DETECTING STRUCTURED EFFECTS IN PREFERENCE DATA: A FISHER INFORMATION PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference comparisons selected to improve reward-model fit need not be informative for detecting interactions among evaluation dimensions. We study this selection problem through nuisance-adjusted Fisher information, which combines preference variance with the interaction signal remaining after accounting for main effects. This criterion links labeling budget to a local detection boundary and motivates Cond-Opt, a frozen-pilot selector accompanied by checks of selected-set identifiability and information retention. In fixed HelpSteer2+3 and UltraFeedback pools, the top 1% of comparisons account for 92.6% and 89.8% of estimated interaction information. In matched-budget simulations with split-null calibration, Cond-Opt achieves higher power than unresidualized interaction selection and one-shot full-parameter Fisher leverage across all six evaluated budget–noise settings at $\gamma^\star = 0.5$. Against uniform sampling, its median gain is 58.1 percentage points within the prespecified uniform-power window of 20–80% in a separate grid. On HelpSteer2+3 features with simulated labels, it reduces the 50%-power detection boundary by 36%. Tests on observed human labels do not confirm the specified interaction at the evaluated scale. Together, these results identify the value of targeting residual interaction information and provide a framework for assessing which comparisons can inform a specified preference question.

1 INTRODUCTION

Pairwise preference data are commonly curated to improve reward-model fit, optimize direct preference optimization (DPO) objectives, or maximize policy win rates (Melikidze et al., 2026). The same data also support questions about how evaluation dimensions interact: does the penalty for verbosity increase when response correctness is low? Answering this question requires separating the interaction from the individual effects of verbosity and correctness. A comparison that helps predict the preferred response need not help make this distinction. Our thesis is that preference selection for interaction detection should target the information remaining after accounting for these main effects. We evaluate this principle with a specified feature that aggregates conflicts across dimensions (Section 3.1); the verbosity–correctness question illustrates the broader inferential motivation.

Information for Interaction Detection. Detection value depends on two factors: uncertainty about the preferred response and interaction signal that cannot be explained by main effects. Near-deterministic preferences carry little information, but uncertainty alone is also insufficient when the target interaction adds no distinct signal. Selection criteria aimed at predictive fit or preference uncertainty (Melikidze et al., 2026; Deng et al., 2025; Feng et al., 2025) therefore need not rank comparisons by their value for detecting a specified interaction. In a local logistic model, nuisance-adjusted Fisher information captures both factors (Equation 2). Its link to detection power yields an analytic boundary connecting sample allocation to the smallest detectable effect.

We turn this criterion into **Cond-Opt**, a budget-constrained selection protocol (Figure 1). A pilot fitted under the no-interaction null estimates preference uncertainty and removes the part of the interaction explained by main effects. Cond-Opt ranks comparisons by frozen information scores without accessing evaluation labels. A separate finite-sample audit reports identifiability and how much of that score is retained after refitting the projection on the selected set. This connects a target-specific detection objective to a concrete choice of which comparisons to label.

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.

The empirical comparisons test what this target contributes beyond generic information-based selection. In all six evaluated budget–noise settings at $\gamma^\star = 0.5$, Cond-Opt yields higher power than unresidualized interaction selection and one-shot full-parameter Fisher leverage. In fixed HelpSteer2+3 and UltraFeedback pools, the top 1% of comparisons account for 92.6% and 89.8% of estimated interaction information. On HelpSteer2+3 features with shared simulated labels, Cond-Opt reduces the 50%-power detection boundary by 36%. Tests on observed human labels do not confirm the specified interaction, motivating a distinction between controlled allocation gains and the detectability of natural effects (Section 4).

Contributions.

1. **A detection-oriented formulation of preference selection.** We specialize nuisance-adjusted Fisher information and local power analysis to a specified preference interaction, linking residual signal, preference variance, and labeling budget to an analytic detection boundary (Section 3).
2. **Evidence for targeting residual interaction information.** Controlled comparisons distinguish Cond-Opt from uncertainty, unresidualized interaction, and full-parameter Fisher selection. Analyses of two preference pools locate this information, while observed-label tests assess the specified effect at available sample sizes (Section 4).
3. **A concrete selection rule with finite-sample diagnostics.** Cond-Opt uses a frozen null pilot to rank comparisons without evaluation labels. Selected-set audits expose projection losses, and split-null calibration controls the benchmark tests' false-positive rates (Sections 3.2 and 4.1).

2 Related Work

Preference selection for fitting and alignment. Active preference methods traditionally select comparisons to improve reward modeling, reduce downstream alignment regret, or accelerate direct preference optimization (Melikidze et al., 2026; Das et al., 2025; Lin et al., 2026). Strategies like ActiveUltraFeedback prioritize pairs with large predicted quality margins (Melikidze et al., 2026), while Active Reward Modeling leverages Fisher-information criteria on linear representation layers to optimize parameter estimation (Shen et al., 2025). Their selection objectives target fitting, policy gains, or global parameter precision rather than the conditional information for one interaction. In contrast, our criterion isolates the information specific to a designated structured interaction after projecting out nuisance main effects, $I_{\gamma|w} = \mathbb{E}[vr^2]$ (Table 10, Appendix D). To compare the selection objectives on shared features, we implement a DRTS-based max-margin proxy in a controlled feature environment; this comparison isolates the selection-rule difference within our benchmark.

Identification, estimation, and detection from preferences. Theoretical properties of pairwise data—including identifiable reward subspaces and Bradley–Terry consistency—have been characterized by recent work (Pukdee et al., 2026; Drago et al., 2026). In parallel, DMLRank develops debiased ranking estimation and uncertainty quantification under budget constraints (Frauen et al., 2026). In evaluation settings, Lee (2026) analyze how margin distributions govern the power to detect overall model-level improvements across prompts. We address a complementary inferential problem: testing specific structured interaction coefficients within preference models. By linking nuisance-adjusted Fisher information to local detection boundaries, our framework enables principled power planning and empirical detectability audits, bridging theoretical limits with observed preference distributions.

Classical information and experimental design. Classical local and logistic interaction designs motivate our criterion; see Section 3.2 and Appendix C (Chernoff, 1953; Ford et al., 1985; Kiefer, 1961; Kiefer & Wolfowitz, 1960; Demidenko, 2008; van Berkum, 1987; Graßhoff et al., 2003).

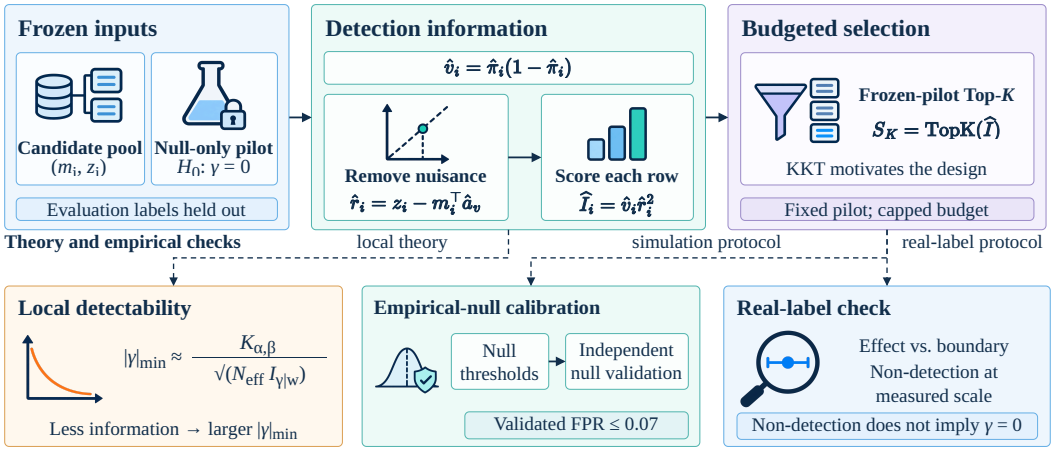

Figure 1: Selecting comparisons for a specified interaction. A frozen null pilot estimates preference variance and squared residual interaction signal; their product defines the pair-level score $\widehat{I}_i = \hat{v}_i \hat{r}_i^2$. Cond-Opt selects the top-K pairs; a separate audit reports identifiability and information retention. Simulated labels measure detection power, while observed-label tests separate pilot estimation and hypothesis testing across prompt clusters.

3 METHOD

Cond-Opt selects comparisons informative about a specified interaction after accounting for main effects. It fits a null model on pilot labels, removes the interaction component explained by main effects, scores the residual information, and selects the top K comparisons. We then relate this rule to a local detection boundary.

3.1 DETECTION TARGET AND NUISANCE-ADJUSTED INFORMATION

Each comparison has normalized differences $m_i \in [-1, 1]^d$ across evaluation dimensions, a binary preference $y_i \in \{-1, +1\}$, and a specified interaction feature $z_i = z(m_i)$. We use the logistic model

$$\Pr(Y_i = +1 \mid m_i) = \sigma(w^\top m_i + \gamma z_i), \qquad \sigma(t) = (1 + e^{-t})^{-1}, \tag{1}$$

where $w \in \mathbb{R}^d$ captures main effects and γ is the target coefficient. The nested likelihood-ratio (LR) test compares $H_0 : \gamma = 0$ with an unrestricted alternative, estimating w in both models.

At a null reference w_0, let $\pi_i = \sigma(w_0^\top m_i)$ and $v_i = \pi_i(1 - \pi_i)$ denote the preference probability and Bernoulli variance. Estimating the main effects removes the part of z that they can explain. The weighted projection, its residual, and the resulting Fisher information are

$$\begin{aligned} a_v &= \arg\min_{a \in \mathbb{R}^d} \mathbb{E}[v(z - a^\top m)^2], \qquad r = z - a_v^\top m, \\ I_{\gamma|w} &= \mathbb{E}[vr^2], \qquad\qquad I_i = v_i r_i^2. \end{aligned} \tag{2}$$

Expectations are under the candidate sampling distribution. The identity follows from the Schur complement of the nuisance-information block, with $\mathbb{E}[vmr] = 0$ (Appendix A.1). Thus a comparison informs the target coefficient through both preference variance and residual interaction signal. High uncertainty alone, or a large unadjusted interaction feature, need not yield high inferential value.

Evaluated interaction. To study conflicts across dimensions, we specify a feature built from the largest positive and negative rubric differences. Its coefficient measures an aggregate conflict effect, distinct from the illustrative named-dimension interaction in the introduction. For $u_i = \max_k m_{ik}$, $\ell_i = \min_k m_{ik}$, and fixed $\tau = 1/12$, define

$$I_\tau(m_i) = \mathbf{1}\{u_i > \tau,\ \ell_i < -\tau\}, \qquad z_i = I_\tau(m_i) u_i \ell_i (u_i + \ell_i). \tag{3}$$

The strict conflict gate can be active while $z_i = 0$ when $u_i + \ell_i = 0$. Specifying the feature and its scale fixes which structured effect is being tested.

3.2 Frozen-Pilot Selection and Finite-Sample Audits

For a candidate pool $\mathcal{P} = \{(m_i, z_i)\}_{i=1}^n$ and budget $K \le n$, Cond-Opt freezes nuisance estimates from a pilot. We fit $\hat{w}_{H_0}$ with $\gamma = 0$ on pilot labels and derive $\hat{v}$ and the weighted projection $\hat{a}_v$ without accessing evaluation labels. Holding $\widehat{I}_i = \hat{v}_i(z_i - m_i^\top \hat{a}_v)^2$ fixed gives a linear objective whose exact optimum selects the top K scoring pairs (Algorithm 1).

Algorithm 1 Cond-Opt: Selection with a frozen null pilot

Input: Candidate features $\mathcal{P}$, pilot data $\mathcal{D}_0$, budget K
1: Fit null model $\hat{w}_{H_0}$ on $\mathcal{D}_0$ with $\gamma = 0$
2: Compute preference probabilities $\hat{\pi}_i = \sigma(m_i^\top \hat{w}_{H_0})$ and variances $\hat{v}_i = \hat{\pi}_i(1 - \hat{\pi}_i)$
3: Estimate the $\hat{v}$-weighted projection $\hat{a}_v$ on reference features
4: Evaluate frozen scores $\widehat{I}_i = \hat{v}_i(z_i - m_i^\top \hat{a}_v)^2$ for all candidates
5: Select $S_K = \mathrm{TopK}(\widehat{I})$; set $q_i = K^{-1}\mathbf{1}\{i \in S_K\}$
6: Compute rank, positive-information, and projection-fidelity audit results on S_K (below)
7: **return** S_K, q, and audit results

Algorithm 1 computes the projection on candidate-pool features in simulations and on pilot features in observed-label evaluations. Both settings exclude evaluation labels; the observed-label design also separates pilot and evaluation data by prompt cluster.

Relation to optimal design. The corresponding local design problem seeks

$$\max_q \sum_{i=1}^n q_i v_i r_i(q)^2, \qquad \sum_i q_i = 1, \quad 0 \le q_i \le 1/K. \tag{4}$$

The cap includes uniform weights on any K distinct comparisons. Here the projection changes with q, unlike Cond-Opt's frozen objective. Classical optimal-design theory (Kiefer, 1961; Kiefer & Wolfowitz, 1960) establishes concavity and characterizes the interior optimum via KKT conditions (Appendix A.3).

Identifiability and projection fidelity audit. To check how selection changes the projection, we compare frozen and refitted information on the selected subset S:

$$\widetilde{I}(S) = \frac{1}{K}\sum_{i \in S} \hat{v}_i(z_i - m_i^\top \hat{a}_v)^2, \qquad I(S) = \frac{1}{K}\sum_{i \in S} \hat{v}_i(z_i - m_i^\top \hat{a}_v(S))^2,$$

where $\hat{a}_v(S)$ is the refitted projection on S. The audit checks (1) full rank of the main-effect matrix m_S and augmented matrix $[m_S, z_S]$; (2) strictly positive profiled information; and (3) projection fidelity $I(S)/\widetilde{I}(S) \ge 0.95$. A local-exchange refinement (**Cond-Opt+repair**) accepts candidate swaps only if they strictly increase re-projected information $I(S)$. The 0.95 threshold is a projection-fidelity diagnostic, separate from identifiability and test calibration. The benchmark retains each selected set and reports its audit status; a failed fidelity check flags information loss beyond the stated tolerance. Repair can improve retained information without reaching the threshold or a global optimum.

3.3 Local Detection Boundary

The detection boundary links this information to the signal strength required for a test. For N independent comparisons, fixed sampling distribution q, and $0 < I_{\gamma|w} < \infty$, the local Gaussian approximation gives

$$|\gamma|_{\min}(\alpha, \beta, N, q, \sigma_J) \approx \frac{K_{\alpha,\beta}}{\sqrt{N I_{\gamma|w}}}, \qquad K_{\alpha,\beta} = z_{1-\alpha/2} + z_{1-\beta}, \tag{5}$$

where $1 - \beta$ is target power, α is the significance level, z_p is the standard-normal quantile, and σ_J indexes covariate noise. The dependence on σ_J enters through the information: when the observed

interaction differs from the label-generating signal, we replace $I_{\gamma|w}$ with the effective information defined below. This specializes classical local-design and logistic-interaction power calculations (Chernoff, 1953; Demidenko, 2008). Clustered observations are handled via cluster-robust inference rather than treating clustered rows as independent trials.

The boundary uses the full residual information: $z = 0$ does not imply $r = 0$. A factorization over the interaction support requires zero residual information outside that support (Appendix A.6). Rescaling z by $c \neq 0$ scales $|\gamma|_{\min}$ by $1/|c|$, preserving scale equivariance relative to the specified feature map.

Measurement noise. With noisy covariates, the observed interaction z_{obs} may differ from the latent signal s generating labels. Let r be the weighted residual of z_{obs}. The simulation diagnostic uses

$$I_{\text{eff}} = \frac{\mathbb{E}[vrs]^2}{\mathbb{E}[vr^2]}, \qquad \mathbb{E}[vr^2] > 0, \tag{6}$$

in place of $I_{\gamma|w}$. When $s = z_{\text{obs}}$, orthogonality recovers $I_{\text{eff}} = I_{\gamma|w}$. In controlled simulations, s is directly observable; in empirical data, selection relies on the observable pilot score $\hat{v}_i \hat{r}_i^2$ (Appendix A.6).

