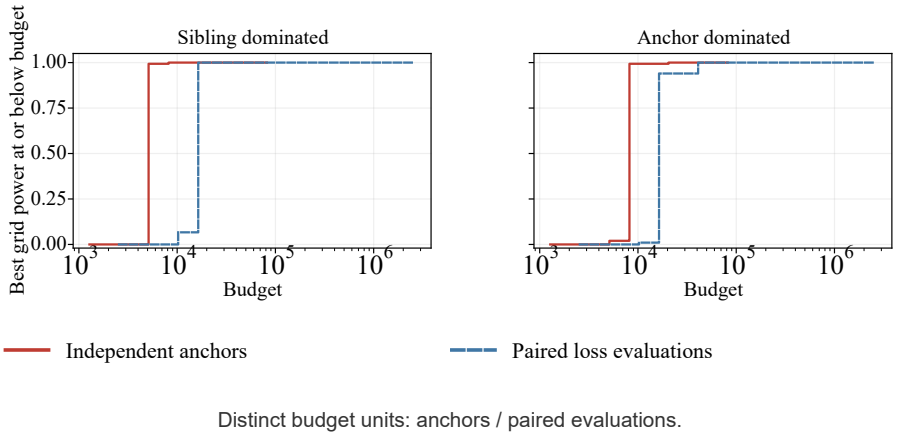


Figure 10: Descriptive frontiers over the full frozen allocation grid. An anchor and a paired view evaluation are different units: more siblings can reduce the first cost while increasing the second. These frontiers do not select an audit allocation after viewing a real test.

F.9 Finite-calibration policy selection

We freeze four policies, $\rho = .05$, family failure probability $\delta = .025$, and $m = n \in \{64, 256, 1024, 4096, 16384\}$ independent calibration and audit anchors, with $K \in \{2, 4, 16\}$

IID siblings per anchor. Each cell has 300 independent replications. Additional pre-specified cells use $J = 2$ and $\rho = .25$. The factual loss is $0.5 + 0.04A + 0.3S$, where independent A, S are uniform signs. The four paired losses are $a_h + 0.04A + b_h S$, with $(a_h, b_h) \in \{(-.08, .02), (-.09, -.45), (-.01, .06), (.02, .01)\}$. Every factual and candidate loss is in $[0, 1]$. The exact robust difference is $a_h + \sqrt{\rho}|b_h|$ for these radii, which satisfy local exactness. Truth is used for scoring only. The algorithm receives finite loss blocks; it estimates nuisance sensitivities and cluster variances with the same module used for the real-model audit.

Five rules share candidates, blocks, radius, and loss: paired certification, separate-model certification, a fully bounded fallback, reference-only mean selection, and a paired reference-validation interval. All three robust methods share the explicit joint event ledger of Proposition D.2; validation rules use the same observations but have no shift guarantee. Selection occurs before scoring by population risk. The released results report beneficial non-factual selection, harmful selection, candidate power, factual fallback, best-safe selection, realized robust difference, certificate width, and simultaneous coverage.

At $n = 1024, K = 4$, paired auditing selects a beneficial policy in $300/300$ replications with zero harmful selections. Separate and bounded auditing always fall back. Reference-only selection is harmful in $265/300$ replications; paired reference-validation is harmful in $17/300$. The paired/separate/bounded upper bounds cover all candidates in every replication of this cell. Zero observed failures in 300 trials is not proof of zero failure probability; its 95% Wilson upper limit is approximately 1.26%. At $n \leq 256$, paired auditing also abstains. This is a constructive cancellation example, not universal dominance. In particular, separate-model failure includes irreducible envelope slack, not solely a worse statistical rate.

F.10 Analytic mechanism checks

These constructions check the comparative-utility statements in settings where every population quantity is known. These are analytic and finite-sample checks, not evidence that an end-to-end transformer optimizer satisfies the assumptions. All numerical values below follow from the stated laws.

F.10.1 A loss-aligned covered-shift witness

Take one task state, $S \in \{-1, +1\}$, and the uniform reference law. Let $g(S) = m + \alpha S$ and let the covered deployment law be $\pi_\delta(S = s) = (1 + \delta s)/2$ for $|\delta| \leq 1$. Then

$$\Delta_\nu = m, \qquad V_\nu^\Delta = \alpha^2, \qquad \bar{\chi}^2_{\pi_\delta} = \delta^2, \qquad \Delta_{\pi_\delta} = m + \alpha\delta.$$

Consequently,

$$|\Delta_{\pi_\delta} - \Delta_\nu| = |\alpha\delta| = \sqrt{\bar{\chi}^2_{\pi_\delta} V_\nu^\Delta}.$$

Panel (a) plots $m = -0.12$ and $\alpha = -0.58$. The line lies on the corresponding endpoint of the certificate interval, so the coefficient one in Theorem B.3 is not a proof artifact. Choosing $|m| + |\alpha| \leq B$ realizes the signed difference as two bounded losses, as shown in the proof.

F.10.2 Local exactness for an arbitrary fixed pair

The local-exactness construction is not restricted to the binary witness. For any fixed pair, let $h = g - \bar{g}$, $V = V_\nu^\Delta$, and $M = \operatorname{ess\,sup} |h|$. For a prescribed $\rho \leq V/M^2$, we evaluate the two density ratios

$$a_\pm = 1 \pm \sqrt{\rho/V}\, h.$$

They are nonnegative, conditionally normalized, and have global chi-square radius exactly ρ. The resulting utility differences are $\Delta_\nu \pm \sqrt{\rho V}$, matching both endpoints of the theorem. This calculation checks the fixed-pair exactness proposition pointwise over the same bounded response function used to define the certificate; it is not a end-to-end training experiment.

F.10.3 WHY PAIRED UTILITY CAN BE TIGHTER THAN TWO SEPARATE BOUNDS

Let the centered reference responses of the two models have unit L^2 norm and angle ϑ. The separate-model radius is always 2, whereas the paired response has norm

$$\|h_C - h_F\|_2 = \sqrt{2 - 2\cos\vartheta}.$$

Thus identical marginal sensitivities can yield a zero paired radius under perfect cancellation and the full separate radius under opposition. Panel (b) plots this exact family. It is a direct numerical check of (B.6), and also explains why estimating two marginal certificates and adding them can hide useful paired information.

F.10.4 FINITE-SAMPLE ANCHOR–SIBLING DECOMPOSITION

For an evaluation difference Z_{ij} with anchor variance $\sigma_A^2 = \mathrm{Var}(\mathbb{E}[Z_{ij} \mid C_i])$ and conditional sibling variance $\sigma_S^2 = \mathbb{E}\,\mathrm{Var}(Z_{ij} \mid C_i)$, the exact variance is

$$\mathrm{Var}(\widehat{\Delta}) = \frac{\sigma_A^2}{n} + \frac{\sigma_S^2}{nK}.$$

Panel (c) uses $n = 128$, $\sigma_A^2 = 0.64$, and $\sigma_S^2 = 0.36$. The values approach $0.64/128$ rather than zero. This is the comparative version of the anchor floor in Theorem B.1; it also explains why the finite certificate in Proposition B.4 contains two radius terms.

F.11 FINITE-SAMPLE INTERPRETATION

The generated manifest records the exact bounded-loss construction used for a finite-sample Monte Carlo sanity check. In that check the two predictors are held fixed, paired losses are evaluated on each fresh anchor and sibling, and the empirical mean and standard deviation are compared with (B.11). This is a finite evaluation of the certificate's probability mode, not an empirical-ERM or language-model optimization result. Keeping the predictors fixed is intentional: it matches the fresh-evaluation conditioning in Assumption A.4 and avoids silently turning a post-training diagnostic into a theorem about model selection.

G PROSPECTIVE RECORD CERTIFICATION

G.1 PROSPECTIVE FINITE-CATALOG PROTOCOL

The ReCoRD protocol is frozen before checkpoint outcomes. The target law is uniform over the eligible article catalog, with one deterministic QA per article, rather than an assumed IID sample of all future news. Eligibility uses source data and tokenizer lengths only: unique normalized passage, lowest-index QA, exactly one placeholder, a valid extractive answer, and all four prompts at most 1024 tokens. Known SQuAD train/development and retrieval contexts are excluded by normalized hashes. The resulting 65,255 eligible articles supply a uniform random disjoint partition of 5,000 calibration, 10,000 audit, and 2,000 sealed test articles (seed 23092643).

Freeze the existing 0.5B Qwen learner, seed-4101, 1000-step factual checkpoint and four QA checkpoints (RGF/Rule, 5000/15000 CF samples), $\rho = .05$, $\delta = .025$, primary loss $1 - \mathrm{F1}$, four equally weighted content-preserving prompt formats, greedy decoding with 32 new tokens, and minimum-negative-endpoint selection with factual fallback. No models are trained or selected using ReCoRD. Dataset release membership does not imply absence from foundation model pretraining; the independence claim concerns this project's auditing and candidate-selection history.

Sampling is without replacement from a finite catalog. We therefore use bounded-MGF comparison, not the IID empirical-variance bound, to justify the prospective endpoints (Bardenet & Maillard, 2015). For paired-KL let $e = \delta/(2J)$, with $J = 4$;

$$U_{\mathrm{PK}} = \bar{M}_{\mathrm{audit}} + \sqrt{2\log(1/e)/n} + \sqrt{\rho H_1(Q_{\mathrm{cal}}; e)}.$$

Q is exactly enumerated over the four styles. For generic DRO, use the same frozen 80-element η grid as the retrospective study and all $m + n$ pre-test articles:

$$U_{\mathrm{GD}} = \min_{\eta}\{\eta\rho + \bar{\phi}_\eta + \sqrt{2\log(JH/\delta)/(m+n)}\}.$$

Both statements follow because the bounded exponential-moment comparison for simple random sampling is no larger than for sampling with replacement. The KL Chernoff inversion and Hoeffding upper tail thus remain valid marginally; their union bound does not require independence of the disjoint roles. No calibration-law variance is transported to a dependent audit block. Each method has a .025 family guarantee, not a joint .025 guarantee across methods.

The decision and its hash are saved before generating test predictions. Test F1/EM and the test-catalog envelope are descriptive held-out outcomes, not exact full-catalog risks. All candidates and negative decisions are retained.

Prospective outcomes.

Policy	Paired upper	Generic upper	Test F1	Test EM
Factual	–	–	8.06	1.81
RGF 5k	0.0096	0.0069	16.08	10.31
Rule 5k	0.0514	0.0471	10.73	5.60
RGF 15k	0.0004	-0.0022	16.44	11.19
Rule 15k	0.0404	0.0360	12.07	6.10

F1/EM are percentages. Every endpoint uses the frozen method-wise family budget .025 and radius .05. The saved decision hash precedes all test predictions.

Why the prospective endpoints differ. For RGF-15k the saved paired upper endpoint is $+.0004275$, whereas the generic endpoint is $-.0021798$. On the pooled pre-test catalog the reference difference is $-.08437$, the moment envelope is $-.0376894$, and the exact robust target is $-.0376938$. These descriptive catalog quantities are not unknown full-catalog truth. The endpoint gap decomposes into an audit-versus-pooled mean difference of $-.001280$, a paired-versus-generic mean-radius difference of $-.001544$, a sensitivity-calibration radius excess of $+.005431$, and a moment-envelope-versus-generic-grid difference of approximately $+.0000004$. Their sum is $+.002607$. This decomposes the original split-versus-pooled comparison, not an inherent limitation of explicit moments. The shared-anchor post-decision ablation (Appendix D.1) reverses the failure and gives a tighter endpoint than finite-catalog DRO. In the original comparison, non-local envelope slack is negligible in these observed records. The comparison uses the frozen without-replacement Hoeffding mean bounds, not the IID empirical-Bernstein estimators of the controlled study. It shows a prospective safe selection by generic paired auditing, not an explicit-KL advantage or a new power estimate for the IID experiment.

