Top-k Identification with Correlated Biased LLM Judges via Anchor Leverage

Sixiong Xie Zhuofan Shi Haiyang Shen
Yun Ma Xiang Jing*
Peking University
National Key Laboratory of Data Space Technology and System
Beijing, China
xsx1001@stu.pku.edu.cn, jingxiang@pku.edu.cn

Abstract

Cheap LLM judges can agree on an incorrect ranking when they share systematic bias. We study fixed-confidence top-k identification with correlated, biased judge scores and costly trusted anchors. Under a Gaussian factor-bias model with known features and covariance, identifiability depends on whether anchors cover the feature contrasts across the top-k boundary. Profiling out unknown judge weights yields an instance-dependent cost characterization with a direct-information term and an anchor-leverage term. Profile Track-and-Stop attains the corresponding oracle asymptotic cost under the stated conditions. On a frozen 12-model, 10-judge instance with 23,545 real API scores and simulated anchors, the reported adaptive allocation reaches 95% boundary accuracy with approximately $2.5\times$ less budget than the Uniform and DS-Active adaptations. Its advantage over human-heavy allocations is smaller. Additional Chatbot Arena and MT-Bench instances use disjoint human vote splits as the anchor channel; the larger Arena instance reaches the same threshold with approximately $2.9\times$ less budget than uniform human labeling, with the latter budget extrapolated beyond the finite vote pool. Bias-test power and calibration limitations bound the scope of this evidence.

1 Introduction

LLM judges make model evaluation affordable by scoring candidate responses at scale. Their scores, however, can share a systematic preference for properties such as verbosity or style. Repeating these evaluations reduces noise but need not correct the resulting ranking. This distinction matters when an evaluation must identify a small set of models for a leaderboard, a deployment shortlist, or further testing, while expensive human annotations remain limited.

Consider two models near the top-k cutoff, one more verbose than the other. If all judges favor verbosity, their agreement cannot distinguish genuine quality from their common preference. An unbiased anchor can resolve this ambiguity only when its features cover the direction that separates the two candidates. The relevant question is therefore:

How should judge queries and trusted anchors be allocated to identify the top-k with confidence $1-\delta$ at minimum expected cost?

Classical best-arm identification answers this question with unbiased observations [14, 9]. Multi-fidelity methods allow cheap proxies under restrictions on their bias [12, 16]. Here, the nuisance bias is unknown, shared through arm features, and accompanied by correlated judge noise. We study this

*Corresponding author: jingxiang@pku.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

setting under a Gaussian factor-bias model with known features and covariance; practical estimation of these quantities requires separate calibration assumptions.

Main results. First, we characterize the ambiguity left by judge-only observations and prove that identifying the top-k requires anchors to cover the cross-boundary feature contrasts, rather than necessarily the entire feature space (Section 4). Second, profiling out the unknown judge weights yields an instance-dependent cost bound with a direct-information term and an anchor-leverage term (Section 5). Third, Profile Track-and-Stop combines estimation, forced exploration, and cost-aware tracking to attain the oracle asymptotic cost under the stated regularity conditions (Section 6). The allocation geometry provides interpretable calibration and anchor-design rules (Section 7).

Empirical evidence. We separate observable recovery from analytical cost predictions. On a frozen 12-model, 10-judge instance built from $23{,}545$ real API scores, with simulated anchors, the reported adaptive allocation reaches 95% boundary-pair accuracy at budget $2{,}860$, compared with approximately $7{,}000$ for the Uniform and DS-Active adaptations. Its advantage over human-heavy allocations at this threshold is smaller and has overlapping intervals. An additional Chatbot Arena instance uses real human vote blocks as anchors ($N = 20$, $k = 8$), and a disjoint calibration/target/anchor split tests differential anchor bias. These tests do not establish universal human unbiasedness. Section 8 reports all three recovery metrics, the hardest boundary contrasts, and the limits of the bias tests.