Information limits. If $v \leq \delta$ on $C = \{|z| > 0\}$, projection gives $I_{\gamma|w} \leq \delta \mathbb{E}[z^2 \mathbf{1}_C]$. If C lies entirely where $v = 0$ under the candidate distribution P, every sampling distribution $q \ll P$ has zero information: reallocating labels cannot restore the unobserved signal. For finite logistic predictors $v > 0$, the bound is strict and scales the required sample size inversely with the available information, $N \propto 1/(\delta\, \mathbb{E}[z^2 \mathbf{1}_C])$, so small positive information translates directly into large labeling budgets (Appendices A.2 and A.4). Testing and null calibration procedures are detailed in Section 4.1.

4 Experiments

The experiments ask where detection information lies, whether targeting residual information improves power beyond generic selection rules, and what the observed labels establish about the specified effect.

4.1 Setup

Data and evidence roles. HelpSteer2+3 (Wang et al., 2024; 2025) supplies 21,750 comparisons in 19,765 prompt clusters, with five-dimensional differences from three rubric judges; 991 pass the conflict gate and 520 have nonzero interaction features. UltraFeedback Binarized (Cui et al., 2024; Hugging Face H4, 2023) adds 3000 comparisons under the same null reference weights. The simulation carrier has 401,734 randomly projected, normalized RLHFlow embedding rows (RLHFlow, 2024). Both empirical pools characterize information concentration without evaluation labels. Simulated labels test power on the carrier and HelpSteer geometry; observed HelpSteer labels assess natural effects. Appendix B.7 details feature construction and provenance.

Comparisons. At matched labeling budgets, the four-strategy benchmark compares **Uniform** sampling without replacement, **Cond-Opt** (maximizing $\hat{v}_i \hat{r}_i^2$), a **DRTS-based max-margin proxy** (maximizing $|\hat{\pi}_i - 1/2|$) (Melikidze et al., 2026), and **APO** (maximizing $\hat{v}_i$) (Das et al., 2025). A separate six-arm diagnostic compares Uniform and Cond-Opt with max-$|z|$, unresidualized vz^2, one-shot full-parameter **Fisher leverage**, and **Cond-Opt+repair** (Section 3.2). The scored selectors use shared frozen null-pilot estimates; the matched budget counts evaluation labels, conditional on the pilot, rather than total acquisition or computation costs.

Testing protocols. At $\alpha = 0.05$, we calibrate each 95th-percentile likelihood-ratio (LR) cutoff on 2000 null replications and validate it on 2000 independent null replications. All 36 four-strategy groups have FPR ≤ 0.07. The same split-null check covers 36 six-arm groups at $K \in \{2000, 5000, 10000\}$ and $\sigma_J \in \{0, 0.1\}$ (maximum FPR 0.066). Alternatives use 2000 shared-world replications and $B = 2000$ paired bootstrap intervals. Judge disagreement estimates $\hat{\sigma}_J \approx 0.03$; the four-strategy grid also tests $\sigma_J = 0.25$ (Lee et al., 2026).

For boundary and empirical-feature simulations, we check asymptotic LR thresholds against simulated nulls. Observed-label tests split prompt clusters and report rejection frequencies. Appendices B and E document calibration, seeds, frozen partitions, and artifact mappings.

4.2 Where Is Detection Information Concentrated?

Detection information is concentrated in both fixed preference pools: the top 1% of comparisons account for 92.6% of total estimated information in HelpSteer2+3 and 89.8% in UltraFeedback (Figure 2). Distributing information uniformly over the nonzero-interaction support would yield only 41.7% and 33.0%, respectively, showing concentration beyond sparse support alone. The synthetic simulation pool shows more diffuse information, with a top-1% share of 6.9% (Table 6).

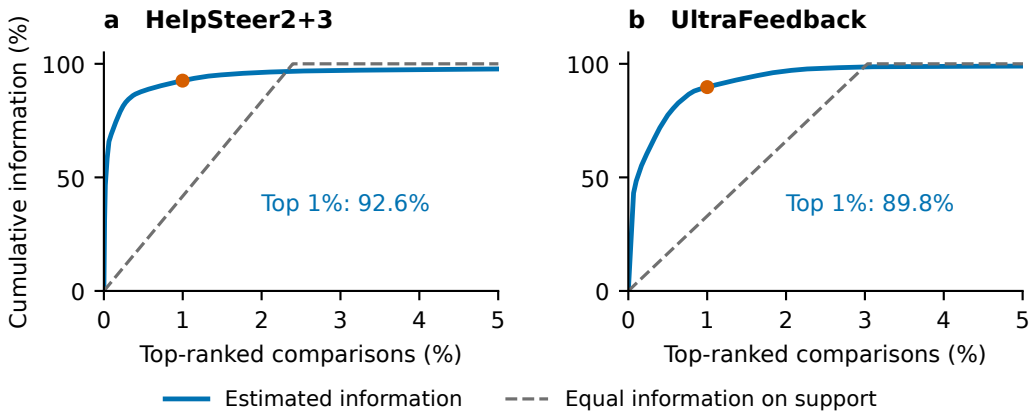

Figure 2: Detection information concentrates sharply in established preference pools. Both panels show the top 5% of comparisons ranked by $\hat{v}_i \hat{r}_i^2$, with cumulative information on the vertical axis. Dashed references allocate information uniformly across the 520 (HelpSteer2+3) or 91 (UltraFeedback) nonzero-interaction pairs. Dots mark the top 1%.

Among conflict-gated comparisons, interaction magnitude and preference variance are negatively correlated (Pearson $\rho = -0.2865$ and -0.4741 under frozen HelpSteer null weights). Across all comparisons, interaction magnitude positively correlates with judge disagreement (Spearman 0.1153 and 0.1255). These associations motivate scoring their joint contribution through residual information rather than either factor alone (Appendix B.7).

4.3 Detection Boundary and Selection Benchmarks

Accuracy of the analytic boundary. The closed-form score-based boundary tracks simulated power with median relative error 1.97% and maximum 6.63% across 35 eligible settings and five target power levels; all 26 evaluable paired directions of boundary shifts agree (Figure 3a). A support-only approximation fails this check, making full residual information necessary for the reported boundary (Appendix B.3).

Matched-budget power gains. Cond-Opt exceeds Uniform power in all 12 four-strategy settings within the prespecified Uniform-power window $[0.2, 0.8]$, with median gain 0.581 (95% paired bootstrap CI $[0.563, 0.595]$). Table 7 reports the complete 27-setting grid, including DRTS proxy and APO. Neither exceeds Cond-Opt anywhere in the grid; ties occur at saturation. DRTS-selected pairs carry mean detection information only $1/16$–$1/18$ of the population mean.

Residualization versus full-parameter information. Among the four selectors, Cond-Opt has the highest power in all six settings at $\gamma^\star = 0.5$ (Table 1). Its paired gains over Uniform exclude zero throughout (Figure 4); these intervals do not directly compare Cond-Opt with Fisher. The full six-arm grid (Table 4) shows similar power after local repair and narrowing differences as signal strength increases.

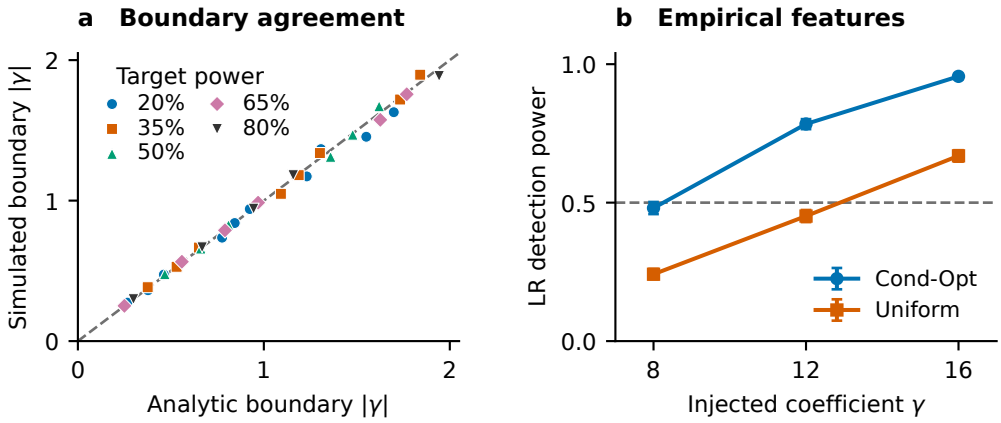

Figure 3: Analytic boundary precision and matched-budget detection power. **(a)** Closed-form versus simulated detection boundaries across 35 eligible settings; dashed line indicates identity. **(b)** Empirical-feature detection power with simulated labels ($N = 10{,}000$, HelpSteer2+3 geometry). Points represent 2000 replications with pointwise 95% Wilson intervals. Cond-Opt reduces the 50%-power detection boundary from 12.89 (Uniform) to 8.26, representing a 36% improvement in detection sensitivity.

Table 1: **Detection power (%) at matched label budgets.** Synthetic carrier with simulated labels at $\gamma^\star = 0.5$, shared frozen pilots, and 2000 alternative replications. Fisher denotes one-shot full-parameter leverage; bold marks the highest point estimate. Independent null validation gives maximum FPR 0.066; Figure 4 reports paired uncertainty.

K	σ_J	Uniform	vz^2	Fisher	**Cond-Opt**
2,000	0	16.40	18.45	43.75	**55.15**
5,000	0	33.65	45.35	80.85	**93.20**
10,000	0	55.60	78.05	95.15	**99.35**
2,000	0.1	14.85	17.95	44.50	**59.95**
5,000	0.1	28.95	36.30	76.35	**93.45**
10,000	0.1	38.95	70.25	89.70	**99.05**

What the selected-set audit establishes. On the synthetic carrier, frozen Cond-Opt retains 92.50%, 95.21%, and 91.05% of its scored information after reprojection at $K = 2000, 5000, 10000$. Repair raises these to 98.16%, 96.97%, and 92.42%, so the largest budget still falls below the 95% fidelity threshold. All sets remain identifiable and are retained in the power comparison, including those flagged by the fidelity audit. The fixed HelpSteer2+3 audit passes at all three budgets (Appendix B.6).

Empirical feature geometry. Using simulated labels on HelpSteer2+3 features tests whether the gain persists on empirical geometry. Cond-Opt lowers the 50%-power boundary from 12.89 to 8.26: a 36% reduction in required signal magnitude (Figure 3b). The median power gain over tested effect sizes is 0.287 (95% CI $[0.2675, 0.3070]$), with null rejection rates 0.057 for Cond-Opt and 0.053 for Uniform. A near-zero-information control ($I = 4.4 \times 10^{-7}$, $N = 2000$) yields rejection rates below 0.015, consistent with the low-information detection limit (Appendix B.8).

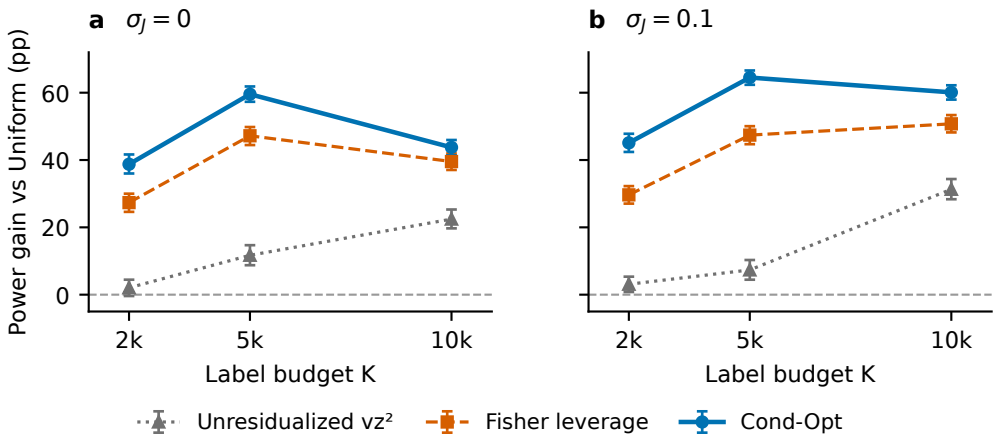

Figure 4: Paired detection-power gains over Uniform at matched label budgets. Panels show all six budget–noise settings on the synthetic carrier with simulated labels at $\gamma^\star = 0.5$. Points are power differences over 2000 shared-world alternative replications; bars are 95% paired bootstrap intervals ($B = 2000$), each relative to Uniform. Fisher denotes one-shot full-parameter Fisher leverage. Independent null calibration and validation precede evaluation; all six-arm validation groups have FPR ≤ 0.066.

4.4 Empirical Detectability and Controlled Validation

Observed-label non-detection. Across 500 out-of-sample prompt-cluster splits on observed HelpSteer human labels, rejection frequencies range from 0.6% to 4.6% among valid fits (Table 2a); these frequencies are not Type-I error estimates. Cond-Opt's row-count plug-in 50%-power thresholds are 13.67–13.77 at $N = 2000$–10000; these are diagnostics, not cluster-robust power guarantees. The retrospective full-pool estimate is $\hat{\gamma} \approx 1.25$ ($p \approx 0.80$). Cond-Opt has no invalid fits in the 500 replications, versus 39 and 2 complete-separation cases for Uniform at the two smaller budgets (Appendix B.9).

Table 2: Observed-label detectability and controlled injections. **(a)** Evaluations across 500 prompt-cluster splits; "Boundary" is a row-count plug-in 50%-power diagnostic, not a cluster-robust guarantee. **(b)** Held-out tests on empirical geometry, one per selected candidate. The primary scan candidate is selected from pre-declared effects; the auxiliary check is diagnostic. Semi-synthetic controls inject $\gamma = 5.915$ and 23.66 with one generated label realization per level; the higher control selects a signal-bearing sibling feature.

(a) Observed labels: repeated pilot/evaluation splits

Rule	N	Valid / 500	Rejections	Rate (%)	Boundary
Uniform	2,000	461	16	3.47	37.27
Cond-Opt	2,000	500	3	0.60	13.77
Uniform	5,000	498	23	4.62	21.53
Cond-Opt	5,000	500	5	1.00	13.68
Uniform	10,000	500	8	1.60	14.41
Cond-Opt	10,000	500	8	1.60	13.67

(b) Held-out checks: one test per selected candidate

Labels / check	Selected feature	$\hat{\gamma}_B$	$\lvert z_B \rvert$	Reject?
Observed / primary	Product, $\tau = 1/4$	12.71	1.11	No
Observed / auxiliary	Gate, $\tau = 1/4$	-0.46	0.81	No
Injected / lower	Gate, $\tau = 1/12$	0.14	1.36	No
Injected / higher	Product, $\tau = 0$	22.21	3.08	Yes

Controlled validation via semi-synthetic injections. Controlled injections check whether the procedure can reject when a specified signal is added to the same empirical geometry (Table 2b). An exploratory screen of fourteen pre-declared candidates selects an interaction at $\tau = 1/4$ with active support 0.082%; it does not confirm on held-out data ($|z_B| = 1.11$). At the higher injection ($\gamma = 23.66$), the scan selects a signal-bearing product sibling at $\tau = 0$, which rejects on held-out data ($|z_B| = 3.08$, $p < 0.01$); the injected $\tau = 1/12$ feature itself is not the selected candidate. At the lower injection ($\gamma = 5.915$), the selected candidate does not reject ($|z_B| = 1.36$). Each injection level uses one generated label realization, so these outcomes check the procedure rather than estimate its detection power (Appendix B.10).

5 Discussion

The value of a preference comparison depends on the inferential question it must answer. In the evaluated mid-power simulations, selecting for a specified interaction yields higher power than selecting for uncertainty or full-parameter Fisher leverage. Combined with the concentration of information in the two empirical pools, this finding motivates evaluating preference datasets by the questions they can resolve, alongside predictive fit or policy win rates (Oh et al., 2026).

The selected-set audit adds a practical qualification to this principle. Ranking frozen information scores is straightforward, but selection changes the geometry used to separate the interaction from main effects. Reporting reprojected information makes this change visible, including cases where local repair improves retention without reaching the fidelity threshold. The audit therefore supports interpretation of the selection rule; split-null calibration separately checks false-positive control.