The experiment uses 340,000 logical model–view evaluations, excluding interrupted or repeated execution work. The selected checkpoint achieves 16.44 test F1 versus 8.06 for factual training, with a descriptive test-catalog moment envelope of $-.03649$. These low absolute scores reflect transfer of fixed SQuAD checkpoints to cloze-style news QA; they are not competitive-benchmark claims. EM is secondary and is not certified by the F1 endpoint.

ReCoRD: attribution, not another prospective method. All post-decision endpoints use only the 15,000 pre-test articles. With a common ledger, RGF-15k endpoints are separate/split $+.02256$, separate/shared $+.01507$, paired/split $+.00118$, and paired/shared $-.00631$. The original saved grid-DRO endpoint is $-.00218$ and its prospective decision remains unchanged. Pairing alone or reuse alone does not change the matched separate/split endpoint's sign; their combination does. These Hoeffding mean radii depend only on range, so paired and summed separate radii are equal: pairing's observed tightening comes from sensitivity, not covariance-adaptive mean estimation. This distinction is essential when interpreting a real-model result.

The three generic procedures remain distinct: sampled-sibling DRO uses IID anchor EB; retrospective finite-view HANS/QA DRO enumerates nuisance views before anchor EB; prospective finite-catalog ReCoRD DRO enumerates views with without-replacement Hoeffding/MGF control. ReCoRD does not validate the IID sampled-sibling proposition. HANS/QA audit failures are reported in Section H.12.

G.2 RECORD ATTRIBUTION ONLY

These are post-decision methodological ablations, never new prospective methods. Only the original 15,000 pre-test records are read. The prospective decision is fixed before these ablations. The four factorial endpoints use identical Hoeffding means, KL sensitivities, and $e = .025/[2(4+1)]$; the generic column retains its original grid ledger. The latter difference prevents attributing its residual entirely to joint estimation. Because the mean radii are range-based here, the separate radii sum to exactly the paired radius: no covariance-dependent *mean* gain is realized by these Hoeffding operators. Pairing tightens sensitivity terms and potentially population slack; reuse tightens both estimation roles. This is not the same mechanism as a variance-adaptive mean audit. No sealed-test outcome enters any endpoint, and the new method comparisons add no prospective-success count.

G.3 RETROSPECTIVE FINITE-VIEW COMPARISON

The same inspected calibration and audit predictions supply both procedures. Generic DRO also receives a same-audit-only comparison in the released records; its primary endpoint uses all available anchors. For each η, compute the exact conditional dual response

$$\phi_\eta(g_C) = \inf_t \{t + \eta + \mathbb{E}_\nu(g_C - t)_+^2/(4\eta)\}.$$

The minimizing t solves $\mathbb{E}_\nu(g_C - t)_+ = 2\eta$. This dual response lies in $[-1, 1]$. Apply anchor empirical Bernstein to ϕ_η over a frozen 80-point grid $[.001, 2]$, adding $\eta\rho$, and use a simultaneous grid and candidate ledger. Both procedures retain the predeclared cross-family budget $.025/24$ per family; no inspected outcome is used to change it. These are retrospective calculations under their stated anchor assumptions, not new independent trials or prospective coverage measurements.

The exact robust functional of the inspected pooled catalog is a diagnostic, not the unknown population target. Of 116 records, 80 catalog targets are nonnegative and 36 are negative but their endpoints remain positive. For the width-limited HANS, HANS-extension, and token-Brier QA records, median expected-moment generic-DRO total-anchor forecasts are respectively 16,518, 13,867, and 17,556.5 (ranges 7,951–328,375, 7,845–324,016, and 7,588–284,235). These substitute empirical moments into a sufficient endpoint; they are neither powered prospective sample-size guarantees nor lower bounds on necessary information. Median generic empirical-Bernstein linear terms are $.1274, .1257, .0479$, versus median total statistical widths $.1429, .1431, .0596$. Removing KL calibration therefore improves the endpoint but does not remove small-margin and bounded-range costs. It is not scientifically justified to attribute every failure solely to missing anchors or solely to KL inversion.

H LANGUAGE-MODEL DIAGNOSTICS AND SCOPE

H.1 EXPLORATORY TRAINING-VOLUME STUDY

Figure 1 reports a separate training-volume study on MultiNLI (Williams et al., 2018), using the `nyu-mll/multi_nli` dataset. We retain entailment (label 0) and contradiction (label 2), excluding neutral examples. All 16 pools share a locked split of 500 factual training examples, 400 newly drawn development examples, and 400 exploratory test examples. The exploratory test set reuses a development set from a preceding exploratory experiment. It is therefore not an independent confirmatory set, and the pools do not provide 16 independently sampled test sets.

Each pool consists of a counterfactual generation library and its associated training runs. The collection contains one initial experiment, seven repeats, and eight second-batch repeats. Generation and training seeds vary across pools. Qwen2.5-1.5B-Instruct generates light, natural English rewrites, and Qwen2.5-0.5B-Instruct serves as the LoRA-tuned classifier. For each $n \in \{100, 300, \ldots, 1900\}$, training uses all 500 factual examples and the first n examples from the pool's generated library. Thus n denotes generated training examples, not independent audit anchors.

Every n condition restarts from the same pretrained classifier checkpoint, rather than continuing from a smaller-n run. Training samples uniformly from the combined factual and generated set for four epochs, with effective batch size 20, LoRA rank 8, LoRA scaling $\alpha = 16$, and learning rate 2×10^{-4}. Within a pool, the training seed is the same for every n. Fixed epochs do not match optimization

steps across dataset sizes: increasing n changes the generated-data fraction and training compute together. The curves describe this training protocol rather than an isolated effect of sample count at fixed compute.

The panels report exploratory-test Brier risk and classification error, source-wise predictive KL in nats, and audited label-disagreement rates. The pale blue curves show individual pools, the red curves their means, and the bands pointwise 95% Student-t confidence intervals across the 16 pools, conditional on the shared split. These bands summarize generation and training variation, not uncertainty from resampling test examples. The predictive KL comparison uses the model's outputs on an original example and a generated edit for $n \leq 500$, but the outputs on two generated edits for $n > 500$. We interpret the $n > 500$ trend within that comparison. The audited rates are descriptive estimates from cached judgments.

H.2 Task and common protocol

We use MultiNLI (Williams et al., 2018). In the source-tag setting, the premise–hypothesis pair is the task state and an appended source cue is the nuisance. Factual training has 95% cue–label agreement, the clean counterfactual condition presents both cues, and the held-out cue–label agreement is 5%. This is an exact cue-reversal construction, not an official genre shift. In the free-text arm, Qwen2.5-1.5B-Instruct rewrites the hypothesis into one of five screened styles while inheriting the original label. We refer to these as automatically screened rewrites rather than semantically certified counterfactuals.

Unless stated otherwise, training uses one BF16 epoch, batch size two, gradient accumulation four, maximum length 384, learning rate 2×10^{-4}, and rank-8 LoRA on `q_proj` and `v_proj`. The main grid uses seeds 2027, 2028, and 2029 on an 8GB RTX 5060 Laptop GPU. All final diagnostic pools are disjoint from training anchors.

H.3 Operational proxies and estimands

The following table makes explicit which population object is and is not observable in the Qwen arm.

H.4 Paired utility audit on locked Qwen predictions

We compute the paired quantities required by Proposition B.4 from the locked crossed-consistency predictions. For every pool and style sibling, the factual $W0$ model and candidate $Wlow$ or $Wstar$ model are evaluated on the same example and shared label using the bounded action loss $1-p_\theta(y \mid x)$. The audit therefore records paired loss differences rather than reconstructing them from marginal risks. The calibration block has 400 anchors and the independent test block has 800 anchors, each with four screened style siblings, across eight locked pools. We report the fully observable bounded radius in (B.13) with $\rho = 0.04$. Because these are automatically screened rewrites, the semantic correction is unknown and is not set to zero.

The mean calibration–test summaries are $\widehat{\Delta} = 0.0016$ and 0.0057 for $Wlow$ and $Wstar$, while the independent-test paired gaps are 0.0023 and 0.0073. The corresponding test anchor components are 0.0015 and 0.0037. Both unadjusted upper certificates remain positive; we therefore report "not certified" rather than converting reduced predictive sensitivity into a robustness claim.

H.5 Cluster scaling and fixed synthetic budget

For each of 24 fitted checkpoints, fresh evaluation uses 80 independent anchors and four sibling edits per anchor. We estimate the variance within each checkpoint by 1,200 cluster resamples per $(n, K) \in \{16, 32, 64\} \times \{1, 2, 4\}$ cell, then summarize these estimates across checkpoints. A least-squares fit to the nine cell means gives RMSE 3.30×10^{-5} for $a/n + b/(nK)$, compared with 5.97×10^{-4} for a sibling-only $c/(nK)$ fit. The nonzero fitted anchor component is the important observation: resampling siblings reduces conditional intervention noise but does not make the anchor uncertainty disappear.

We also evaluate a descriptive fixed-row budget with $nK = 1000$ and $(n, K) \in \{(1000, 1), (500, 2), (250, 4), (125, 8)\}$. These are separate retrains, so the comparison changes

both anchor diversity and sibling diversity; it is not interpreted as a causal estimate of K. The results are included as a budget diagnostic rather than as evidence for a monotone test accuracy law.

H.6 Covered-transfer intervention

We use eight disjoint 400-anchor training pools and match text and token budgets across the intervention conditions. Weighted augmentation changes predictive KL from 0.1704 to 0.1023 (40.0%; one-sided upper bound -0.04840). The held-out style gap changes from 0.003753 to 0.003289, but the one-sided upper bound is $+0.000726$. The reference-risk change has upper bound $+0.012962$, which does not meet the pre-specified 0.01 non-inferiority margin.

The correct conclusion is therefore narrow: consistency can be manipulated to reduce a predictive sensitivity proxy, but this run does not establish a cost-free improvement in realized style stability. The result is consistent with the theorem's role as a transfer certificate. It does not license the reverse implication "KL decreases, therefore deployment risk decreases."

H.7 Matched-error mechanism study

At matched 20% corruption, eight paired LoRA pools compare a cue-reinforcing intervention with a cue-opposing intervention. The mean action-risk difference is 0.0878 with paired 95% interval $[0.0307, 0.1449]$; seven of the eight pool-level differences are positive. This result is useful because the nominal error rate is held fixed while the direction of the induced training signal changes. It is a mechanism stress test outside the exact assumptions of the comparative theorem, not a claim that the theorem predicts the complete nonlinear Qwen optimization path.

Finite-pool transfer slack. A separate retrospective diagnostic uses 1600 frozen Qwen-feature anchors with two source-cue views. After the frozen centering, projection and scaling, let $x_{i,s}$ include an intercept. The reference head w averages every thirteenth head in the archived, ordered predictor family, starting at the first. Let q be the normalized mean development-feature difference between the two cue views, with zero intercept coordinate. For $\alpha \in \{0, .5, 1, 2\}$, use $w_\alpha = w + (\alpha - 1)q(q^\top w)$ and binary probabilities $\mathrm{sigmoid}(2x_{i,s}^\top w_\alpha)$. Thus $\alpha = 0$ removes the cue component and $\alpha = 1$ retains the reference head. The response $r_{i,s}$ is expected zero–one action loss, not argmax error. Under the uniform two-view reference, define $h_{i,s} = r_{i,s} - \bar{r}_i$ and let the test density be 0.1 on the label-aligned cue and 1.9 on the other cue. Writing u for density minus one, set

$$d_i = \mathbb{E}_\nu[u_{i,S} r_{i,S}], \quad L = \frac{1}{n}\sum_i |d_i|, \quad V = \frac{1}{n}\sum_i \mathbb{E}_\nu h_{i,S}^2.$$

The sign-coherence and magnitude-homogeneity factors are $|\bar{d}|/L$ and $L/(.9\sqrt{V})$; their product is the observed gap divided by the global head-aware bound. These are descriptive finite-pool ratios, with no additional model fitting or independent prospective prediction.