Roadmap. Judge scores leave a gauge ambiguity; anchors remove it along their feature directions; top-k certification needs the boundary contrasts covered; the profiled cost prices this coverage; the algorithm tracks an efficient allocation. A small worked example and a known-versus-unknown parameter table make these steps explicit before the general cost characterization.

2 Related Work

Best-arm and top-k identification. Fixed-confidence best-arm identification has matching information-theoretic bounds and Track-and-Stop algorithms [9, 14], with top-k extensions [11, 2] and pure-exploration games [5, 6]. Our lower-bound argument follows this framework, with correlated observations, heterogeneous costs, and a nuisance-profiled KL divergence.

Multi-fidelity and correlated observations. Multi-fidelity bandits exploit cheap biased sources under prescribed bias conditions [12, 16]; correlated-arm models exploit dependence across observations [10]. Our factor-bias model addresses an additional ambiguity: unknown shared judge preferences can change the ranking even with arbitrarily many proxy observations.

LLM-as-a-judge evaluation. MT-Bench, AlpacaEval, Chatbot Arena, and PandaLM [19, 7, 3, 18] study empirical judge reliability and calibration. Crowdsourcing models also aggregate observations from imperfect raters [4]. Our contribution is complementary: under an explicit statistical model, we characterize which anchors are necessary and the evaluation cost of certifying a top-k set. Prediction-powered inference [1] motivates one baseline adaptation, but its estimation objective and assumptions differ from fixed-confidence top-k identification.

Experimental design. Classical optimal design [8, 17] relates informative measurements to estimable contrasts. Here, the target contrasts are determined by the top-k boundary. This explains why a large-norm anchor can be less useful than an anchor that covers the relevant feature direction.

3 Problem Setup

Model and notation. There are N candidate arms with quality vector $\boldsymbol{\mu} \in \mathbb{R}^N$, M judges, and a known feature matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N]^\top \in \mathbb{R}^{N \times r}$. The unknown judge loadings are $\boldsymbol{W} \in \mathbb{R}^{M \times r}$. The instance is $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{W})$; $\boldsymbol{\Phi}$, the positive-definite covariance $\boldsymbol{\Sigma}$, and the human noise variance $\sigma_h^2 > 0$ are fixed model inputs. A judge-vector query J_i and a human query h_i return, respectively,

$$\boldsymbol{X}_{i,\cdot} = \mu_i \mathbf{1} + \boldsymbol{W}\boldsymbol{\phi}_i + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1}$$

and $Y_i = \mu_i + \eta_i$, $\eta_i \sim \mathcal{N}(0, \sigma_h^2)$. Observations are independent across queries and channels. The costs are $c_{J_i} = c_J > 0$ and $c_{h_i} = c_h \geq c_J$.

Table 1: Parameters and allocation quantities in the oracle model.

Quantity	Meaning	Status
$\boldsymbol{\mu}, \boldsymbol{W}$	Quality and judge bias loadings	Unknown
$\boldsymbol{\Phi}, \boldsymbol{\Sigma}, \sigma_h^2$	Features and channel noise	Known in theorems
c_J, c_h	Query prices	Known
τ	Common-shift judge information	Computed from $\boldsymbol{\Sigma}$
$\boldsymbol{\omega}, G, \boldsymbol{b}_{pq}$	Allocation and profiled information	Design-dependent
$T^*(\boldsymbol{\theta})$	Optimal leading certification cost	Instance-dependent

Write $\mathcal{A} = \{J_1, \ldots, J_N, h_1, \ldots, h_N\}$ and $K = |\mathcal{A}| = 2N$. Here $t = \sum_a N_a(t)$ is the total number of action queries, including the pilot. For action counts $N_a(t)$, define

$$C(t) = \sum_{a \in \mathcal{A}} c_a N_a(t), \qquad \Omega_c = \left\{ \boldsymbol{\omega} \geq 0 : \sum_a c_a \omega_a = 1 \right\}. \tag{2}$$

An allocation ω_a is a number of queries per unit cost, and $\hat{\boldsymbol{\omega}}(t) = N(t)/C(t) \in \Omega_c$. Its anchor set is $A(\boldsymbol{\omega}) = \{i : \omega_{h_i} > 0\}$.