Scope and limitations. Our power gains use labels simulated within the logistic family underlying the criterion, on synthetic or empirical features. Their robustness to preference-law misspecification and to selection from an independent finite pilot remains unestablished; the in-pool pilot bootstrap assesses a narrower form of stability (Appendix A.6). Realized annotation savings require prospective evaluation that includes pilot acquisition costs. On observed labels, the specified effect is not detected. Its estimate lies below the row-count plug-in boundary, but cluster-aware power analysis is needed to turn this diagnostic into a sample-size recommendation. The single-realization injections check the procedure without estimating power. Under the evaluated judge-noise models, separate null calibration is essential because uncorrected asymptotic tests can inflate false-positive rates (Appendix B.1).

6 Conclusion

Preference selection can be directed at a specified interaction through nuisance-adjusted Fisher information $\mathbb{E}[vr^2]$. We connect this criterion to a local detection boundary and a frozen-pilot selection rule with selected-set audits. Controlled comparisons show why residual interaction information matters beyond uncertainty and full-parameter information, while analyses of two preference pools reveal strong concentration of this resource. Observed-label tests leave the natural effect unconfirmed at the evaluated scale. Together, the results provide a way to connect a preference question, the comparisons selected to answer it, and the information needed for detection.

Artificial Intelligence (AI) use statement

We used generative AI tools to implement methods and experiment code, assist with analysis of experimental records, and revise the manuscript. We reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts.

Reproducibility statement

The appendix specifies data construction, synthetic generators, seeds, pilot estimation, and test protocols. Table 14 maps reported results to source records. The anonymous supplementary package contains analysis code, frozen run records behind Table 14, and a figure bundle with extracted source data, hashes, and scripts that regenerate Figures 2–4, Tables 1 and 2(a,b), and Appendix Table 7 from the packaged artifacts. Judge-noise comparisons use disjoint calibration and validation replicates; empirical-feature comparisons use shared simulated labels and paired uncertainty estimates. Historical exploratory records are cataloged in designated appendix sections for provenance and are demarcated from the current benchmark evaluations. Re-executing the full experimental grid requires the frozen judge-label files, which are not included in the package.

References

Herman Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24(4):586–602, 1953. doi: 10.1214/aoms/1177728915.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.01377.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient RLHF. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, 2025. URL https://arxiv.org/abs/2402.10500. arXiv:2402.10500; accepted at ECML-PKDD 2025.

Eugene Demidenko. Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine*, 27(1):36–46, 2008. doi: 10.1002/sim.2980. URL https://doi.org/10.1002/sim.2980.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving LLM alignment via preference data selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2502.14560.

Simone Drago, Marco Mussi, and Alberto Maria Metelli. Online compatible reward identification from preference feedback. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/65224.

Yunzhen Feng, Ariel Kwiatkowski, Kunhao Zheng, Julia Kempe, and Yaqi Duan. PILAF: Optimal human preference sampling for reward modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16744–16776, 2025. URL https://proceedings.mlr.press/v267/feng25g.html.

I. Ford, D. M. Titterington, and C. F. J. Wu. Inference and sequential design. *Biometrika*, 72(3): 545–551, 1985. doi: 10.1093/biomet/72.3.545.

Dennis Frauen, Athiya Deviyani, Mihaela van der Schaar, and Stefan Feuerriegel. Nonparametric LLM evaluation from preference data. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/61327.

Ulrike Graßhoff, Heiko Großmann, Heinz Holling, and Rainer Schwabe. Optimal paired comparison designs for first-order interactions. *Statistics*, 37(5):373–386, 2003. doi: 10.1080/0233188031000154812.

Kexin Huang, Junkang Wu, Ziqian Chen, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, and Xiang Wang. Larger or smaller reward margins to select preferences for LLM alignment? In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://icml.cc/virtual/2025/poster/44101.

Hugging Face H4. UltraFeedback Binarized. Hugging Face dataset, 2023. URL https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized. Dataset card, accessed September 20, 2026.

Timo Kaufmann, Yannick Metz, Daniel Keim, and Eyke Hüllermeier. ResponseRank: Data-efficient reward modeling through preference strength learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://neurips.cc/virtual/2025/poster/120043.

J. Kiefer. Optimum designs in regression problems, II. *The Annals of Mathematical Statistics*, 32 (1):298–325, 1961. doi: 10.1214/aoms/1177705160.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi: 10.4153/CJM-1960-030-4.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/63607.

Wilson Y. Lee. How many human judgments are enough? feasibility limits of human preference evaluation. *arXiv preprint arXiv:2601.09084*, 2026. URL https://arxiv.org/abs/2601.09084.

Xiaoqiang Lin, Arun Verma, Zhongxiang Dai, Daniela Rus, See-Kiong Ng, and Bryan Kian Hsiang Low. ActiveDPO: Active direct preference optimization for sample-efficient alignment. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10009521.

Davit Melikidze, Marian Schneider, Jessica Lam, Martin Wertich, Ido Hakimi, Barna Pásztor, and Andreas Krause. ActiveUltraFeedback: Efficient preference data generation using active learning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2603.09692.

Giyeong Oh, Junghyun Lee, Jaehyun Park, Youngjae Yu, Wonho Bae, and Junhyug Noh. Random is hard to beat: Active selection in online DPO with modern LLMs. In *ICLR 2026 Workshop on I Can't Believe It's Not Better (ICBINB)*, 2026. URL https://arxiv.org/abs/2604.02766.

Rattana Pukdee, Maria-Florina Balcan, and Pradeep Ravikumar. What does preference learning recover from pairwise comparison data? In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/63498.

RLHFlow. pair_data_v2_80K_wsafety. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/RLHFlow/pair_data_v2_80K_wsafety. Dataset card, accessed September 20, 2026.

Yunyi Shen, Hao Sun, and Jean-Francois Ton. Active Reward Modeling: Adaptive preference labeling for large language model alignment. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 54410–54430, 2025. URL https://proceedings.mlr.press/v267/shen25c.html.

E. E. M. van Berkum. *Optimal Paired Comparison Designs for Factorial Experiments*, volume 31 of *CWI Tract*. Centrum voor Wiskunde en Informatica, Amsterdam, 1987. ISBN 90-6196-311-7.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer 2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.08673.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2505.11475.

A Appendix: Proofs and Technical Derivations

This appendix supplies full derivations for the information identities and optimality conditions sketched in Section 3. Throughout we work under the pairwise preference logit

$$\Pr(Y{=}1 \mid m) \;=\; \sigma\big(w^\top m + \gamma\, z(m)\big), \tag{7}$$

with nuisance $w \in \mathbb{R}^d$, scalar target $\gamma \in \mathbb{R}$, dimension-difference feature $m \in \mathbb{R}^d$, and structured feature $z = z(m) \in \mathbb{R}$. Write $\eta = w^\top m + \gamma z$, $\pi = \sigma(\eta)$, and $v = \pi(1-\pi)$ for the Bernoulli variance. Expectations are with respect to a design measure P on (m, z) (or a reweighted measure q when indicated). We assume standard regularity for logistic MLE (open parameter space, positive-definite Fisher information on compact neighborhoods, dominated scores) so that score/Fisher calculations and envelope differentiation under the integral are justified. These assumptions specify the local information and design calculations below.

A.1 C1: Nuisance-adjusted Fisher information and the δ-bound

Full Fisher block. For the preference label $Y \in \{-1, +1\}$, write $T = \mathbf{1}\{Y = +1\}$. For a single observation the log-likelihood score of equation 7 is

$$\nabla_{(w,\gamma)} \ell \;=\; (T - \pi) \begin{pmatrix} m \\ z \end{pmatrix}. \tag{8}$$

Because $\mathrm{Var}(T \mid m) = \pi(1-\pi) = v$, the expected outer product yields the $(d{+}1) \times (d{+}1)$ Fisher information

$$\mathcal{I}_{(w,\gamma)} \;=\; \mathbb{E}\Big[v \begin{pmatrix} mm^\top & mz \\ zm^\top & z^2 \end{pmatrix} \Big] \;=\; \begin{pmatrix} B & c \\ c^\top & \mathbb{E}[vz^2] \end{pmatrix}, \tag{9}$$

where we define the v-weighted Gram quantities

$$B \;=\; \mathbb{E}[v\, mm^\top] \in \mathbb{R}^{d \times d}, \qquad c \;=\; \mathbb{E}[v\, mz] \in \mathbb{R}^d. \tag{10}$$

Schur complement (profiled information). The nuisance-adjusted (profiled) information for the scalar parameter γ is the Schur complement of B in equation 9. Whenever B is invertible,

$$I_{\gamma|w} \;=\; \mathbb{E}[vz^2] \;-\; c^\top B^{-1} c. \tag{11}$$

Equivalently, $I_{\gamma|w}$ is the variance of the efficient score for γ after projecting out the w-score in the v-geometry; its reciprocal is the per-observation asymptotic variance factor for efficient estimation of γ.

Residual form. Define the v-weighted least-squares coefficient

$$a_v \;=\; \arg\min_{a \in \mathbb{R}^d} \mathbb{E}\big[v\,(z - a^\top m)^2\big] \;=\; B^{-1} c \tag{12}$$

(the normal equation $Ba_v = c$ is exactly $\mathbb{E}[vm(z - a_v^\top m)] = 0$) and the residual

$$r \;=\; z - a_v^\top m. \tag{13}$$

Expanding the quadratic form gives

$$\begin{aligned}\mathbb{E}[vr^2] &= \mathbb{E}\big[v(z - a_v^\top m)^2\big] \\ &= \mathbb{E}[vz^2] - 2a_v^\top \mathbb{E}[vmz] + a_v^\top \mathbb{E}[vmm^\top]\, a_v \\ &= \mathbb{E}[vz^2] - 2a_v^\top c + a_v^\top B a_v. \end{aligned} \tag{14}$$

Substitute $a_v = B^{-1}c$, so $Ba_v = c$ and $a_v^\top B a_v = c^\top B^{-1} c$:

$$\mathbb{E}[vr^2] \;=\; \mathbb{E}[vz^2] - 2\,c^\top B^{-1}c + c^\top B^{-1}c \;=\; \mathbb{E}[vz^2] - c^\top B^{-1}c. \tag{15}$$

Comparing equation 11 and equation 15 yields the identity used throughout the main text:

$$I_{\gamma|w} \;=\; \mathbb{E}[vr^2] \;=\; \mathbb{E}[vz^2] - c^\top B^{-1}c. \tag{16}$$

If B is singular the same identity holds with B^{-1} replaced by any generalized inverse on the range of c, provided $z \notin \operatorname{span}(m)$ in the v-geometry (otherwise $I_{\gamma|w} = 0$).

δ-bound. Residualization is a contraction in the v-inner product: choosing $a = 0$ in the least-squares definition shows

$$\mathbb{E}[vr^2] \;=\; \min_a \mathbb{E}\big[v(z - a^\top m)^2\big] \;\le\; \mathbb{E}[vz^2]. \tag{17}$$

Let $C = \{|z| > 0\}$ denote the signal support. Then $\mathbb{E}[vz^2] = \mathbb{E}[vz^2 \cdot \mathbf{1}_C]$. If $v \le \delta$ almost everywhere on C,

$$I_{\gamma|w} \;=\; \mathbb{E}[vr^2] \;\le\; \mathbb{E}[vz^2 \cdot \mathbf{1}_C] \;\le\; \delta\,\mathbb{E}\big[z^2 \cdot \mathbf{1}_C\big]. \tag{18}$$

This is the δ-bound (C1).

Tightness. Equality holds in equation 18 when both of the following are met:

1. *No residualization gap:* $c = \mathbb{E}[vmz] = 0$, hence $a_v = 0$ and $r = z$, so $\mathbb{E}[vr^2] = \mathbb{E}[vz^2]$.
2. *Entropy saturation of the bound:* $v = \delta$ almost everywhere on the support of $z^2 \cdot \mathbf{1}_C$.

Under those two conditions, $I_{\gamma|w} = \delta\,\mathbb{E}[z^2 \mathbf{1}_C]$. In general the residualization gap $c^\top B^{-1} c \ge 0$ makes the bound strict whenever z has a nonzero v-projection onto $\operatorname{span}(m)$.

Gaussian detection scale. Under a local Gaussian approximation to the likelihood-ratio (LR) statistic for $H_0 : \gamma = 0$, the standardized local shift after N i.i.d. draws is $\gamma\sqrt{N I_{\gamma|w}}$ (the corresponding noncentral χ^2_1 parameter is $N\gamma^2 I_{\gamma|w}$). The usual Gaussian approximation to the two-sided level-α boundary at power $1 - \beta$ is

$$|\gamma| \;\ge\; \frac{K_{\alpha,\beta}}{\sqrt{N I_{\gamma|w}}}, \qquad K_{\alpha,\beta} \;=\; z_{1-\alpha/2} + z_{1-\beta}, \tag{19}$$

and the δ-bound substitutes to give

$$|\gamma|_{\min} \;\gtrsim\; \frac{K_{\alpha,\beta}}{\sqrt{N \cdot \delta \cdot \mathbb{E}[z^2 \mathbf{1}_C]}}. \tag{20}$$

The operational quantity under a fixed labeling budget is thus $N \cdot I_{\gamma|w}$, not N alone.

A.2 C2: Sample-size insufficiency—precise form and counter-example

A common over-reading of equation 20 is that "tiny positive v cannot be overcome by sample size." We record the precise statement and the counter-example that refutes the over-reading.

Unrecoverable zero-entropy limit. If $v = 0$ almost everywhere on $\operatorname{supp}(z)$, contraction gives $0 \le I_{\gamma|w} \le \mathbb{E}[vz^2] = 0$. The Gaussian scale equation 19 is infinite for every finite N: no sample size restores nontrivial local power. (This is the $v{=}0$ case of C4 below.)

Vanishing-entropy sequence. Suppose instead $v \le \delta_N$ on C with a (possibly N-dependent) envelope $\delta_N > 0$, while the design geometry $\mathbb{E}[z^2\mathbf{1}_C]$ stays of constant order. Then equation 18–equation 20 give

$$|\gamma|_{\min} \gtrsim \frac{K_{\alpha,\beta}}{\sqrt{N\,\delta_N\,\mathbb{E}[z^2\mathbf{1}_C]}}. \tag{21}$$

This upper bound on information gives a lower bound on the detection boundary, so it establishes only the following necessary conditions:

1. *Inconclusive:* $N\delta_N \to \infty$. The lower bound vanishes, but this alone does not imply $|\gamma|_{\min} \to 0$; recoverability also requires residual signal mass outside the nuisance span.
2. *Boundary:* $N\delta_N \to \lambda \in (0,\infty)$. The lower bound remains of positive constant order $K_{\alpha,\beta}/\sqrt{\lambda\,\mathbb{E}[z^2\mathbf{1}_C]}$.
3. *Failure:* $N\delta_N \to 0$. The lower bound diverges, ruling out a shrinking detection boundary under the stated geometry.

A sufficient recoverability condition is a subset $C_0 \subseteq C$ on which $v \ge v_{\min} > 0$ and $\mathbb{E}[r^2\mathbf{1}_{C_0}] > 0$. Then

$$I_{\gamma|w} \;\ge\; v_{\min}\,\mathbb{E}[r^2\mathbf{1}_{C_0}] \;>\; 0 \qquad \Rightarrow \qquad |\gamma|_{\min} \;\propto\; N^{-1/2} \to 0. \tag{22}$$

Counter-example: tiny positive v *is* overcome by N. The informal claim "arbitrarily small positive v cannot be rescued by sample size" is therefore false. A finite-support weighted-information illustration uses four equally weighted atoms ($P(\{i\}) = 1/4$ for $i = 1,\dots,4$) in a scalar-nuisance design $m \in \mathbb{R}$ with no residualization gap:

i	1	2	3	4
m_i	0	0	1	-1
z_i	$+1$	-1	0	0
v_i	ε	ε	$1/4$	$1/4$

for an arbitrarily small $\varepsilon \in (0, 1/4]$. These are prescribed information weights, not the null probabilities implied by a no-intercept logistic model on the displayed m alone. Here $c = \mathbb{E}[vmz] = 0$, so $a_v = 0$, $r = z$, and

$$I_{\gamma|w} \;=\; \mathbb{E}[vz^2] \;=\; \tfrac{1}{4}\varepsilon\cdot 1 + \tfrac{1}{4}\varepsilon\cdot 1 + 0 + 0 \;=\; \tfrac{1}{2}\varepsilon \;>\; 0. \tag{23}$$

The Gaussian detection boundary is

$$|\gamma|_{\min} \;=\; \frac{K_{\alpha,\beta}}{\sqrt{N\cdot\varepsilon/2}} \;=\; \frac{K_{\alpha,\beta}\sqrt{2}}{\sqrt{N\varepsilon}}, \tag{24}$$

which tends to 0 as $N \to \infty$ for every fixed $\varepsilon > 0$. Thus "tiny but positive" entropy is a finite-sample budget issue in this design with nonzero residual signal, not an impossibility result by itself. For a fixed i.i.d. design, expected information remains linear in N even under extreme concentration. Concentration makes the per-label factor $I_{\gamma|w}$ small and can make finite samples variable in their information content. No finite N restores local power when $I_{\gamma|w} = 0$, including the $v{=}0$ limit and designs where residualization leaves $r = 0$.