H.8 Evidence boundary

The controlled numerical appendix verifies the paired identity and its finite-sample decomposition with known quantities. Qwen shows that the interfaces are measurable, but the natural rewrite arm lacks a semantic audit and the prospective intervention does not pass all downstream gates. Broader CFPT utility claims require those additional checks.

H.9 Shift-granularity diagnostic

This locked natural-text audit uses completed predictions without retraining any checkpoint. For each candidate and factual pair, we align the same 256 test anchors and their three nuisance tags, and use $1 - p_\theta(y \mid x)$ as a bounded expected-zero-one loss. The paired difference is then summarized under three nested categorical classes: one common nuisance law, one nuisance law per target-label group, and one nuisance law per anchor. The reference law is uniform over the three tags.

The manifest records 36 candidate/task/size/seed conditions, $2{,}160$ raw class/path records, the fixed radii $\rho \in \{0.05, 0.10, 0.25, 0.50, 1.00\}$, and the prespecified mixture path $\pi_\delta = (1 - \delta)\nu + \delta e_1$ for

$\delta \in \{0, .25, .5, .75, 1\}$. At $\rho = 0.25$, the pooled global versus per-anchor certificate difference is small for the clean condition and remains small relative to the corruption effect for the 20% and 40% conditions. We therefore do not use this P0 result to select a more local class for the main paper. Its value is negative and methodological: it prevents the controlled separation result from being misrepresented as an empirical claim about the current Qwen predictions.

H.10 Failure boundaries: semantic corruption and support escape

The second boundary audit uses exactly the same locked Qwen files to show how semantic corruption changes the sign and magnitude of a candidate's realized loss. Relative to the factual baseline, the clean screened condition is near zero or slightly beneficial, whereas the 20% condition increases held-out loss by approximately 0.19–0.30 across the four task/size cells and the 40% condition increases it by approximately 0.41–0.57. These values are a diagnostic of automatically screened rewrites, not an estimate of a human-verified semantic error rate. Figure 20(a) therefore supports the narrower claim that a low nominal corruption rate is not an adequate substitute for a paired utility audit.

For support escape, let the reference law place all mass on the observed styles and let a deployment law place mass γ on a new style. If the paired loss difference on that unseen style is $u \in [-1, 1]$, the realized gap is γu, while every reference-only sensitivity statistic remains unchanged. Figure 20(b) plots this exact envelope. It is not a failure of the covered-shift theorem: it is the reason coverage is an assumption rather than a cosmetic regularity condition.

H.11 HANS: sensitivity beyond average reference gain

This experiment asks whether paired sensitivity supplies information about shift response beyond average reference gain. It does not seek a universally best CFPT optimizer. Theorem B.3 motivates measuring the nuisance variation of the *risk difference* between each policy and its factual comparator. We compare this quantity with actual risk changes under fixed nuisance reweightings. The two main panels are empirical mechanism diagnostics; they are not finite-sample certification claims.

Models, interventions, and data. The factual RoBERTa model is trained on 1,000 MultiNLI pairs and 600 original HANS pairs. CF training uses four relation-preserving lexical views per HANS anchor, constructed by consistent noun substitutions. The reference weights are $(0.325, 0.325, 0.175, 0.175)$: two aligned-cue views jointly receive probability $q_0 = 0.65$ and the two opposed-cue views receive $1 - q_0$. All policies use actual model predictions and are paired with the same-seed factual model; no prediction-level sensitivity is imposed.

The main comparison retains five policies. Paired P50 uses balanced CF cue exposure and paired-variance regularization of weight two; P995 uses aligned cue probability 0.995. Both assign 40% of supervised loss to CF examples. Cue70, Cue90, and Cue100 instead assign 70%, 90%, and 100% to aligned CF examples, with alignment probability one. Fixed displayed checkpoints are 120 updates for the first four policies and 240 for Cue100, all following the same 240-update schedule. Earlier development selected these checkpoints; they are frozen for the replication. The 60-update paired checkpoint is an ablation of the same method and is omitted from the primary display, with its results preserved in the archive.

The initial cohort comprises four training-order seeds (7201–7204), followed by eight additional seeds (7301–7308). All twelve share initialization and data, and all five policy results are retained for each seed. The evaluation catalog contains 756 label-balanced syntactic structures absent from all available earlier locked HANS datasets at construction, divided into 252 calibration, 252 reference, and 252 test anchors. Figure 21 uses reference risk differences from the reference block and paired sensitivity from calibration; Figure 22 uses the test block. These structures have now been inspected, so the repeated runs replicate training randomness on a fixed catalog, rather than provide a fresh-data confirmation. Views and seeds do not multiply the number of independent test structures.

Shift axis and its scope. For each q, aligned views receive $q/2$ each and opposed views receive $(1-q)/2$ each. The displayed risk difference is the exact weighted average of the same four measured losses. It is therefore affine in q; the evidence lies in the fitted models' measured response slopes, margins, and crossings, not in discovering a nonlinear curve. We first examined $q \in [0.5433, 0.7567]$,

corresponding to $\rho = 0.05$, and now display $q \in [0.40, 0.90]$ on a 0.05 grid while retaining the original endpoints. The chi-square divergence is

$$\chi^2(q\|q_0) = \frac{(q - 0.65)^2}{0.65 \cdot 0.35}.$$

The expanded range reaches approximately 0.2747, so it is not contained in the original $\rho = 0.05$ ball. It stays within the same four-view support. The original interval is shaded, and the reference position is marked. Expansion was motivated by the first four results and is exploratory for that cohort; its grid is fixed before the additional seeds are trained.

Observed mechanism across twelve seeds. Average reference Brier differences are -0.0191 for Paired P50 and -0.0158 for P995, while measured paired sensitivities are 0.00260 and 0.00888, respectively, a factor of approximately 3.4. These point estimates illustrate that nearby average benefits can accompany different sensitivities; they do not establish statistical equivalence of reference gains. Cue70, Cue90, and Cue100 have mean sensitivities 0.01289, 0.01940, and 0.03748. Table 12 reports the completed replication separately from the pooled descriptive summary.

In the expanded interval, Paired P50 remains below zero at every evaluated law in all twelve seeds, including all eight runs in the extension cohort. P995 does so in eleven of twelve: one seed crosses zero near $q = 0.443$, an exception included in both the figure and the table. Cue70 and Cue90 improve the test reference condition in all twelve seeds, yet both cross above zero at sufficiently low cue alignment in all twelve. Cue100 is already harmful at the test reference condition in all twelve, making it a stress case rather than the principal example of a lost reference gain.

Per-seed zero crossings lie in $q \in [0.419, 0.493]$ for Cue70, $[0.541, 0.621]$ for Cue90, and $[0.696, 0.736]$ for Cue100. At $q = 0.40$, mean risk differences are -0.0177, -0.00406, $+0.00521$, $+0.0239$, and $+0.0683$ in the five-policy order above. The mean P995 curve stays negative even though one individual seed fails, illustrating why the faint per-seed traces and cohort counts are needed in addition to averages. Crossing locations follow exactly from the measured two-mixture losses; they are descriptive estimates, not confidence limits.

Replication and limits. Eight additional training-order seeds (7301–7308) completed the frozen five-policy protocol without new CF generation or checkpoint selection. All 48 extension model tasks (eight factual models plus forty CF models) completed without failed or missing jobs in 30.2 minutes, within the 60-minute budget. The original-four, additional-eight, and combined-twelve summaries are retained separately. The main figures show the combined twelve, which are a descriptive summary of both stages, not twelve independent evaluation datasets. The added seeds repeat the controlled-mechanism pattern: Paired P50 improves every tested law in 8/8 runs, while Cue70 and Cue90 lose their test reference advantage somewhere in the expanded range in 8/8 runs.

The supported empirical claim is that paired sensitivity adds information about nuisance-shift response beyond an average reference gain. High sensitivity does not force harm: P995 remains beneficial throughout this range in eleven of twelve runs despite being more sensitive than Paired P50. The displayed laws also do not exhaust all task-conditional nuisance shifts. Formal certification remains a separate limitation: the preceding small-sample audit gave no negative paired or separate bounds among its 24 HANS model pairs, including the 60-update ablation. The finite-sample terms were too wide to certify the

H.12 HANS: WHAT CAN THE EXISTING INFORMATION CERTIFY?

We reuse every candidate from the four-seed fresh-structure study and the twelve-seed extension, including weaker and failing candidates. Calibration, reference audit, and test each contain only 252 noun-normalized structures; four nuisance views are enumerated with weights $(.325, .325, .175, .175)$. The cohorts overlap and cannot be pooled as independent observations. Each training-order seed fixes another model, not another evaluation dataset. The audit calibrates variance rather than substituting a second moment, uses explicit sensitivity upper bounds, and allocates failure probability jointly over all reported families. No new training or radius reduction is used. The sampling roles are stratified partitions of an inspected finite catalog, not newly drawn IID task blocks. These recalculated endpoints are diagnostic; the prospective IID coverage proposition is not asserted for this replay.

Table 13 separates two empirical diagnoses. If $\widehat{\Delta} + \sqrt{\rho \widehat{V}} \geq 0$, the plug-in envelope already fails before uncertainty. If it is negative but $U \geq 0$, the current certificate is width-limited at those plug-in moments. Neither diagnosis identifies the unknown population sign with certainty. We also solve the nonnegative-density finite-catalog DRO problem and report its duality gap and envelope slack; its exact target is that calibration catalog, not the unknown HANS population. Candidate-wise means, sensitivities, paired/separate/bounded endpoints and choices are reported for each cohort. None of the HANS or QA endpoints in this analysis is negative. For width-limited candidates, plug-in planning under this implementation and joint ledger requires approximately $9{,}100$ to several million anchors in *each* block, depending on the margin. These are estimated-moment forecasts, not guaranteed sample sizes. A nonnegative population envelope would not be repaired by more audit data. Replays of smaller n and IID draws of the same four views are explicitly conditional resampling of inspected data.

Sampling scope. These HANS endpoints reuse an inspected structure catalog. Additional training seeds or lexical views do not create independent task structures, so this analysis is retrospective, not a prospective power confirmation.

H.13 QA: ALIGNING THE AUDIT WITH GENERATION QUALITY

The token-Brier audit reuses 30 SQuAD adapters over two training budgets and three seeds. A compute-limited alignment check freezes the complete five-model family at $1{,}000$ updates, seed 4101: factual, RGF and rule-based CF at $5{,}000$ and $15{,}000$ examples. We generate answers on 650 calibration and 650 reference contexts under four fixed content-preserving prompt formats, producing $26{,}000$ answers. Greedy decoding uses at most 32 new tokens. Two A100s complete evaluation in 3.2 minutes, with no additional training or failed jobs.