The scalar $\tau = \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}$ is the Fisher information in one judge-vector query about a common shift of its mean. Equivalently,

$$\frac{1}{\tau} = \min_{\boldsymbol{v}^\top \mathbf{1} = 1} \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}, \qquad \boldsymbol{v}^* = \boldsymbol{\Sigma}^{-1} \mathbf{1} / \tau.$$

Thus $1/\tau$ is the noise variance of the best linear unbiased aggregate of the judges. For $\boldsymbol{\Sigma} = I_M$, $\tau = M$; for a unit-variance equicorrelated covariance, $\tau = M/[1 + (M-1)\rho]$. Positive correlation reduces the information gain from adding judges. This quantity concerns noise after accounting for bias; it does not remove a shared bias in the mean.

Goal. Let $S_k^*(\boldsymbol{\mu})$ be the indices of the k largest qualities and

$$\mathrm{bdry}(\boldsymbol{\mu}) = \{(p, q) : p \in S_k^*(\boldsymbol{\mu}),\ q \notin S_k^*(\boldsymbol{\mu})\}, \qquad \Delta_{pq} = \mu_p - \mu_q.$$

A δ-PAC procedure stops at τ_δ, returns a size-k set $\widehat{S}$, and satisfies $\mathbb{P}_{\boldsymbol{\theta}}(\widehat{S} \neq S_k^*(\boldsymbol{\mu})) \leq \delta$. Our objective is the expected total cost $\mathbb{E}_{\boldsymbol{\theta}}[C(\tau_\delta)]$. The alternative set $\mathrm{Alt}(\boldsymbol{\theta})$ varies only $(\boldsymbol{\mu}, \boldsymbol{W})$ while holding $(\boldsymbol{\Phi}, \boldsymbol{\Sigma}, \sigma_h^2)$ fixed.

Assumption 1. $\boldsymbol{\Phi}$ has full column rank, $\boldsymbol{\Sigma} \succ 0$ is known, and there are no ties at the top-k boundary. For the oracle tracking theorem we also assume that the optimal allocation is unique.

Roadmap. The analysis separates three questions. A gauge transformation shows which quality differences judge scores cannot identify. Anchors remove the gauge directions that matter for the top-k boundary. Profiling the remaining nuisance parameters then gives a convex allocation problem whose value is the optimal leading cost. The upper bound concerns this oracle Gaussian model; the experiments separately examine practical allocations with fitted features and covariance.

4 Identifiability: Gauge and Boundary Coverage

For any $\boldsymbol{v} \in \mathbb{R}^r$, the transformation

$$(\boldsymbol{\mu}, \boldsymbol{W}) \longmapsto (\boldsymbol{\mu} + \boldsymbol{\Phi} \boldsymbol{v},\ \boldsymbol{W} - \mathbf{1} \boldsymbol{v}^\top) \tag{3}$$

leaves every judge distribution unchanged. Human observations on a set A also remain unchanged exactly when $\boldsymbol{\Phi}_A \boldsymbol{v} = 0$. The remaining gauge dimension is $r^*(A) = r - \mathrm{rank}(\boldsymbol{\Phi}_A)$.

Theorem 2 (Full identifiability and uniform impossibility)**.** *With all judge channels available, the full quality vector is identifiable from the judge laws and the human laws on A if and only if $\mathrm{rank}(\boldsymbol{\Phi}_A) = r$. If a top-$k$ boundary contrast is not orthogonal to $\ker(\boldsymbol{\Phi}_A)$, two gauge-equivalent instances can have different top-k sets. No procedure using only those channels can be uniformly δ-PAC over a class containing both instances, for $\delta < 1/2$.*

Theorem 3 (Sort-identifiability)**.** *For $(p, q) \in \mathrm{bdry}(\boldsymbol{\mu})$, the sign of $\mu_p - \mu_q$ is identifiable from these laws if and only if $\boldsymbol{\phi}_p - \boldsymbol{\phi}_q \in \mathrm{span}\{\boldsymbol{\phi}_i : i \in A\}$. Top-$k$ is identifiable exactly when this holds for every boundary pair.*