A.3 C3: Constrained-optimal sampling and the KKT equal-marginal condition

Passive budget problem. Let $\mathcal{P} = \{(m_i, z_i)\}_{i=1}^n$ be a fixed unlabeled pool and let $q \in \mathbb{R}^n$ be a labeling distribution on pool indices subject to

$$\sum_{i=1}^n q_i = 1, \qquad 0 \le q_i \le \frac{1}{K}, \tag{25}$$

where the upper bound $1/K$ encodes sampling without replacement at budget K. Write $B(q) = \mathbb{E}_q[vmm^\top]$, $c(q) = \mathbb{E}_q[vmz]$, and

$$I(q) \;=\; \mathbb{E}_q[vz^2] - c(q)^\top B(q)^+ c(q) \;=\; \mathbb{E}_q\big[v\,r(q)^2\big], \tag{26}$$

with $r_i(q) = z_i - a_v(q)^\top m_i$ the residual in the q-reweighted v-geometry (or the pilot residual when the projection is held fixed).

Concavity. For each fixed $a \in \mathbb{R}^d$ the map

$$q \;\mapsto\; \mathbb{E}_q\big[v(z - a^\top m)^2\big] \;=\; \sum_{i=1}^{n} q_i\, v_i (z_i - a^\top m_i)^2 \tag{27}$$

is affine in q. Therefore

$$I(q) \;=\; \inf_{a \in \mathbb{R}^d} \mathbb{E}_q\big[v(z - a^\top m)^2\big] \tag{28}$$

is a pointwise infimum of affine functions, hence concave in q. (When the residual is frozen at a pilot value the infimum collapses to a single affine—in fact linear—objective $\sum_i q_i v_i r_i^2$.)

Envelope derivative. At any q for which the minimizer $a_v(q)$ is unique, Danskin's theorem (envelope theorem) yields

$$\frac{\partial I}{\partial q_i}(q) \;=\; v_i\, r_i(q)^2, \tag{29}$$

because the partial derivative of the Lagrangian in a vanishes at the optimum. Thus the marginal detection information of row i is exactly the per-row profiled information $I_i(q) = v_i r_i(q)^2$.

KKT system. Maximizing the concave objective $I(q)$ over the compact convex set defined by equation 25 is a convex program. Stationarity with multipliers $\lambda \in \mathbb{R}$ (equality $\sum q_i = 1$), $\mu_i \geq 0$ (lower bounds), and $\nu_i \geq 0$ (upper bounds $q_i \leq 1/K$) gives

$$v_i r_i(q^\star)^2 \;-\; \lambda \;+\; \mu_i \;-\; \nu_i \;=\; 0, \tag{30}$$

together with complementary slackness $\mu_i q_i^\star = 0$ and $\nu_i(1/K - q_i^\star) = 0$. Consequently every optimal $q^\star$ satisfies the equal-marginal condition (C3):

$$\begin{aligned} &v_i r_i(q^\star)^2 = \lambda \quad \text{if } 0 < q_i^\star < 1/K \quad \text{(free active)},\\ &v_i r_i(q^\star)^2 \geq \lambda \quad \text{if } q_i^\star = 1/K \quad \text{(saturated)},\\ &v_i r_i(q^\star)^2 \leq \lambda \quad \text{if } q_i^\star = 0 \quad \text{(inactive)}. \end{aligned} \tag{31}$$

This is *local profiled-Fisher optimality* for the linear/concave information objective under box constraints—not a claim of global finite-sample optimality of the downstream LR test, and not an unrestricted optimum over designs outside $\mathcal{P}$.

Bang-bang form under a frozen pilot. When residuals are evaluated at a fixed pilot, $I_i = v_i r_i^2$ is a constant vector and $I(q) = \sum_i q_i I_i$ is linear. An optimum of the linear objective over the capped simplex is bang-bang: $q_i^\star = 1/K$ on the K largest-I_i rows and 0 elsewhere, with canonical multiplier $\lambda = I_{(K)}$ (the K-th order statistic). Ties at the selection boundary permit multiple optima, including mixtures of tied selections. Any λ in the interval between the largest inactive and smallest saturated information is also valid; the implementation uses $\lambda = I_{(K)}$.

Why $q \propto vr^2$ is not optimal. Proportional allocation can place mass on lower-information rows and can violate the upper cap when a normalized row weight exceeds $1/K$. A two-row counter-example suffices. Take $I = (I_1, I_2) = (10, 1)$ and budget $K = 1$ (so the cap is 1):

$$q^\star = (1, 0) \quad \Rightarrow \quad I(q^\star) = 10, \qquad q^\propto = \left(\tfrac{10}{11}, \tfrac{1}{11}\right) \quad \Rightarrow \quad I(q^\propto) = \tfrac{101}{11} \approx 9.18 < 10. \tag{32}$$

With $K = 2$ and $I = (10, 9, 8, 0.1)$ the proportional rule still places mass on the third row, while the KKT optimum saturates only the top two. Hence $q \propto vr^2$ is strictly suboptimal for the capped linear objective.

Why the $|z|v$ heuristic can misorder. Detection uses vr^2 after residualization, not $|z|v$. The following four-row weighted projection illustrates an order reversal; its weights obey the Bernoulli variance bound. This algebraic example illustrates the ranking operation, rather than specifying a complete null-logit data-generating process. Let

$$m = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad z = (1,\, 0.6,\, 0.5,\, 0.5), \quad v = (0.1,\, 0.1,\, 0.25,\, 0.25). \tag{33}$$

The v-weighted normal equations give $a_v = (0.8,\ 0.5)$, hence

$$r = z - ma_v = (0.2,\ -0.2,\ 0,\ 0), \qquad I = vr^2 = (0.004,\ 0.004,\ 0,\ 0), \tag{34}$$

while

$$|z|\, v = (0.1,\ 0.06,\ 0.125,\ 0.125). \tag{35}$$

Rows 3–4 rank highest under $|z|v$ but carry zero detection information; rows 1–2 are optimal under vr^2. Ranking by $|z|v$ is therefore only a heuristic and can select a zero-information active set. The implementation uses the equal-marginal / top-K rule on pilot $\widehat{I}_i = \hat{v}_i \hat{r}_i^2$.

A.4 C4: Unrecoverable boundary

Statement. If

$$\mathrm{supp}(z)\ \subseteq\ \{m : v(m) = 0\} \qquad \text{almost everywhere under } P, \tag{36}$$

then for every sampling distribution q absolutely continuous with respect to P,

$$I_{\gamma|w}(q)\ =\ 0. \tag{37}$$

Derivation. Absolute continuity $q \ll P$ preserves null sets, so $\mathrm{supp}_q(z) \subseteq \{v = 0\}$ as well. On $\{v = 0\}$ one has $vz^2 = 0$ and $vmz = 0$ coordinates-wise, hence

$$\mathbb{E}_q[vz^2] = 0, \qquad \mathbb{E}_q[vmz] = 0. \tag{38}$$

The Schur identity equation 16 (or its generalized-inverse form) then yields $I_{\gamma|w}(q) = 0$. Equivalently, minimizing over projections gives $0 \le \mathbb{E}_q[vr(q)^2] \le \mathbb{E}_q[vz^2] = 0$, so the nonnegative residual-information integrand vanishes q-almost everywhere.

Logistic interior. Under a strictly logistic model with finite linear predictor, $v = \pi(1-\pi) > 0$ everywhere. Condition equation 36 is therefore a limiting / extended-support control (hard preference constraints, infinite logits, or an idealized $v \to 0$ truncation), not a generic interior regime. No reallocation of a passive budget can create detection information that the design does not contain; the zero-information bound is an analytic lower bound used as a validity check, not a claim about typical preference data.

Misspecification and judge noise. The zero-information bound is a statement about the working model equation 7. Covariate noise, label noise, or link misspecification change the *observed* law and can replace (v, z) by different functionals $(v_{\mathrm{obs}}, z_{\mathrm{obs}})$; the zero-information conclusion then applies to the observed working model, not necessarily to a latent well-specified one. Score-based effective information (Section 3.3) handles the noisy-covariate case operationally.

A.5 C5: Non-monotonic effect of judge noise

Judge noise may corrupt observed features, labels, or both. We record a local expansion that isolates two opposing forces on detection information.

Setup. Fix a latent linear predictor $L = w^\top m$ and write $v(u) = \sigma(u)(1-\sigma(u))$. Consider plug-in noisy covariates that shift the predictor to $L + s\xi$, where $s \ge 0$ is a noise scale and ξ is a mean-zero symmetric perturbation with $\mathbb{E}[\xi^2] = 1$ and $\mathbb{E}[|\xi|^4] < \infty$, independent of the design at the expansion point. (The same second-order structure appears for equicorrelated orientation-antisymmetric feature noise after linearization.)

Derivatives of logistic entropy. With $\sigma' = \sigma(1-\sigma) = v$,

$$v'(u) = v(u)\,(1 - 2\sigma(u)), \tag{39}$$

$$v''(u) = v(u)\,(1-2\sigma(u))^2 - 2v(u)^2\ =\ v(u)\,(1 - 6v(u)). \tag{40}$$

The identity $v'' = v(1-6v)$ follows by substituting $(1-2\sigma)^2 = 1 - 4\sigma + 4\sigma^2$ and $v = \sigma - \sigma^2$.

Second-order expansion of expected entropy. Taylor expansion in the noise scale gives

$$v(L + s\xi) \;=\; v(L) + s\xi\, v'(L) + \frac{s^2}{2}\xi^2 v''(L) + \frac{s^3}{6}\xi^3 v'''(L) + O(s^4|\xi|^4). \tag{41}$$

Taking expectations and using symmetry ($\mathbb{E}[\xi] = \mathbb{E}[\xi^3] = 0$) yields

$$\mathbb{E}\big[v(L + s\xi)\big] \;=\; v(L) + \frac{s^2}{2}\, v(L)\big(1 - 6v(L)\big) + O\big(s^4\mathbb{E}[|\xi|^4]\big)\,. \tag{42}$$

Hence:

- if $v(L) < 1/6$ (preference-saturated region), the s^2 term is positive and noise *raises* expected entropy (de-saturation);
- if $v(L) > 1/6$ (near $\pi = 1/2$; note $\max v = 1/4$), the s^2 term is negative and noise *lowers* expected entropy.

Two opposing contributions to I. Write a schematic noisy information functional $I(s) = \mathbb{E}[v_s r_s^2]$ (or its score-based counterpart $I_{\text{eff}}(s)$) with both entropy v_s and residual signal r_s depending on s. Differentiating under the integral,

$$I'(s) \;=\; \underbrace{\mathbb{E}\big[(\partial_s v_s)\, r_s^2\big]}_{\text{de-saturation (sign indefinite)}} \;+\; \underbrace{\mathbb{E}\big[v_s \cdot \partial_s(r_s^2)\big]}_{\text{signal attenuation (typically} \le 0)}\,. \tag{43}$$

The first term can be positive when mass sits in the saturated region (equation equation 42); the second term is negative whenever noise attenuates or misaligns the structured feature (and therefore the residual). Which force dominates depends on the noise law, the conflict geometry, and the operating γ. There is therefore no universal sign for $\partial|\gamma|_{\min}/\partial\sigma_J$: judge noise is non-monotone for detection (C5).

Empirical measurement layer. Separately from the population expansion, the nested LR instrument used in experiments can itself inflate type-I error under covariate noise. That is a property of the testing instrument, handled by the calibration gate in Appendix B, not a revision of the information identities above.

A.6 C6: Detection-boundary proxy

Gaussian LR starting point. Under the local Gaussian approximation of Appendix A.1,

$$|\gamma|_{\min} \;\approx\; \frac{K_{\alpha,\beta}}{\sqrt{N_{\text{eff}}\, I_{\gamma|w}}}, \tag{44}$$

with N_{eff} the effective count of independent measurement units under sampling q (row count for the synthetic power map; a separately estimated cluster-effective count for clustered designs) and $K_{\alpha,\beta} = z_{1-\alpha/2} + z_{1-\beta}$.

Interaction magnitude and preference variance. To describe the association between interaction magnitude and preference variance, we consider the distribution of pairs satisfying $I_\tau(m) = 1$. Let $\rho = \operatorname{corr}(|z|, v)$ denote their Pearson correlation. When the means and variances are nonzero, the normalized co-occurrence ratio is

$$\kappa_\rho \;=\; \frac{\mathbb{E}\big[|z|\, v\big]}{\mathbb{E}[|z|]\,\mathbb{E}[v]} \;=\; 1 + \rho \cdot \mathrm{CV}(|z|) \cdot \mathrm{CV}(v), \tag{45}$$

where all moments are taken under this conditional distribution and $\mathrm{CV}(X) = \mathrm{sd}(X)/\mathbb{E}[X]$. Thus, $\kappa_\rho < 1$ indicates a negative association between interaction magnitude and preference variance. This descriptive association does not by itself establish a reduction in $I_{\gamma|w} = \mathbb{E}[vr^2]$, which depends on the residualized interaction. Empirical estimates are reported in Section 4.2.

Factorization over the interaction support. For the interaction support $C = \{|z| > 0\}$, write $f_C = \Pr(C)$. In general, residual information can extend outside C:

$$\mathbb{E}[vr^2] = f_C \mathbb{E}_C[vr^2] + \mathbb{E}[vr^2 \mathbf{1}_{C^c}].$$

The residual $r = z - a_v^\top m$ need not vanish where $z = 0$. Under the additional condition $\mathbb{E}[vr^2 \mathbf{1}_{C^c}] = 0$, the factorization over C becomes

$$\mathbb{E}[vr^2] \;=\; f_C \cdot \mathbb{E}_C[vr^2] \;=\; f_C \cdot \bar{v}_C \cdot \overline{r_C^2} \cdot \widetilde{\kappa}_{v,r^2}, \tag{46}$$

where $\bar{v}_C = \mathbb{E}_C[v]$, $\overline{r_C^2} = \mathbb{E}_C[r^2]$, and the co-occurrence factor

$$\widetilde{\kappa}_{v,r^2} \;=\; \frac{\mathbb{E}_C[vr^2]}{\mathbb{E}_C[v]\,\mathbb{E}_C[r^2]} \tag{47}$$

equals 1 under $v \perp r^2$ on C and otherwise absorbs dependence between preference variance and the squared residual interaction. Note this v–r^2 co-occurrence $\widetilde{\kappa}_{v,r^2}$ is a *different* quantity from the $|z|$–v co-occurrence $\kappa_\rho = 1 + \rho \cdot \mathrm{CV}(|z|) \cdot \mathrm{CV}(v)$ (Eq. 45); the two are not numerically equal. Under this support condition, substitution into equation 44 produces the design proxy (C6)

$$|\gamma|_{\min} \;\approx\; \frac{K_{\alpha,\beta}}{\sqrt{N_{\text{eff}} \cdot f_C \cdot \bar{v}_C \cdot \overline{r_C^2} \cdot \widetilde{\kappa}_{v,r^2}}}. \tag{48}$$

Signal scale s_z. If the structured feature is rescaled $z \mapsto cz$ with $c \neq 0$, then $r \mapsto cr$ and $I_{\gamma|w} \mapsto c^2 I_{\gamma|w}$, so the boundary multiplies by $1/|c|$. Equation equation 48 already contains this scale through $\overline{r_C^2}$; multiplying by an additional $1/s_z$ would count the same rescaling twice. Numerical comparisons must nevertheless fix the scale convention for z (in our experiments, the frozen gate map z_τ at $\tau = 1/12$).