Bounded token Brier, $1 - \mathrm{F1}$, and $1 - \mathrm{EM}$ use identical anchors, policies, $\rho = .05$, and a joint $.025$ budget divided equally across the three losses. The $1 - \mathrm{F1}$ reference differences for RGF at 5k/15k are $+.0171/+.0120$, versus $-.0362/-.0457$ for rule-based CF. Token Brier improves for all four candidates. The inspected test agrees with the F1 sign distinction, but is not a new confirmation set. No generation-aligned certificate becomes negative. Alignment changes the target and apparent reference benefit; it does not automatically solve the statistical-width problem.

H.14 A WRONG-SHIFT-CLASS BOUNDARY

A separate frozen boundary witness uses $g = -.04 + .30AS$. Its exact shared-law robust value is $-.04$, but the task-dependent class has value $-.04 + .30\sqrt{.05} = .02708$. Finite calibration of the global sensitivity and the same bounded mean procedure accepts in $300/300$ trials at $n = 4096$; the correctly specified conditional-class rule accepts in $0/300$. The global certificate covers its own target throughout but misses the larger-class target. More data cannot repair a wrong shift class. This boundary uses finite calibration rather than supplying sensitivity to the auditor.

A per-cell nuisance budget and a population-averaged conditional budget define different adversaries and must not be interchanged.

H.15 HANS: AN EXTERNAL HIGH-HEADROOM DIAGNOSTIC

The finite-state result has known truth by construction, but it does not test whether the same separation is visible in a non-synthetic language benchmark. We therefore add a fixed-protocol diagnostic using the Heuristic Analysis for NLI Systems (HANS) benchmark (McCoy et al., 2019). This experiment is deliberately scoped as a controlled mechanism test rather than as evidence that automatically generated counterfactuals preserve semantics. We fine-tune a randomly initialized binary classification head on a local RoBERTa-base checkpoint, with the same optimizer, three epochs, sequence length, and batch size in every condition, and use five independent seeds. The HANS training file supplies three balanced heuristic groups (lexical overlap, subsequence, and constituent); the held-out HANS evaluation file supplies a pre-specified, label-balanced test subset of 500 examples per group, sampled by a fixed seed rule.

The factual condition uses the locked MultiNLI anchors. The natural-CF condition adds the locked $K = 1$ automatically screened siblings. The remaining conditions add a matched HANS intervention pool of 600 examples per group: *HANS-balanced* uses the original labels, while *aligned* and *reversed* corrupt exactly 20% of the pool, with the same pool size, but flip opposite label directions. All sampling rules and corruption indices are fixed from the seed and do not depend on model outcomes. We report bounded Brier risk, NLL, accuracy, and the within-HANS heuristic-group risk profile; the bounded risk is the quantity most directly comparable to the finite-risk certificates above.

Figure 24 gives two useful pieces of evidence. First, the factual and natural-CF conditions remain at chance on HANS (mean Brier risks 0.404 and 0.401), whereas the controlled HANS intervention reaches mean Brier risk 0.008 and mean accuracy 0.990. This is a high-headroom mechanism result: a targeted intervention can change the nuisance response that the certificate is meant to track, while the natural rewrite condition does not automatically do so. Second, the two matched-rate corruption directions have different mean risk profiles (Brier 0.123 versus 0.134; NLL 0.377 versus 0.408), even though they use the same nominal error rate and pool size. The seed intervals are wide for these corruption conditions, so we treat the directional contrast as evidence of heterogeneity rather than as a universal ranking. The group panel makes the relevant unit visible: the question is not only whether average test loss moves, but which nuisance groups move together or separate.

This result has an important boundary. HANS-balanced examples and HANS test examples share a controlled template family, so the large improvement is not a claim of natural deployment robustness. Conversely, the natural-CF row is not a clean semantic-validation result because its rewrites are only automatically screened. The experiment therefore strengthens the mechanism evidence for the transfer and direction interfaces, while leaving the semantic-validity and external-validity questions open.

To probe the coverage boundary with the same trained models, we evaluate the frozen checkpoints on the development portions of Adversarial NLI (ANLI) (Nie et al., 2020). We merge the fixed R1, R2, and R3 development pools and map entailment to one and neutral/contradiction to zero before reading any model output. This is a held-out external diagnostic, not a tuning set and not a claim about a three-class NLI optimum. The external contrast is sharper than the natural-CF row alone: HANS-balanced changes HANS Brier risk from 0.404 to 0.008, but its ANLI Brier risk is 0.342 versus 0.317 for factual training, a paired increase of 0.025 (95% seed-level t-interval $[0.003, 0.048]$). Natural CF changes ANLI Brier risk by only $+0.005$ ($[-0.016, 0.027]$). The matched-rate corruption directions also separate on ANLI: reversed versus aligned corruption changes Brier risk by -0.089 ($[-0.156, -0.022]$). These are mechanism and boundary results: a gain on a covered heuristic family need not transfer to an external adversarial pool, and a nominal corruption rate does not determine the direction of the realized risk.

Table 5: All residual-estimator ablations, 90% grid thresholds. C is fixed before outcomes, not tuned using population or sampled losses. Full 80%/90% confidence sets are released.

Method	0	1	2	3	4	5	6	7
A: Moment, current ledger	3456	3840	6144	6400	2688	3456	2816	3456
B: Grid DRO	4608	3712	8192	6272	3712	3328	3840	3328
C: Fixed dual	2944	2688	5632	4352	2304	2432	2432	2432
D: Fixed dual, matched	3328	2944	6272	4864	2560	2688	2688	2688
Fixed dual, H ledger	4608	4224	8576	6656	3712	3712	3840	3840
Fixed dual, loose range	4096	3712	7296	5632	3328	3328	3456	3456
Marginal EB, fixed dual	4224	3968	7424	6016	3456	3584	3456	3584
Marginal KL, fixed dual	3584	4480	6912	7040	2816	3968	2944	4096
Fixed dual, Hoeffding	> 16384	> 16384	> 16384	> 16384	> 16384	> 16384	> 16384	> 16384

Table 6: The unchanged three-law covariance stress with dense budgets and simultaneous confidence sets.

Method	Covariance	80% threshold	90% threshold
generic	-1	2944 [2880, 2944]	3072 [3008, 3136]
generic	0	3008 [3008, 3072]	3200 [3200, 3264]
generic	1	3136 [3136, 3200]	3392 [3328, 3456]
shared	-1	3328 [3264, 3392]	3520 [3456, 3584]
shared	0	3328 [3264, 3392]	3584 [3520, 3648]
shared	1	3392 [3328, 3456]	3584 [3520, 3648]

Table 7: Total anchors $m + n$ at the first refined grid point attaining 80%/90% empirical power and retaining it at all larger tested budgets. Ratios are grid-resolved Monte Carlo estimates. Pointwise Wilson-lower thresholds are also released; they change these budgets by less than 0.5%.

Gain	Power	Paired KL	Separate KL	Generic paired DRO	Sep./paired
.030	80%	5,680	58,336	3,824	10.3
.030	90%	5,704	58,552	3,848	10.3
.025	80%	9,184	116,744	5,512	12.7
.025	90%	9,248	117,192	5,552	12.7
.020	80%	20,448	352,680	9,832	17.2
.020	90%	20,680	354,024	9,944	17.1

Table 8: Exact values underlying panel (c). The residual reduction from $K = 4$ to $K = 32$ is within-anchor noise; the nonzero limiting term is anchor uncertainty.

K	1	4	32
$\mathrm{Var}(\widehat{\Delta})$	0.0078125	0.0058594	0.0050391
σ_A^2/n	0.0050000	0.0050000	0.0050000

Table 9: Post-decision, pre-test-only ReCoRD attribution. Four moment endpoints use a common event ledger. Prospective endpoints use their original confidence allocation.

Policy	Sep./split	Sep./shared	Pair/split	Pair/shared	Original DRO
rgf_5000	+0.03223	+0.02481	+0.01033	+0.00277	+0.00690
rule_5000	+0.07675	+0.06723	+0.05211	+0.04299	+0.04706
rgf_15000	+0.02256	+0.01507	+0.00118	-0.00631	-0.00218
rule_15000	+0.06485	+0.05554	+0.04113	+0.03189	+0.03602

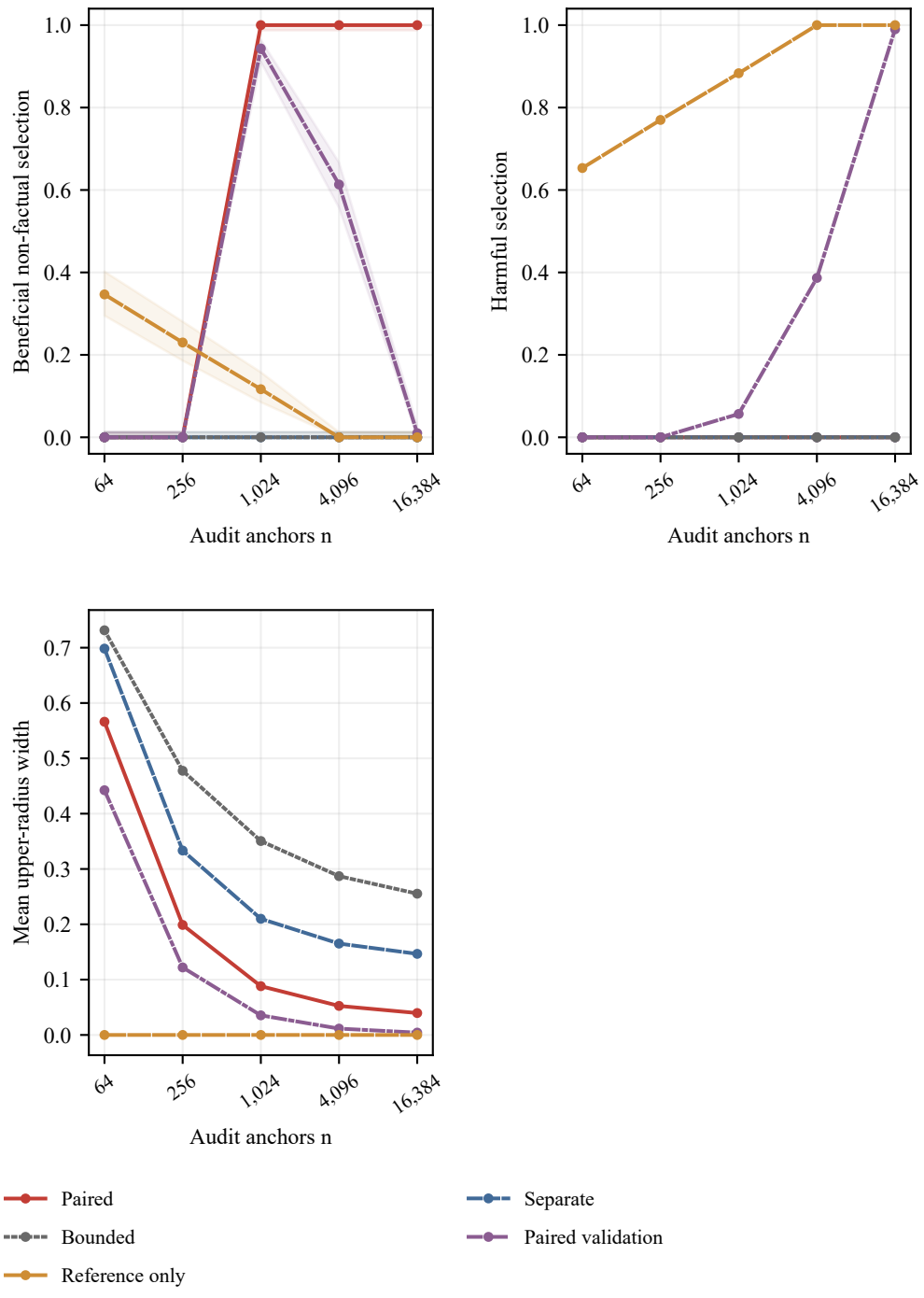
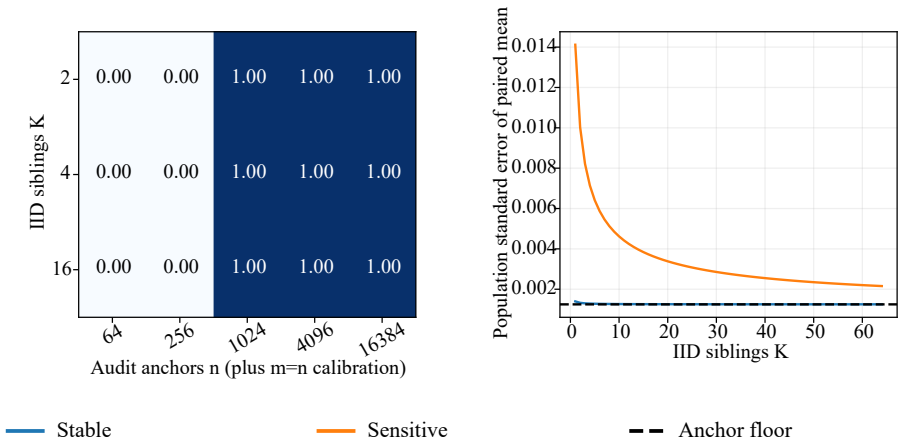