Example 4 (A three-arm verbosity gauge)**.** Let $N = 3$, $k = 1$, $r = 1$, $\boldsymbol{\Phi} = (1, 0, -1)^\top$, $\boldsymbol{\mu} = (0.2, 0.1, 0)^\top$, and let every judge have loading 0.3. Judge means are $(0.5, 0.1, -0.3)$. The gauge shift $v = -0.3$ gives quality $(-0.1, 0.1, 0.3)$ and loading 0.6, with exactly the same judge means, but changes the best arm from 1 to 3. An anchor on arm 2 has feature zero and cannot remove this ambiguity. An unbiased anchor on arm 1 or 3 covers the scalar feature direction.

Repeated judge queries estimate the biased sums precisely, but do not separate quality from the shared preference. If a residual gauge only adds the same constant to every quality, it changes the quality vector without changing its ranking.

Boundary coverage can require less than full quality identifiability. If boundary differences span a d-dimensional subspace, the anchor span must contain that subspace. Whether d anchors suffice depends on the available candidate features; r linearly independent anchor features always suffice.

5 Matching Cost Complexity

5.1 The profiled anchor leverage

Set $A_i(\boldsymbol{\omega}) = \tau\omega_{J_i} + \omega_{h_i}/\sigma_h^2$ and define

$$G(\boldsymbol{\omega}) = \frac{1}{\tau}\sum_{i=1}^{N} \frac{\omega_{J_i}\omega_{h_i}/\sigma_h^2}{\tau\omega_{J_i} + \omega_{h_i}/\sigma_h^2}\,\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top, \tag{4}$$

where a summand with both weights zero is zero. Each coefficient combines judge and human information at an anchor through a harmonic mean. For $A_p, A_q > 0$, let

$$\boldsymbol{b}_{pq}(\boldsymbol{\omega}) = \frac{\omega_{J_p}}{A_p}\boldsymbol{\phi}_p - \frac{\omega_{J_q}}{A_q}\boldsymbol{\phi}_q, \qquad \kappa_{pq}(\boldsymbol{\omega}) = \boldsymbol{b}_{pq}^\top G(\boldsymbol{\omega})^\dagger \boldsymbol{b}_{pq}. \tag{5}$$

The leverage is assigned value $+\infty$ when $\boldsymbol{b}_{pq} \notin \mathrm{range}(G)$. This extended value is essential: a Moore–Penrose quadratic alone does not detect a contrast outside the information range.

5.2 Instance-dependent rate

Theorem 5 (Profiled matching cost bounds)**.** *Under Assumption 1, every δ-PAC procedure satisfies*

$$\mathbb{E}_{\boldsymbol{\theta}}[C(\tau_\delta)] \geq T^*(\boldsymbol{\theta}) \log \frac{1}{2.4\delta}, \qquad T^*(\boldsymbol{\theta}) = \min_{\boldsymbol{\omega}\in\Omega_c} \max_{(p,q)\in\mathrm{bdry}(\boldsymbol{\mu})} T_{pq}^*(\boldsymbol{\omega}), \tag{6}$$

where

$$T_{pq}^*(\boldsymbol{\omega}) = \frac{2}{\Delta_{pq}^2}\left[\underbrace{\frac{1}{A_p} + \frac{1}{A_q}}_{\text{direct boundary information}} + \underbrace{\kappa_{pq}(\boldsymbol{\omega})}_{\text{profiled anchor leverage}}\right]. \tag{7}$$

If $A_pA_q = 0$, the pair cost is $+\infty$. Algorithm 1 asymptotically matches this lower bound on regular instances.