Why the proxy is a heuristic, not a data theorem. Every term on the right-hand side of equation 48 depends on modeling choices (w through v, the residual map through a_v, the conflict definition, and the noise law through the observed design). We use the full information in equation 44 unless the residual-support condition has been checked. Under that condition, C6 separates the roles of sample count, conflict geometry, and co-occurrence.

What is estimable from a frozen dataset. Given a frozen (m, z) design and a pilot nuisance $\hat{w}$ (hence $\hat{v}$), one can estimate without labels for γ:

- nonzero-interaction fraction f_C and empirical top-k shares of $\hat{v}\hat{r}^2$;
- mismatch diagnostics $(\hat{\rho}, \hat{\kappa}_\rho)$;
- plug-in $\widehat{I}_{\gamma|w} = \widehat{\mathbb{E}}[\hat{v}\hat{r}^2]$ and the factorization terms in equation 46.

What cannot be read off without further assumptions is the noise law that produced the observed panel, the true γ against which power is evaluated, and the correctness of the logistic link. In experiments the proxy (or its score-based sibling I_{eff}) is compared to Monte Carlo simulated detection boundaries under a stated data-generating process rather than claimed as an identity. The power-map diagnostic evaluates target powers of 20%, 35%, 50%, 65%, and 80% (Appendix B.3).

Score-based effective information under covariate noise. When the observed feature z_{obs} need not equal the latent signal s entering the label law, the well-specified information $\mathbb{E}[vr^2]$ can misstate detection power. The score-based correction used in the main text is

$$I_{\text{eff}} \;=\; \frac{\mathbb{E}[v\, r\, s]^2}{\mathbb{E}[vr^2]}, \tag{49}$$

with r the v-weighted residual of z_{obs} on m. When $s = z_{\text{obs}}$ one recovers $I_{\text{eff}} = I_{\gamma|w}$. The operational boundary estimator is then $|\gamma|_{\min} \approx K_{\alpha,\beta}/\sqrt{N_{\text{eff}} I_{\text{eff}}}$.

Pilot-information uncertainty (full-pipeline cluster bootstrap). The pilot allocation is a finite-sample estimate; its induced information $I(\hat{q})$ is therefore also random. To characterize its variation within the frozen pool, we resample prompt clusters with replacement ($B = 2000$, seed `20260825`; the independent sampling unit is the prompt cluster) and, on each resample, re-estimate the *entire* pilot pipeline—the stage-1 H0 logit refit (4-decimal contract against the frozen weights, verified exactly), the $\hat{v}$/$\hat{a}_v$/$\hat{r}$ maps, and the bang-bang top-10^4 selection—so that no pilot quantity is held frozen beyond the data. On 2,000/2,000 successful draws, the mean of $I(\hat{q})$ is 4.11×10^{-6}, with a 95% bootstrap percentile interval of $[2.25, 6.47] \times 10^{-6}$, bracketing the frozen-pool point $I(\hat{q}_0) = 4.11 \times 10^{-6}$; the per-row information ratio over the uniform allocation is 2.17 $[2.16, 2.18]$, stable across draws; and the selected top-10^4 set overlaps the frozen selection by a median 0.591 (min 0.447; 10^4 is 46% of the 21,750-row pool, so the overlap is a selection-stability diagnostic, not a set-identity claim). This resampling characterizes in-pool stability; it does not establish performance with an independent finite pilot. All outputs are `local_hypothesis` (`RUNN7PILOTBOOTSTRAP2026082501`).

B Supplementary Experiment Protocols

This section records the frozen measurement protocols behind Section 4. All Monte Carlo experiments use two-sided level $\alpha = 0.05$, `master_seed` $= 20260809$, and $R = 2000$ repetitions unless noted. Evidence role for reported empirical numbers remains `local_hypothesis`.

B.1 Type-I calibration gate

The nested logistic LR instrument is approximately calibrated at $\sigma_J = 0$ but inflates false-positive rate (FPR) under covariate noise. We therefore apply an explicit admission gate before any nominal-scale power comparison:

$$\text{FPR} \;\le\; 0.07 \;=\; \alpha + 4 \cdot \text{SE}(R{=}2000), \tag{50}$$

where the binomial standard error at the null is $\text{SE} = \sqrt{\alpha(1-\alpha)/R} \approx 0.00487$, so $\alpha + 4 \cdot \text{SE} \approx 0.0695$ (rounded to 0.07). A design cell enters confirmatory power comparisons only if its Monte Carlo FPR under $H_0 : \gamma = 0$ satisfies equation 50. Cells that fail the gate are reported as measurement failures, not as evidence about $I_{\gamma|w}$.

Empirically, over the 36 power-map design cells (asymptotic χ^2 threshold; N up to 50,000): mean FPR ≈ 0.052 at $\sigma_J = 0$ (calibrated); mean FPR ≈ 0.132 (max 0.548, at $N = 50{,}000$) at $\sigma_J = 0.1$; mean FPR ≈ 0.359 (max 0.999, at $N = 50{,}000$) at $\sigma_J = 0.25$. Table 3 reports the corresponding four-strategy grid ($N \le 10{,}000$), whose maximum is 0.867; the two maxima are drawn from different run grids. The asymptotic threshold therefore cannot support confirmatory power comparisons under covariate noise.

Null-calibration instrument (P0-1 revision). To restore type-I control at $\sigma_J > 0$ we re-run each $(\text{strategy}, N, \sigma_J)$ group at $\gamma = 0$, store the per-replicate LR statistics, and take their empirical 95% quantile as the rejection threshold for that group; $\gamma > 0$ replicates are judged against the same threshold, and invalid (non-converged/separated) fits never reject. The historical post-reconciliation diagnostic gives FPR 0.0500 on all 12/12 groups at $\sigma_J = 0.1$ by construction of the per-group empirical quantile and keeps FPR ≤ 0.07 at $\sigma_J = 0.25$ (run `RUN-N2-NOISE-ROBUST-20260810-01`). A cluster-robust Wald variant was evaluated and did not control FPR under this noise model (FPR 0.052–0.879), so null calibration is the primary instrument. The admission gate equation 50 remains as a post-hoc check on every reported cell. The I3 replay raises the split to $R = 4000$ and independently checks the first/last 2000 replicates: all 36 strategy–budget–noise validation groups pass FPR ≤ 0.07, yielding 12 gated Cond-Opt/Uniform cells. The resulting median paired gap is 0.581 (95% CI $[0.56325, 0.59451]$); the artifact is a calibrated local observation on the declared synthetic carrier and grid.

Structured-noise stress tests. Two additional models are evaluated separately. Cluster-correlated noise assigns 50% of its variance to a shared cluster component and 50% to a row component. Heteroscedastic noise doubles the baseline standard deviation where $|z_\tau| > 10^{-12}$. The grid combines Uniform and Cond-Opt, $N \in \{2000, 10000\}$, and $\sigma_J \in$

Table 3: Calibration scope. "Pass" requires FPR ≤ 0.07 in every evaluated group. Additive-noise asymptotic LR and Wald ranges use all 4000 null replications per group; empirical-null LR uses 2000 calibration and 2000 independent validation replications from the same current protocol-repair run. Its displayed range is the independent validation FPR, not the 5% rate imposed on the calibration sample. The two non-additive noise models are separate stress tests, each using 2000 calibration and 2000 validation replications per setting.

Instrument	Noise model	Status	FPR range
Asymptotic χ^2 LR	additive $\sigma_J = 0$	pass	$[0.042, 0.057]$
Asymptotic χ^2 LR	additive $\sigma_J \in \{0.1, 0.25\}$	fail	$[0.046, 0.867]$
Cluster-robust Wald	additive $\sigma_J \in \{0, 0.1, 0.25\}$	fail	$[0.042, 0.865]$
Empirical-null LR (primary)	additive $\sigma_J \in \{0, 0.1, 0.25\}$	pass	$[0.0355, 0.0670]$, 36/36
Empirical-null LR	cluster-correlated or gate-dependent variance	pass	$[0.039, 0.069]$, 16/16
Asymp. LR / Wald	cluster-correlated or gate-dependent variance	fail	$[0.059, 0.630]$
Cluster-resampled null gate	any	stated extension	—

$\{0.1, 0.25\}$, with 2000 calibration and 2000 validation replications per model and setting. All 16 calibrated validation rates lie in $[0.039, 0.069]$, while asymptotic LR or Wald reaches 0.6295. The source is `RUN-I5-NONADDITIVE-NOISE-20260914-01/nonadditive_calibration.json`; these simulations retain the source pool's stated cluster policy rather than resampling the empirical prompt clusters.

Validation direction: cluster-robust Type-I gate. The real pool's units are non-independent: 21,750 rows span 19,765 prompt clusters (§B.7), and the prompt cluster is the sampling and inference unit. A cluster-robust Type-I gate that resamples clusters with replacement under H_0 and re-estimates the admission threshold on each cluster-resampled world is therefore a natural extension of the row-level null-calibration instrument adopted here. We state this as a validation direction, not as a defect of the current instrument or as a new result.

B.2 Pilot protocol (no oracle leakage)

Constrained-optimal, DRTS, and APO arms compute selection scores from *frozen null weights only*. Let $\hat{w}_{H_0}$ denote the stage-1 null fit (numeric value used in the synthetic four-strategy run: $[4.3365,\ 2.0123,\ -0.2074,\ 0.3107,\ 0.0608]$ to four decimals). The pilot map is

$$\begin{aligned} \hat{\pi} &= \sigma(m^\top \hat{w}_{H_0}), \\ \hat{v} &= \hat{\pi}(1-\hat{\pi}), \\ z &= z_\tau(m) \quad (\tau = 1/12 \text{ frozen}), \\ \hat{a}_v &= \arg\min_a \sum_i \hat{v}_i (z_i - m_i^\top a)^2, \\ \hat{r} &= z - m\hat{a}_v, \\ \widehat{I}_i &= \hat{v}_i \hat{r}_i^2. \end{aligned} \tag{51}$$

No true γ, no H_1 fit, and no label-side oracle enters the pilot. The constrained-optimal allocation is the top-K rule on $\widehat{I}$ (Appendix A.3); DRTS (Double Reverse Thompson Sampling, here the paper-based proxy) ranks by near-certainty $|\hat{\pi} - 1/2|$; APO (Active Preference Optimization) ranks by $\hat{v}$. Active sets are fixed across replicates at a given budget (without-replacement draws within the selected set).

B.3 POWER-MAP GRID (324 CELLS) AND THE N=40000 SLOT

The analytic-vs-simulated boundary comparison sweeps a Cartesian grid

$$\begin{aligned} N &\in \{1000, 5000, 10000, 50000\}, \\ \gamma &\in \{0, \pm 0.25, \pm 0.5, \pm 1, \pm 2\} \quad (9 \text{ values}), \\ \sigma_J &\in \{0,\ 0.1,\ 0.25\}, \\ f_C &\in \{\text{natural},\ 0.02,\ 0.05\}, \end{aligned} \tag{52}$$

i.e. $4 \times 9 \times 3 \times 3 = 324$ cells, each with $R = 2000$ LR replicates on the frozen G3-derived carrier under $\tau = 1/12$.

N=40000 at the 50000 slot. Sparse-conflict arms ($f_C \in \{0.02, 0.05\}$) place $N = 40000$ at the nominal 50000 slot. Reason: under without-replacement sampling the $\tau = 1/12$ carrier supplies only about 41k non-conflict rows, so a stratified draw at $N = 50000$ with small f_C is infeasible. The natural-conflict arm retains $N = 50000$. The audit funnel is 324 executed cells $\to$ $19/36$ FPR-gate-surviving group arms (171 underlying cells) $\to$ 35 reachable boundary pairs at target powers of 20%, 35%, 50%, 65%, and 80% $\to$ 35 final comparisons to the I_{eff} surface (median relative error 1.97%, maximum 6.63%). The $26/26$ direction count is trajectory-level (17 N pairs, 3 calibrated-noise pairs, and 6 f_C pairs), not a second cell count. The I0 compare-only replay recomputed the current score-based I_{eff} consumer from the byte-preserved power map; an independent audit checked 1038 provenance/formula/report items and $36/36$ proxy groups. The separate $I_{\text{conf_only}}$ oracle failure is retained. These values are a synthetic, row-sampled local diagnostic characterizing the analytic boundary under controlled settings; they condition on the stated feature geometry and noise model.

B.4 FOUR-STRATEGY GRID (144 CELLS)

The noise-robust sampling comparison uses

$$\begin{aligned} &N \in \{2000, 5000, 10000\}, \quad \gamma^\star \in \{0.5, 1, 2\}, \quad \sigma_J \in \{0, 0.1, 0.25\}, \\ &\text{strategies} \in \{\text{uniform},\ \text{cond-opt},\ \text{DRTS},\ \text{APO}\}, \end{aligned} \tag{53}$$

together with $\gamma = 0$ calibration cells for every (strategy, N, σ_J) combination. Counting gives $3 \times 3 \times 3 \times 4 = 108$ powered cells plus $3 \times 3 \times 4 = 36$ null cells, i.e. 144 cells total, each with $R = 2000$ (run `RUN-N2-NOISE-ROBUST-20260810-01`). Control arms (conflict-only, max-$|z|$) and the small-budget grid ($N \in \{200, 500, 1000\}$) extend this protocol without changing the admission rule. The $N = 200$ empirical-null calibration is near-degenerate in the retained low-budget diagnostic, with a near-zero LR threshold; these cells are therefore not used as standalone confirmatory evidence.[1] P3 admission requires: both treatment and uniform reference calibrated under equation 50, and reference power in $[0.2, 0.8]$. The predeclared win rule is median $\Delta = \text{Power}_{\text{cond-opt}} - \text{Power}_{\text{uniform}} \geq 0.15$ with bootstrap 95% CI excluding 0.

B.5 CLUSTER BOOTSTRAP PROCEDURE

Power-difference CI (synthetic four-strategy diagnostic). Let $R_{c,r} \in \{0, 1\}$ denote the rejection indicator for comparison cell c on replicate $r = 1, \ldots, R$ ($R = 2000$). For bootstrap draw $b = 1, \ldots, B$ with $B = 2000$ and seed stream derived from `master_seed` = 20260809:

1. draw replicate indices uniformly with replacement within each cell (paired across strategies that share the cell index);
2. recompute per-cell power gaps and their median $\Delta^{(b)}$;
3. form the percentile CI from the 2.5 and 97.5 percentiles of $\{\Delta^{(b)}\}_{b=1}^{B}$.

The historical exploratory post-reconciliation median gaps under null calibration are $\Delta = 0.515$ ($\sigma_J = 0$, CI [0.4925, 0.537], n=3), $\Delta = 0.615$ ($\sigma_J = 0.1$, CI [0.5935, 0.6315], n=5), and $\Delta = 0.6645$ ($\sigma_J = 0.25$, CI [0.65, 0.678], n=4) (run `RUN-N2-NOISE-ROBUST-20260810-01`). The fresh

[1] The near-zero threshold at $N = 200$ is an instrument-level calibration condition, not evidence that the structured effect is absent.

I3 E3 replay gives a paired-index observation of median $\Delta = 0.581$ (95% CI $[0.56325, 0.59451]$) over 12 gated cells. The independent split-null audit passes all 36 validation groups; the result is reported as a calibrated local observation on the declared carrier, grid, and simulated-label regime.

Real-anchor simulation and cluster-aware uncertainty. E5 generates row-indexed Bernoulli labels on frozen covariates and applies the nested logistic LR test with its asymptotic χ^2_1 threshold. The $\gamma = 0$ simulation checks this rejection rule, giving I3 FPR 0.053 for Uniform and 0.057 for Cond-Opt. Prompt-cluster identifiers enter the coefficient standard-error calculation; they do not turn the LR threshold into a cluster-bootstrap threshold. The power-gap CI resamples aligned Monte Carlo replicate indices across strategies. Cluster resampling of the pilot information map is a separate N7 analysis (Appendix A.6).