Figure 11: Finite-calibration policy selection, $J = 4, K = 4, \rho = .05$, 300 replications per budget. Calibration uses an additional $m = n$ independent anchors. Shading denotes 95% Wilson Monte Carlo intervals, not the audit certificate. Reference-only and reference-validation rules have no covered-shift guarantee. Separate and bounded rules abstain in this population because their robustness envelopes remain too wide.



Left: measured power. Right: analytic population standard error.

Figure 12: Independent anchors versus siblings. Left: measured paired safe selection across the frozen n, K cells. Right: the known population standard error $\sqrt{(0.04^2 + b_h^2/K)/n}$ at $n = 1024$. The horizontal floor cannot be removed by more siblings. The right panel is analytic, not a confidence interval estimated from the audit.

Table 10: Three distinct generic robust estimators. The post-decision ReCoRD ablations do not supply another prospective decision or validate the IID sampled-sibling certificate.

Procedure	Observation regime	Concentration	Evidence
Sampled-sibling generic DRO	K IID siblings	Anchor empirical Bernstein	Controlled laws; Proposition 2
Enumerated finite-view DRO	Known finite nuisance views	Anchor empirical Bernstein	Retrospective HANS/QA
Finite-catalog prospective DRO	Enumerated views; SRS without replacement	Hoeffding/MGF comparison	Original ReCoRD decision

Table 11: Mapping from the comparative theorem to the LLM diagnostic. The distinction prevents an unpaired risk table from being presented as a paired utility certificate.

Object	Operational treatment
Anchor/sibling variance	Resample fresh anchor clusters and report the fitted $a/n + b/(nK)$ components.
Single-model sensitivity	Report predictive KL and held-out style risk under pre-specified covered mixtures.
Paired g and V_ν^Δ	Computed only when factual and CFPT losses are available on the same fresh examples; otherwise marked unavailable rather than reconstructed from two marginal risks.
Generator fidelity	Automatic screening is reported as a filter statistic; no exact semantic-preservation claim is made without an independent audit.
Semantic contamination	The correction $2B\eta_{\text{sem}}$ is not set to zero for the free-text arm; a valid upper bound would require an independent audit.

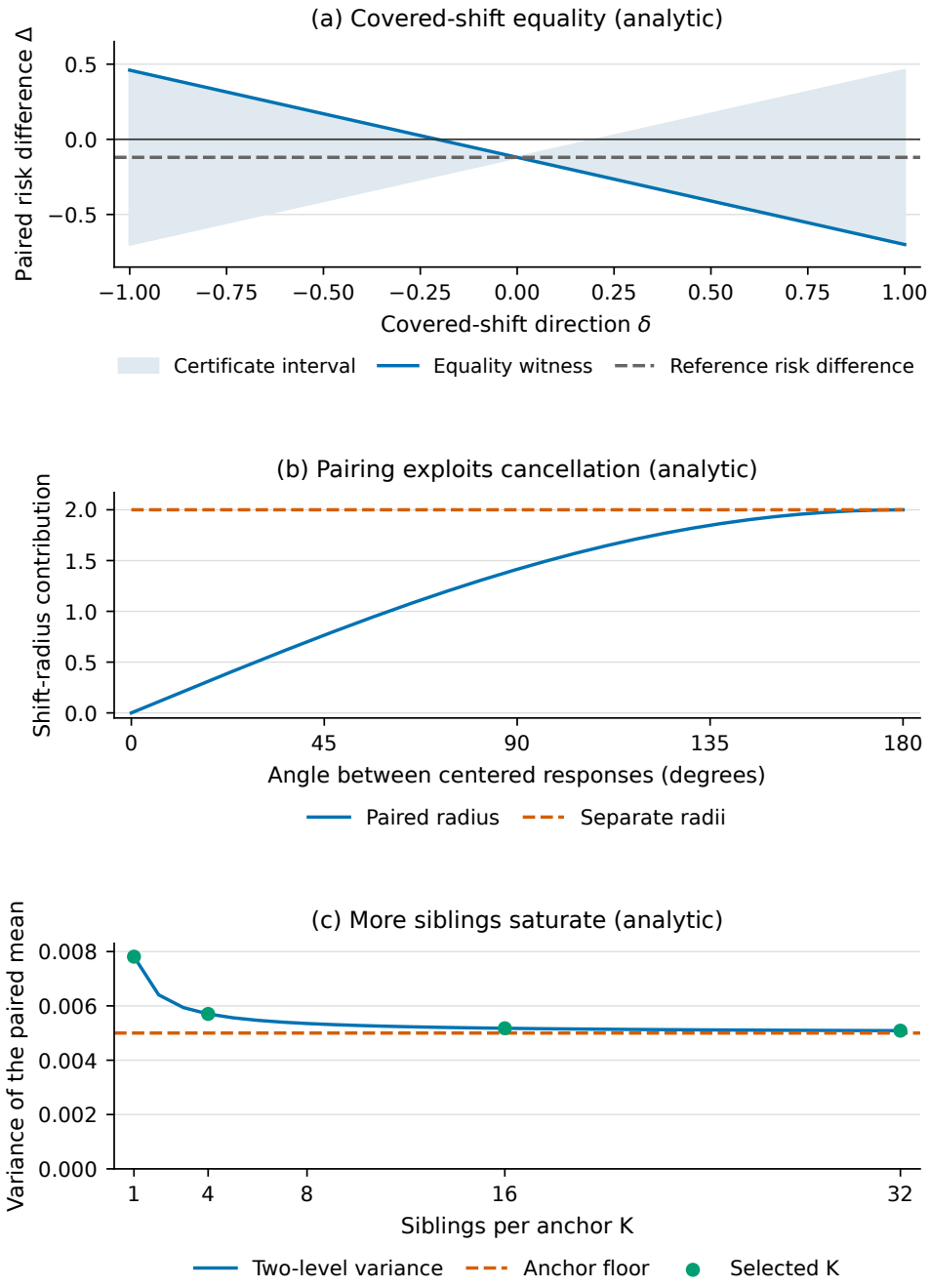


Figure 13: Controlled checks of the comparative-utility theorem. (a) In the two-style witness, the paired risk difference follows the endpoint of the covered-shift interval for every shift direction, demonstrating equality in the Cauchy–Schwarz bound. (b) Two models with the same separate nuisance radii can have a much smaller paired radius when their centered responses cancel; the gap is controlled by their alignment, not by marginal sensitivity alone. (c) With $n = 128$, increasing the number of siblings removes only the conditional sibling term and approaches the anchor floor. All curves are computed directly from the displayed formulas; no fitted parameters are used.

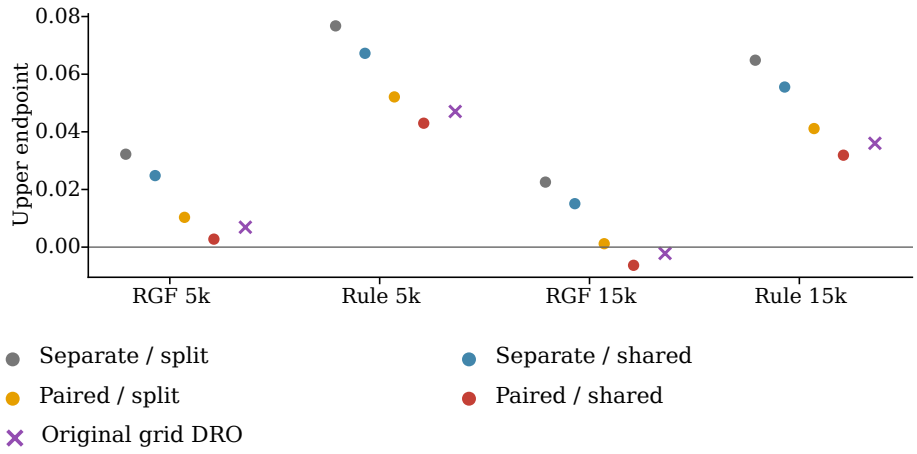


Figure 14: ReCoRD post-decision endpoints on identical pre-test records. The four moment methods have matched ledgers; original DRO retains its predeclared grid ledger. No new prospective selection or sealed-test reuse is claimed. All policies and failed endpoints remain visible.

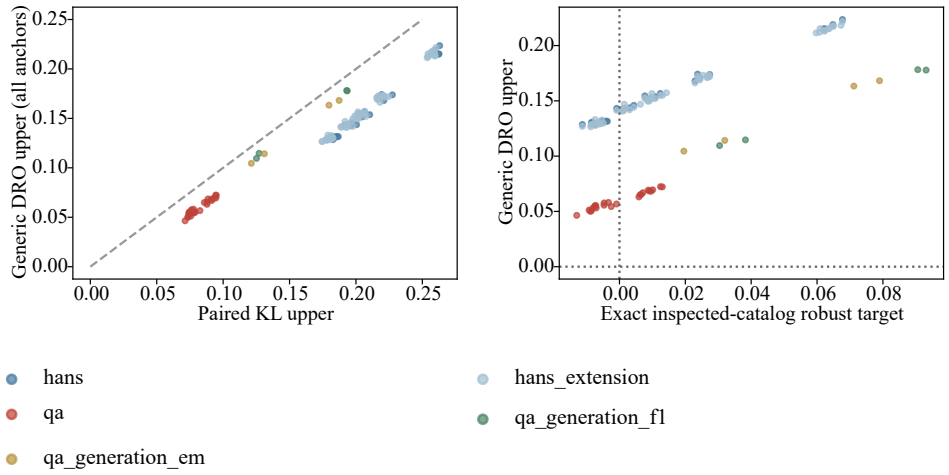


Figure 15: Retrospective comparison on identical inspected HANS/QA records. Generic endpoints are tighter but all remain positive. Right: inspected-catalog robust targets separate negative-target, width-limited records from records already unfavorable in the catalog. Neither axis supplies prospective evidence.