If neither endpoint receives human queries, then $\boldsymbol{b}_{pq} = (\boldsymbol{\phi}_p - \boldsymbol{\phi}_q)/\tau$ and the bracket in (7) becomes

$$\frac{1}{\tau}\left(\frac{1}{\omega_{J_p}} + \frac{1}{\omega_{J_q}}\right) + \frac{(\boldsymbol{\phi}_p - \boldsymbol{\phi}_q)^\top G^\dagger(\boldsymbol{\phi}_p - \boldsymbol{\phi}_q)}{\tau^2}.$$

This special case separates correlated-judge noise from the price of removing shared bias. Formula (7) also includes direct human information when a boundary arm is anchored. Costs enter through Ω_c; they do not change the per-query likelihood.

Algorithm 1 Oracle Profile Track-and-Stop

Require: δ, $\mathbf{\Phi}$, $\mathbf{\Sigma}$, σ_h^2, c_J, c_h

Choose $P_0 \subseteq [N]$ with $|P_0| = r$ and $\operatorname{rank}(\mathbf{\Phi}_{P_0}) = r$
Query each judge action once and each human action in P_0 once
Set $t \leftarrow N + r$
repeat
  Compute the joint Gaussian MLE $\hat{\boldsymbol{\theta}}_t = (\hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{W}}_t)$
  Compute $\boldsymbol{\omega}_t \in \arg\min_{\boldsymbol{\omega} \in \Omega_c} \max_{(p,q) \in \mathrm{bdry}(\hat{\boldsymbol{\mu}}_t)} T^*_{pq}(\boldsymbol{\omega}; \hat{\boldsymbol{\theta}}_t)$
  if $\exists a \in \mathcal{A} : N_a(t) < \sqrt{C(t)/(Kc_a)}$ **then**
    $a_{t+1} \in \arg\min_a c_a N_a(t)$ ▷ Forced exploration
  else
    $a_{t+1} \in \arg\min_a \|(N(t) + \boldsymbol{e}_a)/(C(t) + c_a) - \boldsymbol{\omega}_t\|_\infty$ ▷ Cost-aware tracking
  end if
  Pull a_{t+1}, update counts and cost, and set $t \leftarrow t + 1$
  Recompute $\hat{\boldsymbol{\theta}}_t$ from the updated data
  $Z(t) \leftarrow \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\theta}}_t)} \sum_a N_a(t) \mathrm{KL}(\nu_a^{\hat{\boldsymbol{\theta}}_t} \| \nu_a^{\boldsymbol{\lambda}})$
until $Z(t) \geq \beta(t, \delta)$
return $S_k^*(\hat{\boldsymbol{\mu}}_t)$

5.3 Convexity across changes in support

Lemma 6. *The extended pair cost T^*_{pq} is convex and lower semicontinuous on all of Ω_c, including allocations at which the information rank changes. The minimum defining $T^*(\boldsymbol{\theta})$ is finite and attained.*

For the full Fisher information $\mathcal{I}(\boldsymbol{\omega})$ and the contrast vector $\widetilde{\boldsymbol{e}}_{pq}$ that selects $\mu_p - \mu_q$, the bracket in (7) equals the extended variance

$$\mathcal{V}(\mathcal{I}(\boldsymbol{\omega}), \widetilde{\boldsymbol{e}}_{pq}) = \sup_{\boldsymbol{y}} \{2\widetilde{\boldsymbol{e}}_{pq}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathcal{I}(\boldsymbol{\omega}) \boldsymbol{y}\}.$$

Information is affine in the allocation, so this supremum is convex and lower semicontinuous. It is finite precisely when the contrast is estimable. Appendix C proves this representation and attainment without a fixed-support restriction.

6 Algorithm: Profile Track-and-Stop

Theorem 7 (Upper bound on expected cost)**.** *Under Assumption 1, Algorithm 1, with the valid threshold specified in Appendix B.1, is δ-PAC and*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\theta}}[C(\tau_\delta)]}{\log(1/\delta)} \leq T^*(\boldsymbol{\theta}).$$

The fixed pilot makes the first joint MLE identifiable. Subsequent forced exploration supplies unbounded observations on every channel while spending only $O(\sqrt{C})$ cost. The tracking step compares the allocation after each possible next action with the plug-in oracle target. Appendix B expands the cost-indexed concentration and last-exit argument supplied in the rebuttal, which is needed to pass from almost-sure convergence to expected cost.