DRTS information diagnostic. The fit-versus-detect separation reports mean pilot information on DRTS-selected rows versus the population mean, with a row-level bootstrap CI on the difference ($B = 2000$, same master seed). The synthetic selector diagnostic gives $\mathbb{E}_{\mathrm{DRTS}}[vr^2] = 4.34/4.41/4.71 \times 10^{-5}$ at $N = 2000/5000/10000$, versus population 7.68×10^{-4} (approximately 16–18$\times$ lower; each difference CI excludes 0).

B.6 Same-space Fisher baseline and selected-set safety audit

To isolate the contribution of residualization, we ran a separate paired contrast on the same $401{,}734$-row synthetic carrier. The six arms were Uniform, max-$|z|$, vz^2, the original frozen Cond-Opt, a bounded Cond-Opt+repair arm, and a one-shot Fisher-leverage score. For the latter, $x_i = (m_i, z_i)$ and

$$s_i^{\mathrm{Fisher}} = v_i x_i^\top \left(\frac{1}{n} \sum_j v_j x_j x_j^\top \right)^+ x_i.$$

This is a same-feature-space comparator; it is not a reproduction of an iterative active-reward implementation. Each (K, σ_J) group used 2000 null replications for threshold calibration and 2000 independent null replications for validation, followed by 2000 shared-world alternative replications at $\gamma^\star \in \{0.5, 1, 2\}$. All 36 validation groups passed FPR ≤ 0.07 (maximum 0.066), and every fit-validity denominator was 2000. At $\gamma^\star = 0.5$, the Cond-Opt–Uniform deltas were 0.388–0.645 across the six (K, σ_J) groups, compared with 0.274–0.508 for Fisher-leverage; the paired intervals and complete power grid are in Table 4. At $\gamma^\star = 1$ the two arms were often close, and at $\gamma^\star = 2$ they were saturated. The repair arm remained close to the original Cond-Opt, so it is retained as a local information-improvement variant without a separate performance claim. The complete run and independent audit are `RUN-BASELINE-CONTRAST-FORMAL-20260921-01` and `formal_audit.json`; both retain the local-hypothesis evidence role.

The label-free selected-set audit reports identifiability separately from projection fidelity. It refits the v-weighted projection on the selected rows, checks $\mathrm{rank}(m_S) = d$ and $\mathrm{rank}([m_S, z_S]) = d + 1$, requires positive profiled information, and uses $I(S)/\widetilde{I}(S) \geq 0.95$ as the projection-fidelity threshold. The exact six-row counterexample in the review pool has frozen score $1/81$ but reprojected selected information below 10^{-32}; the audit flags this failure, whereas an alternative four-row set has information $1/112$. On the current HelpSteer2+3 pool, the original top-K sets pass at $K = 2000, 5000, 10000$ with retentions 0.9868, 0.9983, and 0.99995. On the synthetic carrier, the original rule has retentions 0.9250, 0.9521, and 0.9105; the bounded repair improves these to 0.9816, 0.9697, and 0.9242, leaving the largest budget below the fidelity threshold. All sets remain in the power comparison, including those below this threshold. The original rule remains a required arm and the repair is described only as a local heuristic; the audit detects a failure but does not turn local exchanges into a global optimum.

B.7 Data card (real anchor)

Data. HelpSteer2 (nvidia/HelpSteer2, 6,586 pooled rows) and HelpSteer3 (nvidia/HelpSteer3 preference, 15,164 pooled rows), both CC-BY-4.0, with signed ordinal preference strength $\{-3, \ldots, 3\}$

Table 4: Same-space baseline contrast on the synthetic carrier. Entries are LR rejection power (percent) from $R = 2000$ alternative replications after separate $R = 2000$ null calibration and validation. All six strategies share row-indexed label/noise worlds; "Fisher" is the one-shot full-parameter Fisher-leverage score, not a reproduction of an iterative active-reward pipeline. The final column is the largest validation FPR across the six arms for the (K, σ_J) group.

σ_J	K	$\gamma^\star$	Uniform	$\max\|z\|$	vz^2	Fisher	Cond-Opt	Repair	max FPR
0.0	2,000	0.5	16.4	22.2	18.4	43.8	55.1	58.8	0.056
0.0	2,000	1	49.8	67.8	60.8	95.9	99.2	99.5	0.056
0.0	2,000	2	97.3	99.8	99.6	100.0	100.0	100.0	0.056
0.0	5,000	0.5	33.7	45.4	45.4	80.8	93.2	93.5	0.059
0.0	5,000	1	83.4	96.2	95.5	100.0	100.0	100.0	0.059
0.0	5,000	2	100.0	100.0	100.0	100.0	100.0	100.0	0.059
0.0	10,000	0.5	55.6	78.8	78.0	95.2	99.4	99.2	0.066
0.0	10,000	1	99.1	100.0	100.0	100.0	100.0	100.0	0.066
0.0	10,000	2	100.0	100.0	100.0	100.0	100.0	100.0	0.066
0.1	2,000	0.5	14.8	20.7	17.9	44.5	60.0	60.2	0.059
0.1	2,000	1	36.2	59.1	50.9	92.7	99.2	99.3	0.059
0.1	2,000	2	81.3	98.9	96.8	100.0	100.0	100.0	0.059
0.1	5,000	0.5	28.9	41.4	36.3	76.3	93.5	93.0	0.060
0.1	5,000	1	65.1	91.5	87.3	99.9	100.0	100.0	0.060
0.1	5,000	2	99.4	100.0	100.0	100.0	100.0	100.0	0.060
0.1	10,000	0.5	39.0	69.2	70.2	89.7	99.1	99.0	0.064
0.1	10,000	1	88.2	99.8	99.6	100.0	100.0	100.0	0.064
0.1	10,000	2	100.0	100.0	100.0	100.0	100.0	100.0	0.064

(0 excluded from the binary likelihood); the symbol contract maps strength $< 0 \Rightarrow Y{=}1$, strength $> 0 \Rightarrow Y{=}0$; both datasets follow the same convention.

Pooling and cluster unit. $21{,}750$ rows over $19{,}765$ prompt clusters (ratio 1.10 rows/cluster), after cross-source deduplication: HelpSteer3 context lists and HelpSteer2 prompt strings were matched explicitly (0 shared items; 182 shared prompts merged), and cluster IDs were unified across sources. Prompt clusters identify dependence for real-label splitting and coefficient uncertainty. E5 samples rows under conditionally independent simulated labels and checks an asymptotic LR rule; N7 separately resamples clusters to assess pilot information.

Noise-scale anchor. The mean cross-expert standard deviation is 0.1077318532 on the raw 0–4 scale. The frozen feature normalization gives

$$\hat{\sigma}_J = \frac{0.1077318532}{4-0} = 0.0269329633 \approx 0.03.$$

The source is the N1 information-efficiency run's `sigma_J_anchor_measurement.json`. This anchor describes the measured scale, not all possible annotation-noise regimes.

Judge panel. Three frozen LLM judges, each producing label-free rubric scores per dimension: gpt-oss-20b, Qwen3-30B-A3B, and Qwen2.5-Math-7B-Instruct, with 100% coverage of the pooled rows (gpt-oss shard truncation leaves a documented 63% rubric-coverage residual on a subset; agreement statistics use available pairwise scores). Mean pairwise judge disagreement is 0.739 on signal rows versus 0.547 on non-conflict rows (Spearman coupling with $|z|$: $+0.1153$, $p \approx 2.8 \times 10^{-65}$).

Structured feature. $z = z_\tau(m) = I_\tau(m) \cdot \max_k m_k \cdot \min_k m_k \cdot (\max_k m_k + \min_k m_k)$, with I_τ the conflict indicator $\{\max_k m_k > \tau\} \cap \{\min_k m_k < -\tau\}$ and the frozen threshold $\tau = 1/12$ inherited from the prior instrument contract; $m \in \mathbb{R}^5$ is the dimension-difference vector from the judge panel (helpfulness, correctness, coherence, complexity, verbosity). 991 rows satisfy the strict conflict gate; 520 of these have nonzero z.

UltraFeedback subset. We sample 3000 unique prompts without replacement from the `train_prefs` split of HuggingFaceH4/ultrafeedback_binarized using seed 20260810, retaining one response pair per prompt. Rubric scores from available judges are transformed and averaged using the same feature-construction procedure as for HelpSteer2+3. All pairs have scores

from at least two judges: 2954 have scores from all three and 46 from two. The Qwen3-30B-A3B judge uses the GPTQ-Int4 checkpoint for this subset; the other two judge checkpoints match the HelpSteer panel. The reported information diagnostics use the fixed HelpSteer null-pilot coefficients. Dataset sources are UltraFeedback (Cui et al., 2024) and its binarized derivative (Hugging Face H4, 2023). The subset contains 149 conflict-gated comparisons and 91 with nonzero interaction features. Its sampling manifest, judge records, and analysis are archived under `RUN-N1-SECOND-CORPUS-20260810-01`.

Full-pool association statistics. Interaction magnitude is $|z|$. Pairwise judge disagreement is the fraction of available judge pairs with different signs of their summed rubric-score differences; zero versus nonzero signs count as disagreement. HelpSteer2+3 has gate-conditional Pearson $\rho = -0.2865$ (under the frozen null weights) and full-pool Spearman correlation 0.1153 with disagreement (row-level $p = 2.77 \times 10^{-65}$). UltraFeedback has the corresponding values -0.4741 and 0.1255 (row-level $p = 5.32 \times 10^{-12}$). These p-values are not adjusted for prompt clustering. For UltraFeedback, an equal-information reference on its 91 nonzero-interaction comparisons places $30/91 \approx 33.0\%$ in the top 1%; the measured share is 89.8%. The HelpSteer2+3 reference is $217/520 \approx 41.7\%$, compared with a measured share of 92.6%.

Features for controlled simulations. A separate pool contains 401,734 pairs of 4096-dimensional response embeddings from the archived RLHFlow/pair_data_v2_80K_wsafety source (RLHFlow, 2024). We construct five-dimensional features using a fixed random matrix $W \in \mathbb{R}^{5\times 4096}$ with orthonormal rows. For pair i, let e_{iA}, e_{iB} denote its embeddings and $\epsilon_i \in \{-1, +1\}$ a fixed, independently sampled orientation sign. The feature map before numerical clipping is

$$m_{ik} = \tanh\left(\frac{\epsilon_i[W(e_{iA} - e_{iB})]_k}{c_k}\right), \qquad k = 1, \ldots, 5,$$

where c_k is the sample standard deviation of the projected differences in coordinate k, computed on the first 10,000 pairs. The projection is obtained by orthonormalizing a Gaussian matrix with seed 20260805; orientation signs use seed 20260807. Components are clipped to the float32 interior of $(-1, 1)$ before storage. The mapping uses embeddings and pair identifiers without reading preference labels or reward scores. The archived mapping record specifies the projection, scales, orientation signs, and array hashes.

Pilot. v and r for all selection rules are computed from the frozen null fit $\hat{w}_{H_0} = [4.3365, 2.0123, -0.2074, 0.3107, 0.0608]$ (reproduced to four decimals), with no H_1 fit and no label-side oracle; pilot–DGP information alignment Spearman $= 0.905$.

Dependence and sample counts. Prompt clusters define the split and covariance units for observed-label analyses. The ratio of 1.10 rows per cluster alone does not determine an effective sample size. The power-map simulations use their stated row-based data-generating process; observed-label boundary diagnostics also use selected row counts and are not cluster-robust power guarantees.

Artifacts. Frozen mapping *m.npy* / *item_ids.npy* and the mapping contract (SHA256 recorded) live in `RUN-S3-REPAIR-CHAIN-20260805-01/t3_mapping/`; the pooled analysis set in `RUN-S2-POOLING-20260808-05/audit_fix/`; all experiment runs record commands, configs, seeds, logs, and raw outputs under `docs/research/runs/`.

B.8 E4: TINY-v VALIDITY CONTROL (DGP AND FULL RESULTS)

DGP. 2000 rows, 10% signal rows, maximum $v \approx 1.8 \times 10^{-3}$ on signal rows, $\gamma \in \{0.5, 1, 2\}$, R=2000 seeded replicates (`RUN-N1-INFO-EFFICIENCY-20260808-01/t5_unrecoverable`). This is the tiny-positive-v regime of C2 ($\gamma_{\min} \approx 100$ at N=2000), not the exact v=0 C4 boundary.

Results. Empirical $I = 4.4 \times 10^{-7} > 0$; per-$\gamma$ powers 0.0145 / 0.0125 / 0.013, flat in γ and below the nominal 0.05 (near-zero-information LR). $I \leq \delta$-bound 7.6×10^{-7}; the exact v=0 case gives $I = 0$ by C4. All 2000/2000 fits are valid at each effect size. In HelpSteer2+3, the 520 comparisons with nonzero interaction features have median $v \approx 0.20$ under the frozen null weights, compared with $v \approx 0.223$ across all 991 conflict-gated comparisons. The control therefore characterizes a different variance regime; it does not establish the cause of limited detection in the empirical pool.

B.9 E6: REAL-LABEL DETECTION PILOT (PROTOCOL)

Run. `RUN-P0-REAL-LABEL-PILOT -20260827-01` (script `p0_real_label_pilot.py`; R=500 replicates, master seed `20260809`, 8 workers; runtime 468 s). **Protocol.** Per replicate: (i) prompt clusters are split $50/50$ into pilot/eval by a seeded permutation; (ii) the pilot stage fits H_0 (γ=0 logit on m) on pilot rows only, computes $\hat{v}_i$ from $\hat{w}_{H_0}$ on eval rows, forms $\hat{r}_i$ from the v-weighted projection $\hat{a}_v$ fit on pilot rows, and ranks eval rows by $\widehat{I}_i = \hat{v}_i \hat{r}_i^2$; the pilot never sees eval labels; (iii) Cond-Opt takes the top-N eval rows by $\widehat{I}_i$, Uniform takes N seeded-random eval rows, for $N \in \{2000, 5000, 10000\}$; (iv) the eval stage runs the frozen nested-LR instrument ($\tau = 1/12$, real prompt clusters) on the selected rows with real judge labels, reporting $\hat{\gamma}$, CI, LR statistic, and asymptotic rejection; (v) the realized selected-set information I_{eff} (pilot $\hat{v}$, frozen z) gives a row-level $|\gamma|_{\min}$ diagnostic at 50% power via the C6 proxy; because the implementation uses selected row count, this is not a cluster-effective boundary or a cluster-robust power guarantee.

Null diagnostic. A row-wise y-permutation diagnostic (shuffle labels over the selected rows, B=100, first 25 replicates per arm/budget) produces rejection rates of 4.5–6.4% versus nominal 5%. It is not a formal nested-H_0 Type-I calibration: the row-wise permutation breaks the additive nuisance relation and does not preserve cluster dependence.

Fit validity. Across the 3,000 real-label replicate-arm-budget records, 2,959 fits are valid and 41 are marked separation. Cond-Opt has no invalid fits; Uniform has $39/500$ invalid fits at N=2000, $2/500$ at N=5000, and none at N=10000. Rejection rates in the table use valid-fit denominators.

Reference. Full-pool re-fit in the same run: $\hat{\gamma} = 1.253$ ($p = 0.80$, CI $[-8.15, 10.66]$), against the frozen reference 1.2540097738077371 (`RUN-S2-POOLING-20260808-05`); the two agree to three significant figures, and the main text quotes the frozen reference.

Evidence status. This pilot records non-detection at the measured scale, matching the analytic boundary's prediction for the sub-threshold full-pool $\hat{\gamma}$; the selected-set diagnostic boundary falls from 37.27 (Uniform) to 13.77 (Cond-Opt) at $N = 2000$, and from 21.53 to 13.68 at $N = 5000$, showing that Cond-Opt concentrates measured detection information on this observed-label geometry.

B.10 E7: HELD-OUT DETECTABILITY SCREEN (PROTOCOL AND FULL SCAN)

Split. The 19,765 prompt clusters are partitioned once into a scan half A and a confirmation half B by a seeded random permutation of the sorted unique prompt-cluster identifiers (seed `20260914`): $|A|$ = 9,882 clusters (10,935 rows), $|B|$ = 9,883 clusters (10,815 rows). The split is fixed before any label on B is examined; all exploratory selection below uses A only.