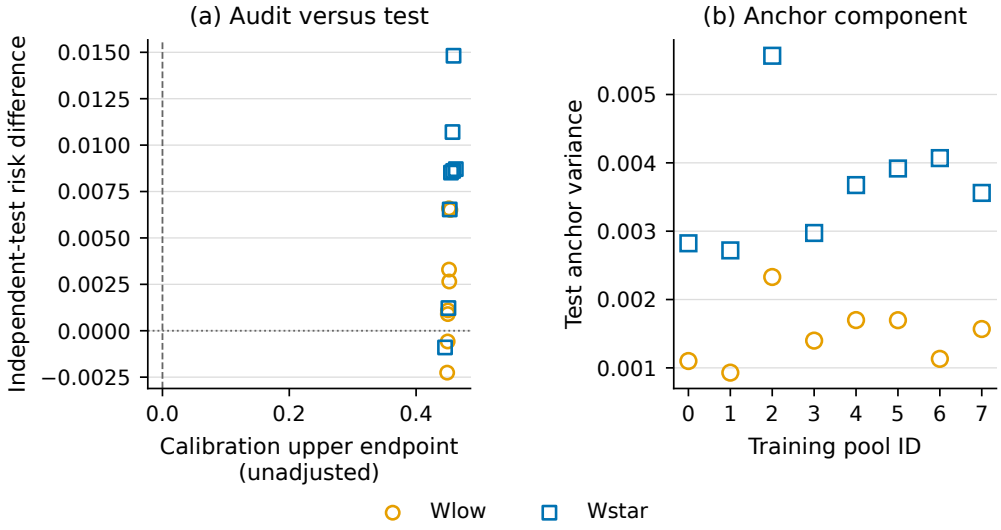


Figure 16: Paired utility audit computed from locked Qwen/MultiNLI predictions. Each point in (a) pairs the unadjusted upper endpoint from the 400-anchor calibration block with the paired risk gap on the independent 800-anchor test block, for one of eight training pools. Both blocks use four sibling views per anchor. Panel (b) reports the test anchor component computed from the same paired loss arrays. The bounded radius is conservative and the semantic correction is unknown, so this figure establishes computability of the diagnostic, not a theorem-valid selection result for the rewrite arm.

Table 12: Completed HANS replication. Numerical columns are descriptive means over all twelve training-order seeds. The last two columns count seeds with negative empirical risk differences at every tested $q \in [.40, .90]$, for the eight-seed extension and all twelve seeds respectively. Counts are not confidence guarantees; all seeds share the structure catalog and initialization.

Policy	Ref. ΔR	$\widehat{V}_\nu^\Delta$	$\Delta R_{q=.40}$	New eight	All twelve
Paired P50 (120)	-0.0191	0.00260	-0.0177	8/8	12/12
P995	-0.0158	0.00888	-0.0041	7/8	11/12
Cue70	-0.0115	0.01289	+0.0052	0/8	0/12
Cue90	-0.0020	0.01940	+0.0239	0/8	0/12
Cue100	0.0216	0.03748	+0.0683	0/8	0/12

Table 13: Retrospective diagnosis at $\rho = .05$. Signs of the empirical plug-in envelope are not population classifications. Cohorts overlap; records must not be interpreted as independent task samples.

Cohort	Records	Envelope ≥ 0	Width-limited	$U < 0$
HANS 4-seed subset	24	15	9	0
HANS 12 seeds	60	47	13	0
QA 6 families	24	10	14	0

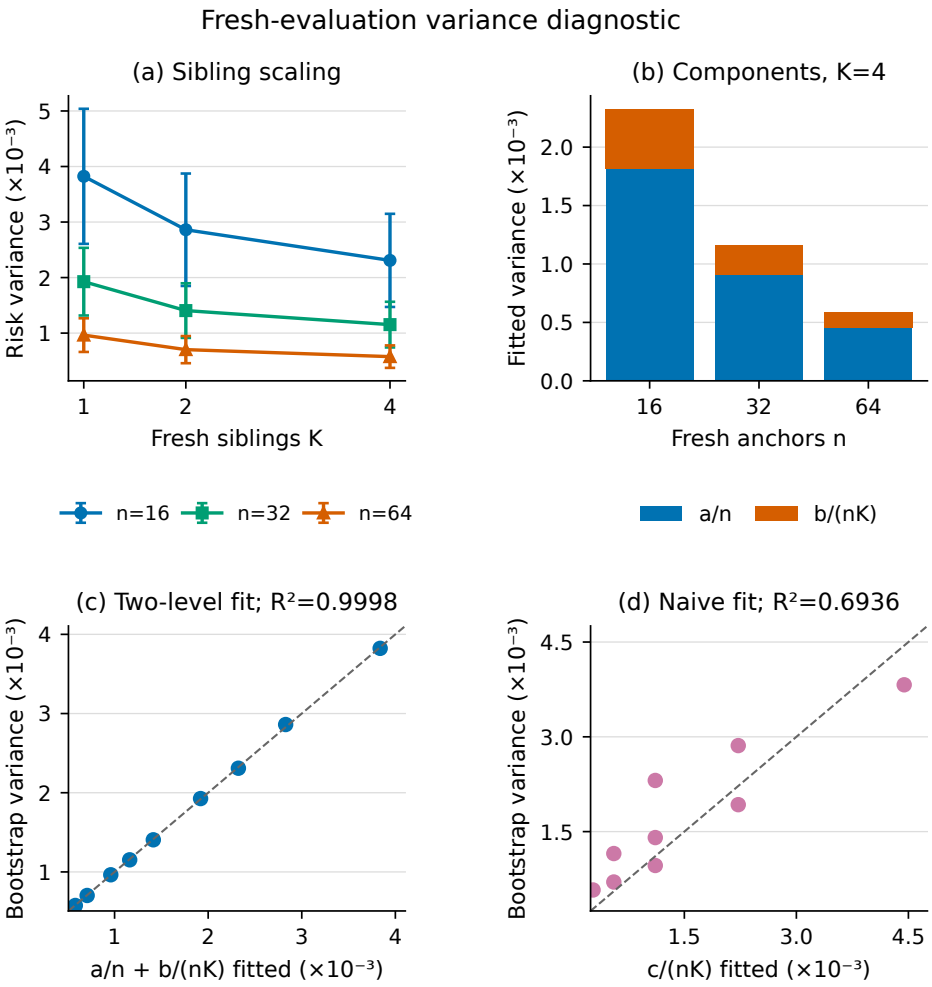


Figure 17: Qwen fresh-evaluation variance diagnostic. The fitted anchor term remains nonzero as sibling multiplicity grows, consistent with the two-level decomposition. In (a), points and error bars are means and standard deviations across 24 fitted models of the bootstrap variance estimate, using 1,200 resamples per model and (n, K) cell. Panel (b) uses $K = 4$. Panels (c,d) compare least-squares fits on the same nine cell means; the axes express variance in units of 10^{-3}. This is a finite diagnostic of the fresh branch, not an estimate of the adaptive mutual-information term.

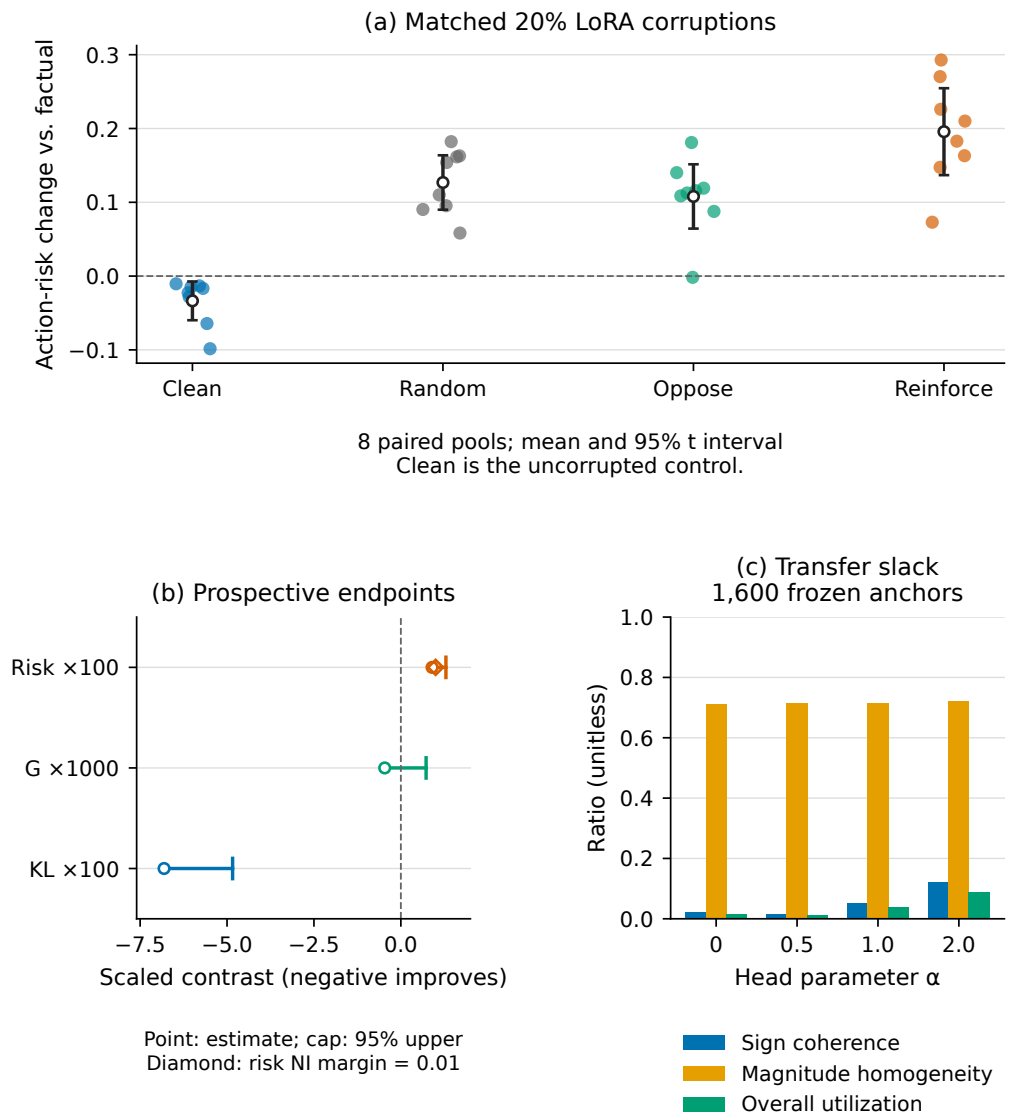


Figure 18: Three complementary Qwen/MultiNLI diagnostics. **(a)** Matched-error LoRA interventions: individual paired pools and mean action-risk changes, with 95% t-intervals over eight pools. **(b)** The independent-pool sensitivity intervention: open points are contrast estimates and caps are one-sided 95% upper bounds, not two-sided intervals. The diamond marks the .01 reference-risk non-inferiority margin, applying only to that row. Only the KL criterion passes. **(c)** Retrospective cue-reversal slack: sign coherence and magnitude homogeneity multiply to the fraction of the head-aware bound used. The three panels concern distinct protocols; none treats a reduced consistency statistic as sufficient evidence of a deployment gain.