The oracle correctness statement uses known inputs and the conservative threshold (17), proved in Appendix B.1. The reported adaptive experiment truncates Algorithm 1 at a fixed budget with fitted inputs. Its measured recovery does not establish the oracle finite-δ guarantee. The stopping experiments use the reported practical stitched threshold. The convex solve considers at most $k(N-k)$ boundary pairs.

6.1 Estimated features and covariance

The factor coordinates are arbitrary: $\boldsymbol{\Phi} \mapsto \boldsymbol{\Phi} R$ and $\boldsymbol{W} \mapsto \boldsymbol{W} R^{-\top}$, for invertible R, leave the model unchanged. The information-based objective and the profiled leverage are invariant to this coordinate change.

Corollary 8 (Coordinate-invariant plug-in guarantee)**.** *Fix a calibration phase of K_{cal} joint judge-vector observations to estimate a rank-r feature subspace and residual covariance $\hat{\boldsymbol{\Sigma}}$. Assume a nonzero singular-value gap, bounded fourth moments, and $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$. The submitted calibration rates are*

$$\inf_{R \in \mathrm{GL}(r)} \|\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi} R\|_F = O_P(K_{\mathrm{cal}}^{-1/2}), \qquad \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} = O_P(K_{\mathrm{cal}}^{-1/2}).$$

On a calibration event where these errors are at most ϵ, the plug-in allocation inflates the oracle profiled cost by $1+O(\epsilon)$ on regular instances. For $K_{\mathrm{cal}} \to \infty$ and $K_{\mathrm{cal}} = o(\log(1/\delta))$, the submitted asymptotic statement is

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\theta}}[C(\tau_\delta) + K_{\mathrm{cal}} c_J]}{\log(1/\delta)} \le T^*(\boldsymbol{\theta}).$$

Appendix B.6 explains the coordinate invariance and conditional allocation perturbation underlying this plug-in statement.

Consistency of an SVD estimate requires an identifiable calibration signal and a singular-value gap. A low-rank fit alone cannot distinguish a shared quality component from a shared judge-bias component. A finite-δ guarantee with estimated model inputs requires a valid nuisance confidence set or another analysis of calibration uncertainty. Our oracle theorem treats these inputs as known; the frozen-data experiments evaluate fitted allocation rules rather than online calibration uncertainty.

7 Allocation Structure and Design Rules

7.1 Calibration-only budget split

Corollary 9 (Geometric-mean split in the calibration subproblem)**.** *Suppose an interior calibration arm affects the objective only through G, so that it is not an endpoint of an active target contrast. Its optimal split is*

$$\frac{\omega_{h_i}}{\omega_{J_i}} = \sigma_h \sqrt{\tau \frac{c_J}{c_h}}, \qquad \frac{c_h \omega_{h_i}}{c_J \omega_{J_i}} = \sigma_h \sqrt{\tau \frac{c_h}{c_J}}. \tag{8}$$

For an anchored endpoint of an active boundary pair, both the direct terms $1/A_p, 1/A_q$ and $\boldsymbol{b}_{pq}$ depend on its split. The full program determines the optimum there; (8) is a design heuristic rather than a universal formula for every anchored arm.

7.2 Anchor design for fixed target contrasts

After imposing the calibration split, a calibration-only anchor-design subproblem takes the form

$$\min_{\boldsymbol{\alpha} \ge 0,\ \sum_i \alpha_i = B} \max_{(p,q)} \boldsymbol{d}_{pq}^\top \left(\sum_i \alpha_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \right)^\dagger \boldsymbol{d}_{pq}, \tag{9}$$

with infinite value for an uncovered target contrast. Here $\boldsymbol{d}_{pq}$ is a fixed boundary direction and B is design mass, not an independently specified monetary budget. This finite-contrast design problem connects anchor placement to classical optimal design [8, 17]. It favors coverage and conditioning in the directions being tested. The general top-k allocation still uses (7), since its anchors can themselves be boundary endpoints.