Candidate family. Fourteen prespecified effects use the same strict conflict gate as Section 3: with $u = \max_k m_k$ and $\ell = \min_k m_k$, define $I_\tau(m) = \mathbf{1}\{u > \tau, \ell < -\tau\}$. The two forms are the gated product interaction $z_\tau(m) = I_\tau(m) u\ell(u + \ell)$ and the gate indicator $h_\tau(m) = I_\tau(m)$, each evaluated at $\tau \in \{0, 1/24, 1/16, 1/12, 1/8, 1/6, 1/4\}$. Several thresholds induce identical features on the discrete observed values; the table retains all fourteen declared candidates.

Selection rule. For each candidate, the specified-effect fit $\hat{\gamma}_A$, its cluster-robust standard error, the pooled detection information $I_A = \text{mean}(\hat{v}\, r^2)$, the 50%-power boundary $|\gamma|_{\min} = 1.96/\sqrt{N_A I_A}$, and the detectability margin $|\hat{\gamma}_A|/|\gamma|_{\min}$ are computed on A. The single candidate maximizing the margin is advanced to confirmation. As a descriptive secondary diagnostic, the candidate maximizing the scan-side Wald $|z|$-score is also evaluated on B; it is not part of the selection rule.

Confirmation rule. One prespecified cluster-robust Wald test of $\gamma \neq 0$ is used for formal confirmation on B, for the selected candidate. Confirmation requires $|z| > z_{0.975} \approx 1.96$. The additional Wald-leader and original-$\tau = 1/12$ fits on B are descriptive diagnostics and do not enter this decision.

Confirmation outcome. The selected candidate (product, τ=1/4) fits $\hat{\gamma}_B = 12.71$ with se $= 11.48$ on B ($z = 1.11$, LR $p = 0.38$); its B-side information gives a boundary of 32.2. The descriptive Wald leader (gate, τ=1/4; $z_A = 2.05$) gives $|z_B| = 0.81$. The selected candidate is not confirmed, and the auxiliary Wald-leader check is also non-significant. **Multiple comparisons.** The scan evaluated fourteen candidates on A and one prespecified confirmation test on B; the scan

Table 5: E7 scan half (A, 10,935 rows). "$\hat{\gamma}$" is the specified-effect fit; $I = \mathrm{mean}(\hat{v}r^2)$ is the pooled detection information; boundary is $1.96/\sqrt{N_A I}$; margin is $|\hat{\gamma}|$/boundary. The product column at $\tau \leq 1/16$ and the gate columns at adjacent τ are identical features on the frozen scan sample. Selected row: τ=1/4 product (margin 1.56).

τ	Form	$\hat{\gamma}_A$	$\lvert z_A \rvert$	z-nonzero	I_A	Boundary	Margin
0	product	0.33	0.06	7.883%	2.76e−6	11.27	0.03
0	gate	−0.043	0.71	11.285%	2.57e−2	0.12	0.37
1/24	product	0.33	0.06	7.883%	2.76e−6	11.27	0.03
1/24	gate	−0.043	0.71	11.285%	2.57e−2	0.12	0.37
1/16	product	0.33	0.06	7.883%	2.76e−6	11.27	0.03
1/16	gate	−0.043	0.71	11.285%	2.57e−2	0.12	0.37
1/12	product	0.584	0.10	2.378%	2.51e−6	11.83	0.05
1/12	gate	0.079	0.77	4.563%	9.56e−3	0.19	0.41
1/8	product	0.584	0.10	2.378%	2.51e−6	11.83	0.05
1/8	gate	0.079	0.77	4.563%	9.56e−3	0.19	0.41
1/6	product	−4.33	0.56	0.338%	2.01e−6	13.22	0.33
1/6	gate	0.344	0.95	0.430%	8.69e−4	0.64	0.54
1/4	product	−26.1	1.01	0.082%	1.25e−6	16.73	1.56
1/4	gate	1.62	2.05	0.165%	3.02e−4	1.08	1.50

stage is exploratory, and its nominal statistics (including the $z_A = 2.05$ gate candidate) carry no confirmatory weight.

Interpretation limits. The held-out result concerns the selected candidate and does not establish non-detectability of all fourteen candidates, absence of interactions, or failure at a larger budget. The selected candidate has the largest scan-side coefficient magnitude and smallest information, illustrating why a large estimate alone cannot establish detection. Information varies with τ within each feature form; product and gate features have different scales, so their raw information values do not isolate threshold sensitivity. The concentration findings in Section 4.2 are conditional on the fixed $\tau = 1/12$ feature map.

Positive control (semi-synthetic injection). To examine the pipeline under specified signal injections, we inject the paper's specified effect (product interaction, τ=1/12) into regenerated labels on the frozen real geometry: $y_i \sim \mathrm{Bernoulli}(\sigma(\hat{w}_{\mathrm{ref}}^{\top} m_i + \gamma_{\mathrm{inj}} z_i))$, where $\hat{w}_{\mathrm{ref}}$ is the H_0 fit on the real labels of scan half A only and $\gamma_{\mathrm{inj}} \in \{0.5, 2.0\}\times$ the real-label boundary (11.83). The identical split, candidate family, selection rule, and confirm rule are then rerun on the synthetic labels. At $\gamma_{\mathrm{inj}} = 5.9$ the largest scan margin is 0.64 and the held-out test does not reject ($z = 1.36$); at $\gamma_{\mathrm{inj}} = 23.7$ the scan selects the τ=0 product feature—a sibling that shares support with the injected feature and carries its diluted signal (the injected feature itself ranks fourth, margin 3.23, behind three tied product siblings at 3.69)—and the held-out test confirms it ($\hat{\gamma}_B = 22.2$, se $= 7.2$, $z = 3.08$, boundary 15.9). The selected feature carries signal but differs from the injected feature, so its fitted coefficient does not estimate the injected coefficient. Each injection level uses one generated label realization; the two outcomes are procedure checks rather than estimates of power. Run: `RUN-I5-POSITIVE-CONTROL-20260914-01`; semi-synthetic labels, controlled procedure check.

C Implementation Checks

The derivations in Appendix A and protocols in Appendix B define the statistical instrument. The following cross-check assesses its numerical implementation.

Framework cross-validation. The in-house logistic fitter, cluster-robust sandwich SE, and nested LR statistic are numerically validated against `statsmodels` 0.14 (`GLM` binomial; `cov_type='cluster'`) on synthetic data: coefficients and log-likelihood agree to $< 10^{-8}$, the cluster-robust SE to $< 1\%$, and the LR statistic exactly; the in-house separation detector is kept because `statsmodels` only warns (and its coefficients explode) under perfect separation, whereas

the simulation pipeline requires an explicit per-replicate failure state (test `tests/unit/test_statsmodels_crosscheck.py`).

D SUPPLEMENTARY EXPERIMENT TABLES

Table 6: Information concentration and detection-boundary diagnostics. The upper panel summarizes information concentration and the association between interaction magnitude and preference variance. The lower panel compares boundaries based on the full score-based information I_{eff} with simulated boundaries at target powers of 20%, 35%, 50%, 65%, and 80%, subject to the stated error-control and power-bracketing criteria. Reported false-positive rates use asymptotic LR thresholds.

Information concentration	**HelpSteer2+3**	**Synthetic carrier**
Top-1% information share	92.6%	6.9%
Top-0.4% information share	86.7%	3.1%
ρ: all / nonzero-interaction comparisons	-0.0456 / -0.1495	-0.0072 / -0.3356
ρ: conflict-gated comparisons	-0.2865	—
Nonzero-interaction proportion f_C	$\approx 2.4\%$	$\approx 89.7\%$
Analytic versus simulated detection boundaries		
Simulation grid / groups meeting the FPR criterion	324 cells / 19 of 36 groups (171 cells)	
Eligible boundary comparisons	35, after FPR and 0.2–0.8 power-bracketing criteria	
Median / maximum rel. error	1.97% / 6.63%	
Mean FPR, $\sigma_J = 0/0.1/0.25$	0.052 / 0.132 / 0.359	

The tables below retain historical per-cell and control-arm records for provenance. They do not expand the current simulation results reported in Sections 4.3 and 4.3; those results and their sources are recorded separately in Table 14. All allocation contrasts remain diagnostic observations.

Reporting protocol. The historical numbers in Tables 8 and 9 were generated by a deterministic reporting protocol (run `RUN-N5-REPORTING-PROTOCOL-20260821-01`; protocol-file SHA-256 prefix `d00308f7990a1f82`). Generation is a single deterministic function of the per-replicate artifacts (R=2000 reject indicators) of `RUN-N2-NOISE-ROBUST-20260810-01`, `RUN-N2-CONTROL-ARMS-20260810-01`, `RUN-N2-SMALL-BUDGET-FIXED-20260810-01`, and `RUN-N2-CONSTRAINED-OPTIMAL-20260808-01`. However, post-freeze reconciliation changed stored rejection indicators and invalidated their hashes against the frozen manifest. Deterministic generation from those artifacts therefore does not establish an immutable raw-to-table chain; the affected E3 and E5 values are retained only as historical exploratory diagnostics. The fresh I3 replay provides the current verified provenance record: its independent E3 split-null and E5 paired-array audits pass, with both benchmarks scoped to the declared synthetic/simulated domains.

P1-1 reward-subset fit columns (recompute provenance). In the historical reward-subset diagnostics, DelibRM denotes the structured score combining a frozen main-effect estimate with a fitted interaction coefficient and the specified conflict gate. With the same frozen main-effect estimate, setting $\gamma = 0$ recovers the B-FUSION reference; B-WFUSION is the comparator with trainable main-effect weights.

All fit metrics of Table 12 (accuracy, AUC; trained B-WFUSION and untrained B-FUSION reference) are deterministic recomputations from the frozen raw of `RUN-N6-REWARD-SUBSET-20260824-01`: archived head weights scored on the archived hold-out (frozen split, selection, weights, and data arrays untouched; exact average-rank AUC, float32), with the accuracy 95% CI from a row-resampling bootstrap (B=2000, seed `20260825`). This fit-column recomputation does not validate the retained historical γ diagnostics; those fields remain historical exploratory diagnostics pending their own raw replay.

The following historical table is superseded for current provenance by the verified I3 paired bundle: on the frozen geometry, simulated powers at $\gamma = 8/12/16$ are $0.2415/0.4515/0.6690$ (Uniform)

Table 7: Current independently calibrated four-strategy grid. Power is reported in percent for all 27 settings, using 2000 evaluation replications per arm and setting. The 12 settings within the prespecified Uniform-power band form the summary reported in Section 4.3. Source: the original LR arrays and split-null audit in the current protocol-repair record. Historical tables below retain their original values and provenance.

N	γ	σ_J	Uniform	DRTS proxy	APO	Cond-Opt
2,000	0.5	0	17.60	16.10	11.80	60.45
2,000	0.5	0.1	12.40	12.35	9.85	63.00
2,000	0.5	0.25	7.50	7.75	6.30	45.55
2,000	1	0	47.75	43.00	27.85	99.45
2,000	1	0.1	28.25	27.95	20.60	99.50
2,000	1	0.25	11.45	13.10	8.80	92.15
2,000	2	0	96.85	94.70	72.80	100.00
2,000	2	0.1	76.90	72.95	57.75	100.00
2,000	2	0.25	21.75	25.35	14.80	100.00
5,000	0.5	0	33.05	29.15	21.90	91.35
5,000	0.5	0.1	29.35	22.50	15.25	89.00
5,000	0.5	0.25	8.65	11.05	8.75	63.50
5,000	1	0	84.35	81.60	66.10	100.00
5,000	1	0.1	68.30	59.00	45.85	100.00
5,000	1	0.25	16.50	18.80	15.40	99.25
5,000	2	0	100.00	100.00	99.85	100.00
5,000	2	0.1	99.55	98.15	94.15	100.00
5,000	2	0.25	37.85	46.15	33.05	100.00
10,000	0.5	0	53.00	52.45	47.15	99.20
10,000	0.5	0.1	40.65	38.60	34.20	98.55
10,000	0.5	0.25	13.55	12.70	13.40	79.15
10,000	1	0	98.75	98.50	94.70	100.00
10,000	1	0.1	87.70	86.25	78.20	100.00
10,000	1	0.25	28.25	30.60	23.35	100.00
10,000	2	0	100.00	100.00	100.00	100.00
10,000	2	0.1	100.00	100.00	99.95	100.00
10,000	2	0.25	64.85	70.95	56.70	100.00

and $0.4805/0.7840/0.9560$ (Cond-Opt). Shared row-indexed label worlds and paired replicate arrays give a median gap 0.287 (95% CI $[0.2675, 0.3070]$) and raw interpolation boundaries 12.89195 and 8.25700; this remains simulated-label evidence, not real-label power.

Table 13: Historical exploratory post-reconciliation diagnostic record of real-anchor reweighting per-cell detail (E5). Budget N=10000, R=2000. Boundary bootstrap uses the two bracketing grid cells around 50% power.

Quantity	Uniform	Cond-opt
Power at γ=8	0.277	0.62
Power at γ=12	0.483	0.92
Power at γ=16	0.668	0.985
Per-cell Δ $(8/12/16)$	0.343 / 0.437 / 0.317	
Median Δ	0.343 (CI $[0.3185, 0.371]$)	
50% boundary (interp.)	$\gamma \approx 12.4$	$\gamma \approx 6.5$
50% boundary (bootstrap)	12.37 $[12.04, 12.76]$	6.49 $[6.29, 6.72]$
Boundary ratio	1.90 $[1.83, 1.99]$	
Implied N for $\hat{\gamma}$=1.25	$\approx 9.8 \times 10^5$	$\approx 2.7 \times 10^5$
Null-grid powers ($\gamma \in \{0.5, 1, 2\}$)	0.046–0.086 ($\approx \alpha$; first run)	

Table 8: Historical exploratory record of combined per-cell, small-budget, and control-arm results for the four-strategy comparison (E3). Δ is cond-opt minus uniform power under null calibration unless noted. Mid-band cells enter the reported diagnostic median; control arms and small-budget cells are mechanism/regime checks.

Cell	Uniform	Cond-opt	Δ / note
Mid-band per-cell Δ at σ_J=0 (diagnostic median 0.515)			
$(N, \gamma) = (2000, 1)$	—	—	0.515
$(5000, 0.5)$	—	—	0.6115
$(10000, 0.5)$	—	—	0.465
Noise layer medians (bootstrap 95% CI)			
σ_J=0 (n=3)	—	—	0.515 [0.4925, 0.537]
σ_J=0.1 (n=5)	—	—	0.615 [0.5935, 0.6315]
σ_J=0.25 (n=4)	—	—	0.6645 [0.65, 0.678]
σ_J=0.1 per-cell range	—	—	+0.242 to +0.691
Illustrative noise path at $(N, \gamma) = (2000, 0.5)$			
$\sigma_J: 0 \to 0.1$ (uniform)	$0.1625 \to 0.134$	—	uniform absorbs FPR penalty
$\sigma_J: 0.1 \to 0.25$ (cond-opt)	—	$0.6515 \to 0.4585$	slight fall; ceiling elsewhere
Control arms (`RUNN2CONTROLARMS2026081001`); ex. $(N, \gamma) = (2000, 1)$			
max-$\lvert z \rvert$ power	0.48 (unif)	0.99 (KKT)	max-$\lvert z \rvert$ =0.70
conflict-only power	0.48 (unif)	0.99 (KKT)	conflict-only =0.34
Small-budget (`RUNN2SMALLBUDGETFIXED2026081001`); $N \in \{200, 500, 1000\}$			
median Δ at σ_J=0	—	—	0.4635 [0.436, 0.4905]
median Δ at σ_J=0.1	—	—	0.563 [0.5425, 0.584]
every band cell	—	—	$\Delta \geq 0.21$ (all 5 cells each)
Fit vs. detection diagnostic (σ_J=0)			
DRTS $\mathbb{E}[vr^2]$	pop. 7.7×10^{-4}	DRTS 4.3×10^{-5}	$\approx 17\times$ lower
APO separations (N=2000)	—	—	174/2000

E Artifact Traceability and Historical Comparisons

Run reports are under `docs/research/runs/`; values below map reported quantities to their corresponding run reports for full provenance. Rows marked historical catalog earlier exploratory runs; I0 E2 provides the analytic boundary replay benchmark, and I3 provides the verified protocol-complete E3/E5 benchmark suite. Plain strings `RUN-*` below enable grep audit.