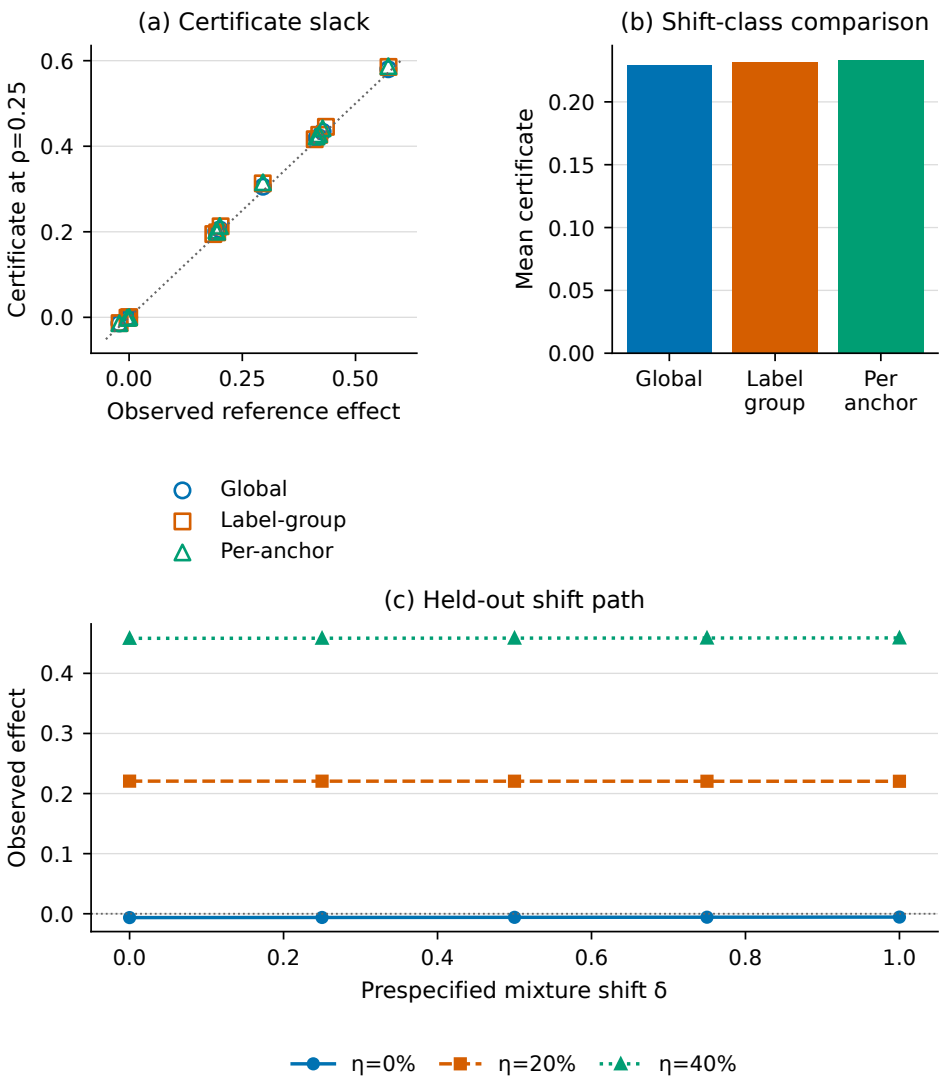


Figure 19: P0 shift-granularity audit on locked Qwen predictions. (a) At $\rho = 0.25$, each point is a three-seed mean for one model-size, task, and corruption condition. The three shift classes lie close to the identity in this finite pool. (b) Bars average certificates over the same model-size, task, and corruption conditions; more local classes give slightly wider certificates. (c) The prespecified mixture shift path is reported separately for corruption rates $0\%, 20\%, 40\%$. This is a real-LLM operational diagnostic, not evidence that one class is universally correct.

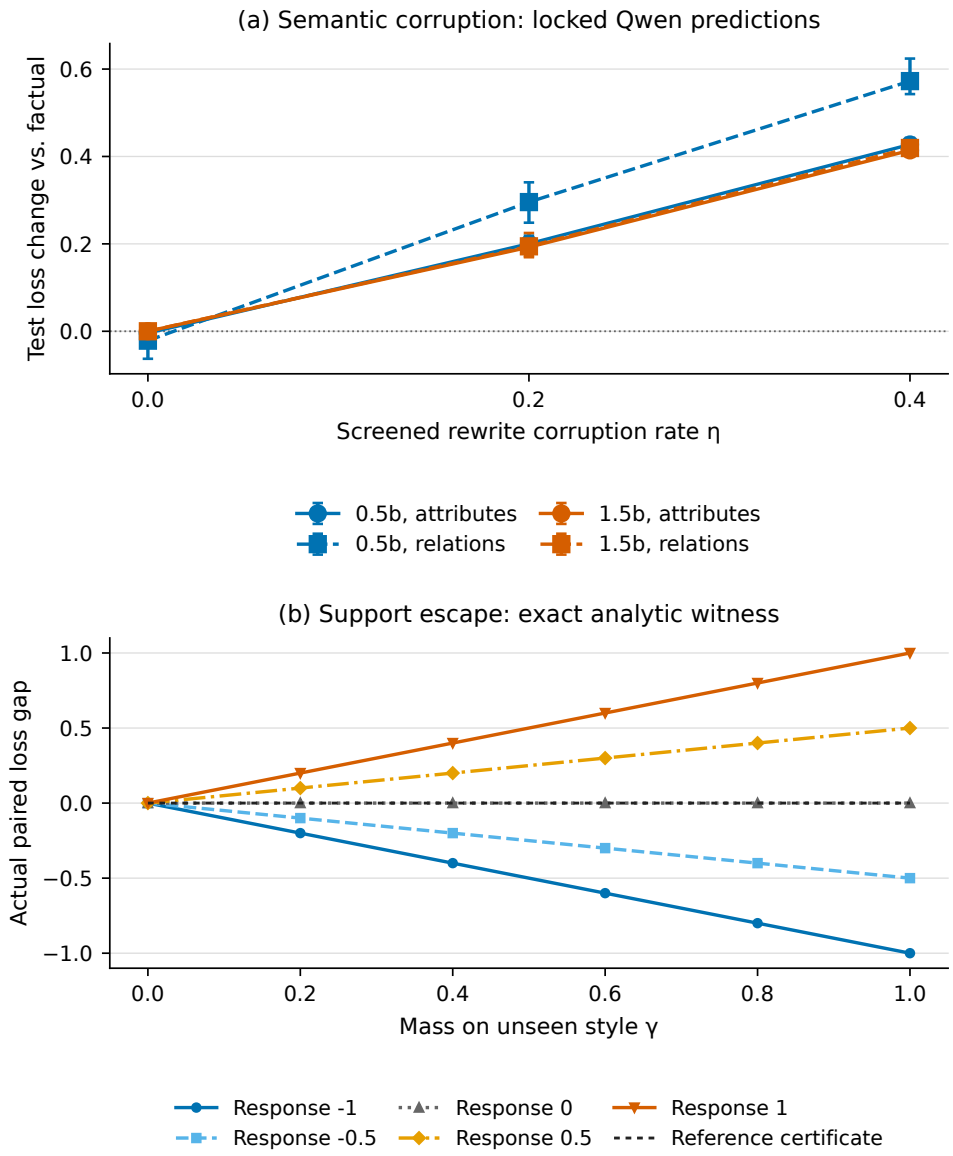


Figure 20: Two failure boundaries. (a) Locked Qwen/MultiNLI predictions show the held-out loss change relative to factual training as the screened rewrite corruption rate increases. Points are means over three training seeds; error bars are percentile 95% intervals from 2,000 bootstrap resamples of those seeds. The plot is not a human semantic audit. (b) In an exact support-escape witness, the reference-only certificate remains zero because the reference law assigns no mass to the unseen style, while the realized risk gap varies linearly with off-support mass and can have either sign.

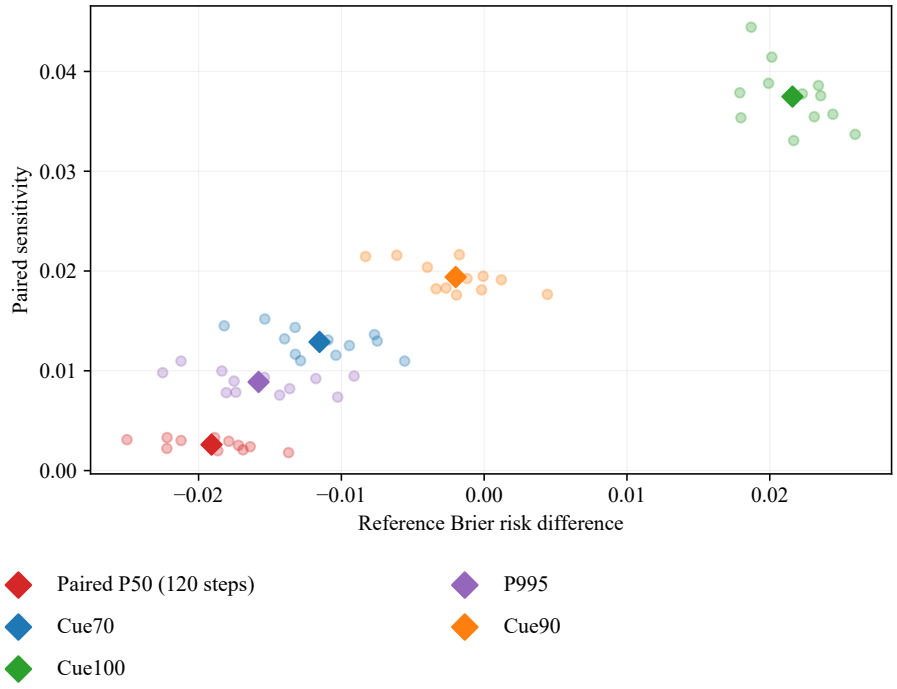


Figure 21: Reference gain and paired sensitivity on the HANS structure catalog, all twelve training-order seeds (original four plus additional eight). Faint points are individual seeds and diamonds are means. Similar mean benefits can coexist with different paired sensitivities. The figure does not assert reference equivalence or isolate a causal effect of sensitivity. Only the 120-update paired policy appears in the primary comparison.

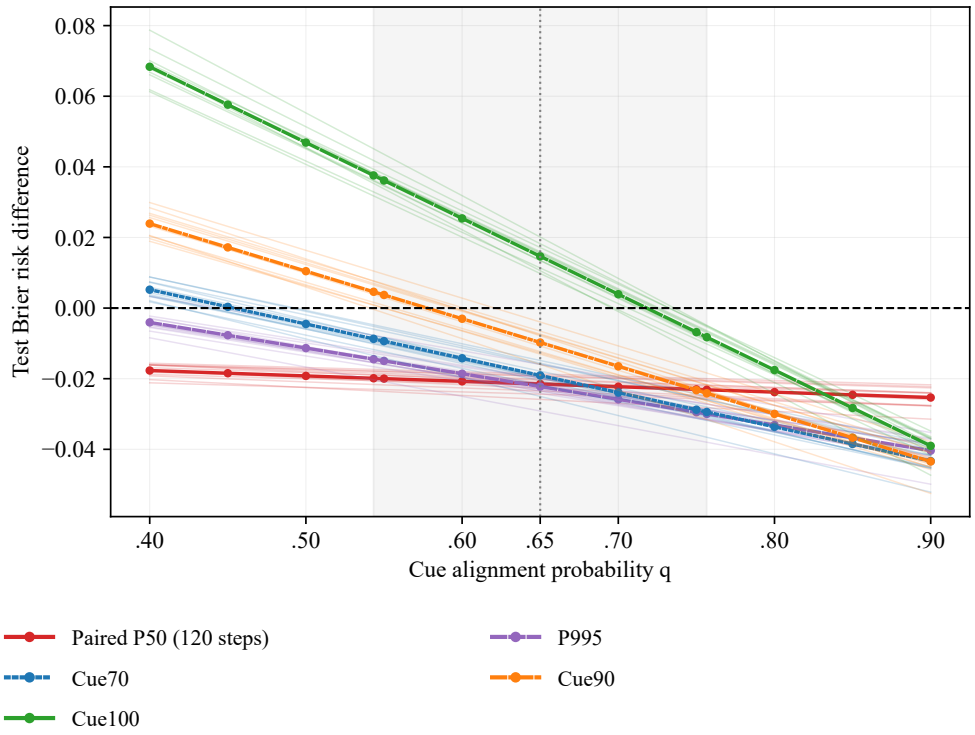


Figure 22: Expanded empirical response to cue-alignment shifts for the same twelve seeds. Negative values indicate improvement over factual training. Bold lines are means; faint lines are individual seeds. The gray band is the original $\rho = 0.05$ interval, and the vertical dotted line marks $q_0 = 0.65$. The expanded range has maximum chi-square divergence approximately 0.2747. Low sensitivity accompanies flatter response in these candidates, while Cue70 and Cue90 lose their test reference advantage at sufficiently low alignment. P995 has one individual-seed crossing despite its negative mean curve. All twelve seeds and five primary policies are shown; the redundant 60-update checkpoint remains archived.