7.3 Practical interpretation

The oracle allocation balances direct information on difficult boundary pairs against information that calibrates shared bias. Active pairs have equalized optimal difficulty, and the marginal decrease in their weighted cost is balanced against each action's price. The calibration split and contrast-aware anchor placement explain practical initialization rules; neither is a substitute for solving the general oracle program.

8 Experiments

We evaluate recovery directly, separately from the analytical allocation objective $T^*(\boldsymbol{\omega})$. The first instance uses $23{,}545$ frozen Arena-Hard-v2.0 judge scores ($N = 12, M = 10$, top-5) with benchmark-derived quality and *simulated* human anchors. Two additional instances use disjoint *real human votes* to define the target and supply anchors. All results below use estimated features and covariance; they are empirical checks of the practical pipeline, rather than finite-confidence guarantees for estimated nuisance parameters. Appendix D gives the protocols and additional diagnostics.

Comparators and metrics. Uniform distributes observations over both channels; Pure-human uses only anchors, and Human-only oracle uses an oracle human allocation. DS-Active uses judge reliability and broad anchor coverage; Estimate-and-correct anchors current top candidates. PPI-inspired and MF-BAI are adaptations of prediction-powered inference and multi-fidelity identification, respectively, rather than their original algorithms [1, 16]. The static methods use oracle allocations held fixed within a run; adaptive P-T&S updates plug-in estimates and is truncated at the same budget. We report exact top-k recovery, set overlap, and accuracy across the $k(N-k)$ target boundary pairs. A separate budget sweep measures cost to reach an average accuracy threshold; it is not a δ-PAC stopping cost.

Frozen judges with simulated anchors. At $c_h/c_J = 30$, budget $B = 5{,}000$, and 200 seeds, adaptive P-T&S attains 96.8% boundary accuracy and 47.5% exact top-5 recovery (Table 2). Its budget to 95% boundary accuracy is $2{,}860$, versus $7{,}030$ for Uniform and $7{,}050$ for DS-Active. Human-heavy methods are more competitive and their intervals overlap. Estimate-and-correct has the largest nominal exact recovery rate (53.5%), despite lower overlap and boundary accuracy; P-T&S therefore does not lead every metric (Fig. 1). Independent-proxy records no exact successes and 70.5% boundary accuracy at this budget, consistent with the anchor-free failure mode.

The hardest tercile contains the 12 smallest gaps among 35 boundary pairs (gaps 0.10–0.98). Adaptive P-T&S reaches 91.0% accuracy there (Table 3). The paired cost ratio for reaching 90% hardest-pair accuracy is Pure-human/adaptive 1.38 $[1.01, 1.90]$ and Human-only oracle/adaptive 1.30 $[0.81, 1.64]$. These results support a modest advantage on the difficult boundary pairs; the oracle comparison remains uncertain.

Table 2: Direct R1 recovery with simulated anchors ($B = 5{,}000$, 200 seeds). Percentages are point estimates; full 95% intervals and all methods appear in Appendix D.2. The last column is a separate threshold sweep.

Method	Exact top-5	Overlap	Boundary accuracy	$B_{95\%,\mathrm{all}}$
P-T&S adaptive	47.5	89.5	96.8	2,860
P-T&S static	42.5	87.2	96.5	3,270
Human-only oracle	34.5	86.1	95.9	3,130
Pure-human	38.0	87.4	96.2	3,420
Uniform	26.5	83.5	93.6	7,030
DS-Active	29.0	84.5	94.1	7,050
Estimate-and-correct	53.5	79.3	85.4	9,840
Independent-proxy	0.0^*	56.7	70.5	not reached

*Zero successes in 200 runs; this does not imply a zero population success probability.

Table 3: R1 hardest-tercile boundary pairs: accuracy at $B = 5{,}000$, separate budget to 90% accuracy, and paired budget ratio to adaptive P-T&S. Reported 95% intervals. Other methods' hardest-tercile results were not reported.