Table 14: Recorded values and artifact sources. Δ = cond-opt − uniform median power (bootstrap 95% CI); E3 FPR uses per-group null calibration; E5 FPR checks an asymptotic LR rule; top-k shares from $I = vr^2$.

Claim / status	**Recorded value**	**Source run and field**
Top-0.4% / Top-1% I share (E1)	86.7% / 92.6% (2%: 96.2%)	`RUN-N1-INFO-EFFICIENCY-20260808-01 t4_info_efficiency/info_efficiency_report.json real.measures.topk_info_share [0.004]=0.8667, [0.01]=0.92594`
Gate-row $\rho = \mathrm{corr}(\lvert z \rvert, v)$	-0.2865 (signal -0.1495, $\kappa_\rho = 0.8646$)	same `t4_info_efficiency/info_efficiency_report.json mismatch_gate_rows.rho=-0.286476`
Spearman($\lvert z \rvert$, disagr.)	$+0.1153$ ($p \approx 2.8\text{e}{-65}$; mean 0.739/0.547)	same `disagreement_coupling .spearman=0.115322`

Continued on the next page

Table 14 continued: artifact traceability

Claim / status	Recorded value	Source run and field
Analytic $\|\gamma\|_{\min}$ surface (I0 scoped replay benchmark)	median rel. err. 1.97% (max 6.63%), 26/26 directions; funnel $324 \to 171 \to 35 \to 35$	`RUN-I0-E2-EVIDENCE-RECOVERY-20260908-01` `compare_only/power_map_report.json` + `e2_gate_audit.json` current I_{eff} consumer; $I_{\text{conf_only}}$ oracle failure retained
Tiny-positive-v control; exact v=0 limit	power ≤ 0.0145 (0.0145/0.0125/0.013), empirical $I = 4.44\text{e}{-7} > 0$, δ-bound 7.6e−7; exact-limit $I = 0$ is analytic	`RUN-N1-INFO-EFFICIENCY-20260808-01` `t5_unrecoverable/unrecoverable_report.json` `powers_by_gamma/analytic.I_empirical=4.44e-07`; exact v=0 follows from C4
Efficacy map $\gamma = 1@N \approx 10k$	22% (σ_J=0) / 6.8% (σ_J=0.1)	same `t3_power_map` (synthetic power at $N = 10k, \gamma = 1$)
Four-strategy Δ (historical exploratory benchmark; superseded by I3)	0.515 $[0.4925, 0.537]$ at 0; 0.615 $[0.5935, 0.6315]$ at 0.1; 0.6645 $[0.65, 0.678]$ at 0.25; FPR 0.0500 on 12/12 at 0.1	`RUN-N2-NOISE-ROBUST-20260810-01` `noise_robust_report.json` current fields `sigma_levels.*.bootstrap.observed_median_delta`; historical only, superseded for current status by I3 audit
DRTS information by budget	4.34/4.41/4.71e−5 at $N = 2000/5000/10000$ vs. population 7.68e−4 (16–18× lower)	`RUN-N2-CONSTRAINED-OPTIMAL-20260808-01/synthetic` `three_arm_report.json` `drts_info_diagnostic`
Real reweight (historical exploratory benchmark; superseded by I3)	$\Delta = 0.343$ $[0.3185, 0.371]$ (0.343/0.437/0.317 at 8/12/16); boundary 12.4→6.5 (−47%, bootstrap 12.37 vs 6.49)	`RUN-N2-CONSTRAINED-OPTIMAL-20260808-01/real` `real_track_power.jsonl` + `real_reweight_report.json`; historical only, superseded for current status by I3 paired-CI audit
I3 E3 protocol-repair benchmark (split-null validated)	144/144 cells; 36/36 validation FPR groups ≤ 0.07; paired-index median $\Delta = 0.581$ $[0.56325, 0.59450625]$ on 12 gated cells	`RUN-I3-E3-E5-PROTOCOL-REPAIR-20260908-01/remote201` `e3_formal/` + `e3_gate_audit.json`; current protocol-complete local observation
I3 E5 protocol-repair benchmark (paired bootstrap)	powers Uniform/Cond-Opt 0.2415/0.4515/0.6690 vs 0.4805/0.7840/0.9560; boundary $12.89195 \to 8.25700$; paired $\Delta = 0.287$ $[0.2675, 0.3070]$	`RUN-I3-E3-E5-PROTOCOL-REPAIR-20260908-01/remote201` `e5_paired/` + `e5_gate_audit.json`; simulated-label, protocol-complete
Six-arm baseline contrast (frozen-pilot comparison)	36/36 validation groups pass FPR ≤ 0.07 (max 0.066); at $\gamma = 0.5$, Cond-Opt–Uniform $\Delta = 0.388$–0.645, Fisher–Uniform $\Delta = 0.274$–0.508; all denominators 2000	`RUN-BASELINE-CONTRAST-FORMAL-20260921-01` `sigma0/baseline_contrast.json`, `sigma01/baseline_contrast.json`, and `formal_audit.json`; six-arm frozen-pilot comparison
Finite-sample identifiability and projection audit	Review six-row counterexample: frozen 1/81, re-projected $< 10^{-32}$; current HelpSteer2+3 retention 0.9868/0.9983/0.99995 at $K = 2k/5k/10k$; synthetic repair 0.9816/0.9697/0.9242	`RUN-SELECTION-SAFETY-20260921-01`, `RUN-SELECTION-REPAIR-SMOKE-20260921-01`; label-free rank/information audit, no power claim
Historical scaling at $\hat{\gamma} = 1.25$	$N \propto \gamma^{-2}$: ≈ 9.8e5 rows (uniform) / ≈ 2.7e5 (reweighted) vs. the 10^4-row budget (≈ 98× / ≈ 27×)	`real_track_power.jsonl` boundary interpolation + design scaling

Continued on the next page

Table 14 continued: artifact traceability

Claim / status	Recorded value	Source run and field
Pilot-information uncertainty (E5)	$I(\hat{q}_0) = 4.11\text{e}{-6}$; $\mathbb{E}[I(\hat{q})] = 4.11\text{e}{-6}$ $[2.25, 6.47]\text{e}{-6}$ (cluster bootstrap B=2000, full-pipeline re-estimation); uniform ratio 2.17 $[2.16, 2.18]$; median top-10^4 overlap 0.591 (min 0.447)	`RUN-N7-PILOT-BOOTSTRAP-20260825-01` `pilot_bootstrap.json` + `bootstrap_draws.jsonl`
Historical real-track FPR	0.053 (uniform) / 0.059 (cond-opt) ≤ 0.07	same `real_track_power.jsonl`
Real-label pilot (E6)	2,959 valid / 41 separation; real-label rejection 0.6–4.6%; auxiliary row-permutation diagnostic 4.5–6.4%; row-level C6 proxy $13.8/13.7/13.7$ (Cond-Opt) vs. $37.3/21.5/14.4$ (Uniform) at $N = 2k/5k/10k$; non-detection at measured scale	`RUN-P0-REAL-LABEL-PILOT-20260827-01` `report.json`; all evidence `local_hypothesis`
Held-out detectability screen (E7)	14 candidates on scan half (10,935 rows); selected product τ=1/4 margin 1.56; held-out confirm $\hat{\gamma}_B = 12.71$, se $= 11.48$, $z = 1.11$; selected candidate not confirmed	`RUN-I5-DETECTABILITY-SCAN-20260914-01` `scan_results.jsonl` + `detectability_scan.json`; exploratory diagnostic screen
Weighted-projection implementation correction (E7)	Replaced the v^2 least-squares solve by the intended v-weighted solve using $\sqrt{v}$; 14/14 candidate order and selected candidate unchanged; selected information changed by 0.0015%–0.2313%	`RUN-I5-DETECTABILITY-SCAN-20260921-PROJECTION-FIX-01` `execution_note.md`; code/test correction, not a new outcome
Screen positive control (semi-synthetic)	injected $\gamma \in \{5.9, 23.7\}$ ($0.5\times$ / $2\times$ boundary) on real geometry; below: max margin 0.64, held-out $z = 1.36$ no reject; above: selected sibling feature confirmed $\hat{\gamma}_B = 22.2$, $z = 3.08$	`RUN-I5-POSITIVE-CONTROL-20260914-01` `positive_control.json` + `receipt.json`; semi-synthetic positive control
Structured-noise calibration (S2)	empirical-null LR FPR $[0.039, 0.069]$, 16/16 cells ≤ 0.07; asymptotic/Wald up to 0.63; separate cluster-correlated and heteroscedastic noise models, $\sigma_J \in \{0.1, 0.25\}$, $N \in \{2k, 10k\}$	`RUN-I5-NONADDITIVE-NOISE-20260914-01` `nonadditive_calibration.json` + `receipt.json`; calibration evidence
Noise-robust small-budget	$\Delta = 0.4635$ $[0.436, 0.4905]$ at 0; 0.563 $[0.5425, 0.584]$ at 0.1 ($N \in \{200, 500, 1000\}$, 5 cells each)	`RUN-N2-SMALL-BUDGET-FIXED-20260810-01` `noise_robust_report.json` `sigma_levels.0.0/0.1.bootstrap`
Second corpus (UltraFeedback)	top-1% 0.8975 (89.8%), $\rho = -0.4741$, Spearman 0.1255 ($p = 5.32\text{e}{-12}$); three-expert-only $\rho = -0.4996$	`RUN-N1-SECOND-CORPUS-20260810-01` `report.md` + `e1_report.json` `verdict.spearman=0.12546,rho_gate=-0.47406`
Old research decision	Δ(B-FUSION)=+0.035 $p = 0.024$; Δ(B-WFUSION)=−0.001 $p = 0.747$	`UNVERIFIED`: prior S3 run identifier and field are not recorded in this table
RewardBench	three judges 2985/2985, Group A 0.859 / Group B 0.797 (secondary; A=GPT-4/human, B=rule/adversarial/verifiable)	`RUN-REWARDBENCH-COMPLETION-20260808-01`; file and field not specified here

Table 9: Historical exploratory record of full four-strategy per-cell null-calibrated power (E3; run `RUN-N2-NOISE-ROBUST-20260810-01`, R=2000). U/C/D/A = uniform / Cond-Opt / DRTS / APO. Mid-band cells (reference uniform power in $[0.2, 0.8]$) are marked $*$; they enter the reported diagnostic median. The table makes the per-cell composition of the mid-band Δ visible: Cond-Opt is at or near the ceiling in every $*$ cell at every σ_J, so the mid-band gain is $(1 -$ comparator power$)$ minus a term ≤ 0.074; at σ_J=0, 3/9 all cells have $\Delta \leq 0.038$.

Cell (N, γ)	U	C	D	A
σ_J=0				
$(2000, 0.5)$	0.163	0.602	0.166	0.103
$(2000, 1)$ $*$	0.478	0.993	0.433	0.269
$(2000, 2)$	0.962	1.000	0.949	0.731
$(5000, 0.5)$ $*$	0.321	0.932	0.294	0.215
$(5000, 1)$	0.842	1.000	0.802	0.658
$(5000, 2)$	1.000	1.000	1.000	0.998
$(10000, 0.5)$ $*$	0.528	0.993	0.554	0.449
$(10000, 1)$	0.988	1.000	0.988	0.942
$(10000, 2)$	1.000	1.000	1.000	1.000
σ_J=0.1				
$(2000, 0.5)$	0.134	0.651	0.118	0.092
$(2000, 1)$ $*$	0.300	0.991	0.271	0.204
$(2000, 2)$ $*$	0.758	1.000	0.742	0.585
$(5000, 0.5)$ $*$	0.297	0.926	0.223	0.163
$(5000, 1)$ $*$	0.691	1.000	0.576	0.443
$(5000, 2)$	0.993	1.000	0.987	0.945
$(10000, 0.5)$ $*$	0.367	0.983	0.363	0.330
$(10000, 1)$	0.902	1.000	0.853	0.810
$(10000, 2)$	1.000	1.000	1.000	1.000
σ_J=0.25				
$(2000, 0.5)$	0.079	0.459	0.085	0.063
$(2000, 1)$	0.104	0.931	0.116	0.101
$(2000, 2)$ $*$	0.234	1.000	0.264	0.157
$(5000, 0.5)$	0.085	0.633	0.089	0.080
$(5000, 1)$	0.175	0.996	0.179	0.140
$(5000, 2)$ $*$	0.403	1.000	0.460	0.327
$(10000, 0.5)$	0.135	0.789	0.161	0.126
$(10000, 1)$ $*$	0.269	1.000	0.289	0.247
$(10000, 2)$ $*$	0.656	1.000	0.691	0.577

Table 10: Estimand comparison with the closest fit-layer selection neighbors.

Work	Selection criterion	Estimand optimized
ActiveUltraFeedback / DRTS (Melikidze et al., 2026)	max predicted quality gap	downstream fit / win rate
Less is More (Deng et al., 2025)	margin maximization + Bayesian aggregation	DPO fit (AlpacaEval2)
Reward margins (M_{AP}) (Huang et al., 2025)	implicit–explicit margin gap	DPO fit
Active Reward Modeling (Shen et al., 2025)	Fisher info (final RM layer)	reward-model fit
ActiveDPO (Lin et al., 2026)	theory-grounded criterion (nonlinear RM)	DPO fit
ResponseRank (Kaufmann et al., 2025)	strength-ranked relative signals	cardinal utility learning
CPRD (Pukdee et al., 2026)	margin, connectivity of comparison graph	*identification*: conditional preference distribution
CompatRewardID (Drago et al., 2026)	query basis / index complexity	*identification*: compatible reward (RL)
DMLRank (Frauen et al., 2026)	debiased-ML collection policies	*estimation*: ranking scores + UQ
This work	v-weighted residual Fisher info $I_{\gamma\|w}$	***detection*: calibrated test of a specified effect**

Table 11: E6 real-label pilot (leak-free cluster split, R=500, real judge labels, frozen nested-LR instrument). $\hat{\gamma}$ median / CI-width median / MSE over valid replicates vs. the full-pool reference 1.254; $|\gamma|_{\min}$ is a realized row-level C6 proxy at 50% power, not a cluster-effective boundary; reject rates are asymptotic χ^2 on real labels versus the auxiliary row-wise y-permutation diagnostic (B=100, 25 replicates), not a formal Type-I calibration.

N	Arm	$\hat{\gamma}$	CI width	MSE	$\|\gamma\|_{\min}$	Rej. (asym./perm.)
2000	Cond-Opt	1.405	26.6	28.9	13.8	0.6% / 5.2%
2000	Uniform	−0.158	77.7	409.2	37.3	3.5% / 4.5%
5000	Cond-Opt	1.245	27.6	31.9	13.7	1.0% / 5.5%
5000	Uniform	0.720	44.5	142.6	21.5	4.6% / 6.4%
10000	Cond-Opt	1.411	28.7	35.0	13.7	1.6% / 6.1%
10000	Uniform	1.566	30.4	40.1	14.4	1.6% / 6.0%

Table 12: Historical reward-subset diagnostic on frozen HelpSteer2+3 (common holdout; exploratory). Fit columns use the frozen-raw recomputation described above; their confidence intervals are row-level, not cluster-robust. APO at N=2000 has all-zero margins and rank failure (n.i. = not identifiable), and is excluded from the fit comparison.

Strategy	N	Acc.	Acc. 95% CI	AUC	$\hat{\gamma}$	CI width	MSE
Uniform	2k	0.647	[0.633, 0.661]	0.725	−11.4	54.100	314.827
Uniform	5k	0.649	[0.634, 0.663]	0.725	4.6	34.700	91.440
Cond-Opt	2k	0.649	[0.634, 0.663]	0.725	2.9	28.099	39.878
Cond-Opt	5k	0.649	[0.634, 0.663]	0.724	3.5	18.099	17.849
DRTS	2k	0.646	[0.633, 0.661]	0.720	8.2	57.201	394.549
DRTS	5k	0.646	[0.633, 0.661]	0.721	7.9	36.099	190.081
APO	2k	0.489	[0.474, 0.504]	0.500	n.i.	—	—
APO	5k	0.649	[0.635, 0.663]	0.726	14.9	42.701	241.875