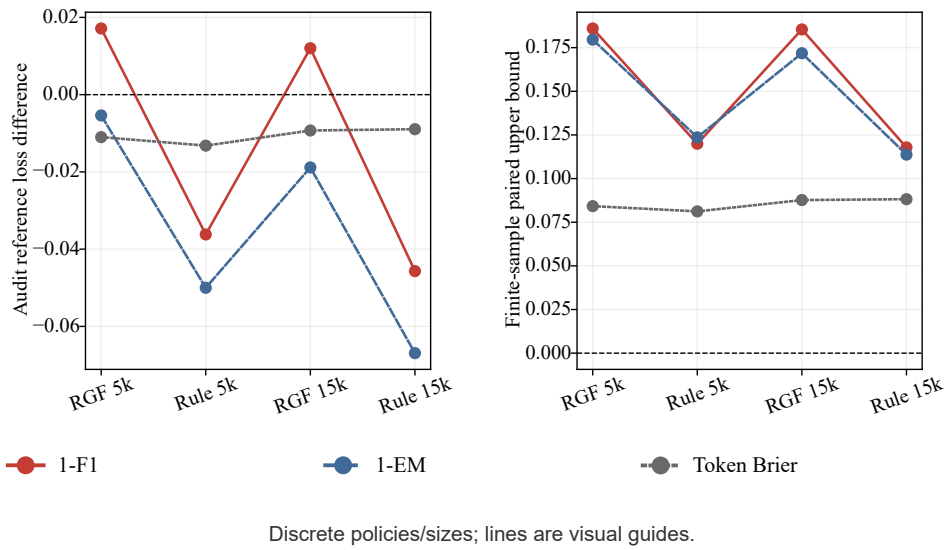


Figure 23: Same anchors and candidates, different bounded QA audit losses. Token-Brier improvement need not imply F1 improvement. All upper certificates remain nonnegative. This is a retrospective alignment study with newly computed calibration/reference generations. The horizontal categories are distinct policy and training-size combinations; connecting lines are visual guides, not interpolation over a common continuous parameter.

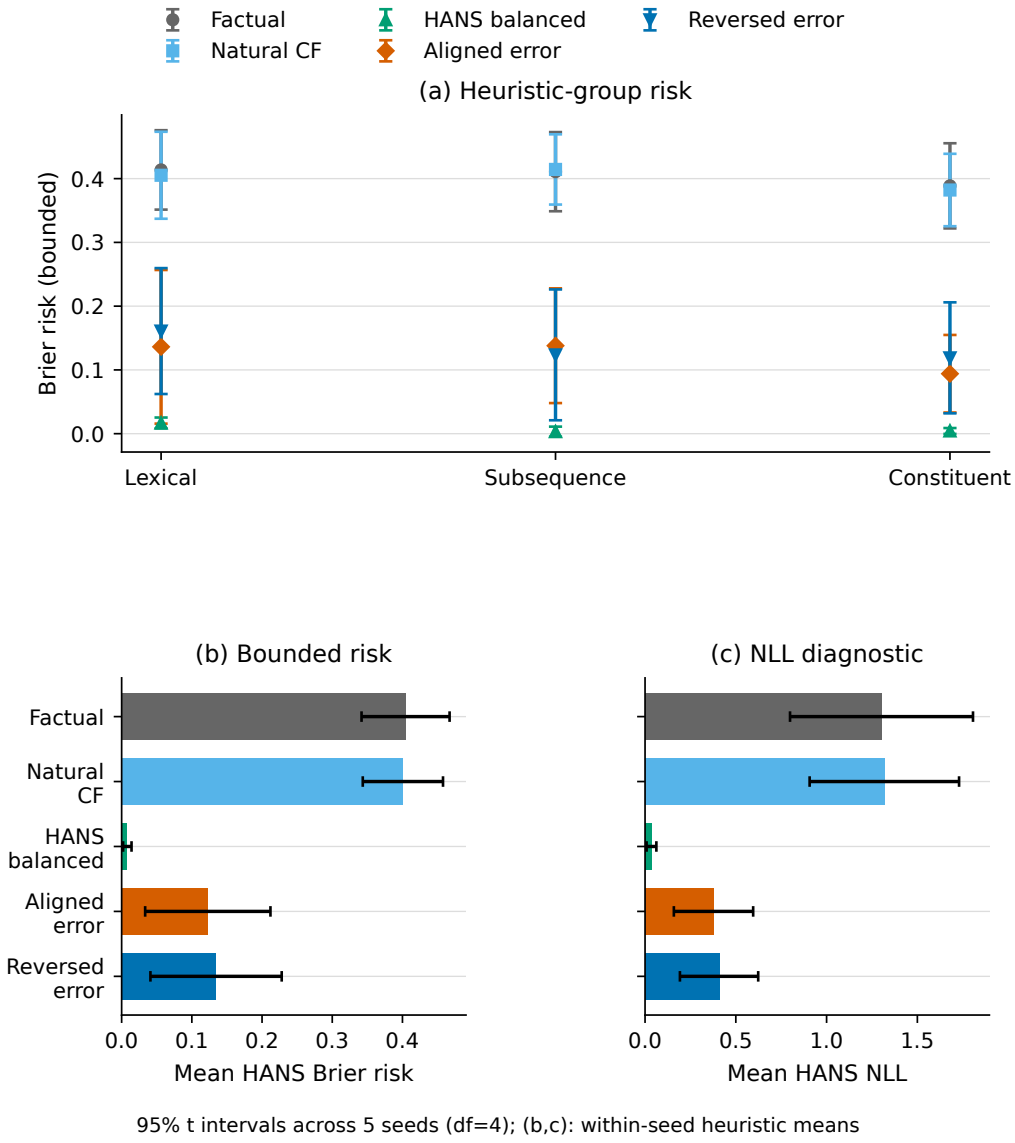


Figure 24: Controlled language diagnostic on HANS with five independent seeds. (a) Brier risk for the three heuristic groups; the shared-scale profile shows whether an intervention changes all groups similarly or leaves a group-specific residual. (b) Mean bounded HANS risk by training condition. Factual and automatically screened natural-CF training remain near chance, while the HANS-balanced intervention has high headroom; the two 20% corruption directions have different risk profiles despite matched rate and pool size. (c) NLL exposes calibration-sensitive differences that accuracy alone can hide. Points/bars are means across five training seeds; error bars are 95% t-intervals with four degrees of freedom. In (b,c), the three heuristic groups are averaged within each seed before computing the interval. HANS-balanced is a controlled template-family intervention, not evidence of natural-language semantic preservation or arbitrary deployment-shift robustness.

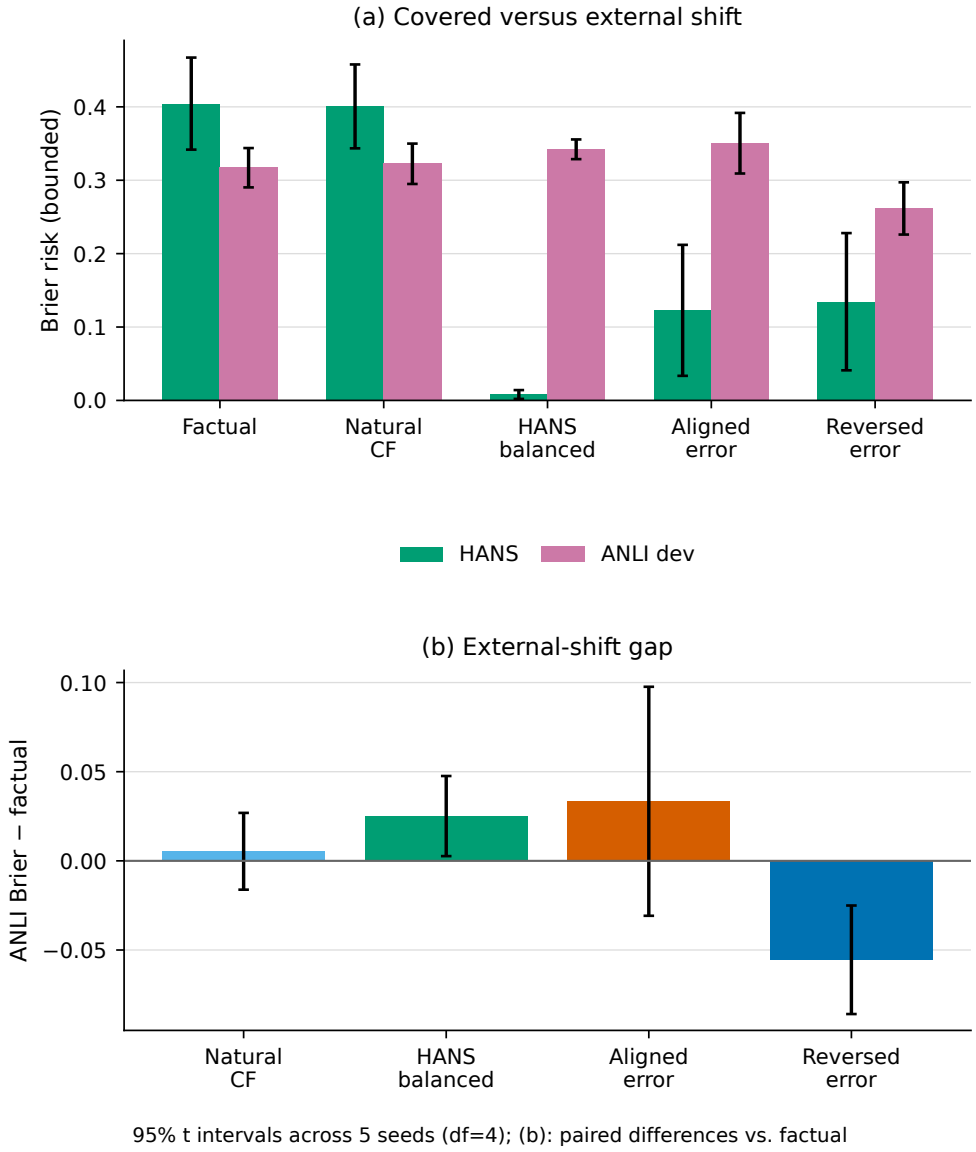


Figure 25: External-shift check using the same five-seed models. (a) HANS is the controlled heuristic family used for the intervention, whereas ANLI dev (Nie et al., 2020) is a frozen external evaluation after the fixed binary entailment reduction. HANS-balanced sharply reduces the covered HANS risk but does not reduce ANLI risk. (b) ANLI Brier change relative to factual training; the HANS-balanced increase is positive, while the natural-CF interval overlaps zero and the two matched-rate corruption directions separate. Error bars are 95% t-intervals with four degrees of freedom over five independent seeds. The panel illustrates coverage and direction effects; it is not a claim that any one corruption direction is universally better.