Method	Hardest accuracy (%)	$B_{90\%,\mathrm{hard}}$	Budget/adaptive
P-T&S adaptive	91.0 [89.5,92.5]	4,319 [3,300,5,284]	1 (reference)
P-T&S static	90.6 [88.8,92.1]	4,634 [4,002,5,614]	1.07 [0.85,1.43]
Human-only oracle	89.0 [87.2,90.4]	5,618 [3,279,6,218]	1.30 [0.81,1.64]
Pure-human	89.1 [87.5,90.7]	5,977 [4,727,7,551]	1.38 [1.01,1.90]

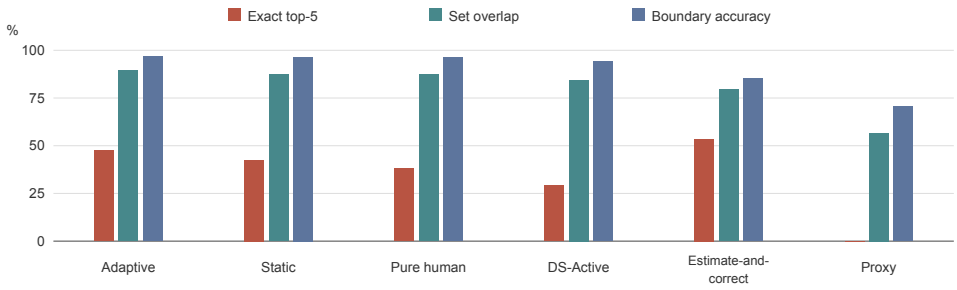

Figure 1: Three recovery metrics at the same fixed budget on the R1 instance. Adaptive and Static denote P-T&S; Proxy denotes Independent-proxy. Bars reproduce the reported point estimates; intervals appear in Table 5. Real judge scores are paired with simulated anchors. This comparison does not establish asymptotic infeasibility empirically.

Real human anchors and their scope. MT-Bench and Chatbot Arena votes are split at the prompt/session level into disjoint calibration, target, and anchor sets. Calibration fixes opponent strengths and normalization; the target defines Bradley–Terry–Elo quality. Each anchor observation fits one arm from a block of 20 votes, priced as $c_h = 30c_J$ for the *whole block*; judges score the original responses seen by voters. No vote serves two roles. Human distortion along the estimated judge-bias directions has normalized amplitude 0.12 on Arena and 0.13 on MT-Bench, with one-sided 95% upper bounds 0.24 and 0.30 and test p-values 0.57 and 0.52. These are non-rejections of differential bias, not proof of unbiasedness. At injected amplitude 0.35, rejection power is 86% on Arena and 60.5% on MT-Bench. Shared voter preferences enter the target by construction and cannot be detected by this split test. Appendix D.1 gives construction and power results; Appendix D.4 examines the effect of biased anchors.

Larger real-vote instance. On Chatbot Arena ($N = 20$, $M = 10$, top-8, $B = 20{,}000$, 200 seeds), adaptive P-T&S achieves 71.5% $[65.6, 77.0]$ exact recovery and 96.5% $[95.8, 97.2]$ boundary accuracy. Its budget to 95% boundary accuracy is $13{,}400$ $[12{,}300, 14{,}700]$, compared with DS-Active's $23{,}100$ $[21{,}000, 25{,}600]$. Independent-proxy reaches 0.5% exact recovery and 71.9% boundary accuracy. Full results and finite-vote-pool extrapolations appear in Appendix D.3. This larger roster supplies a second empirical instance; differing rosters and prompts prevent a controlled scaling claim. Analytical bootstrap uncertainty, heavy-tail sensitivity, and frozen-allocation audits are reported separately in Appendix D.5.

9 Discussion and Limitations

The theory distinguishes collecting more proxy scores from collecting information that resolves the ranking ambiguity. Anchor placement matters because only cross-boundary contrasts need certification. The experiments support this allocation mechanism on the reported instances; analytical certification-cost ratios and fixed-budget recovery improvements measure different quantities.

Oracle assumptions and empirical calibration. The exact cost characterization assumes known features and covariance, a stationary Gaussian location model, and unbiased anchors. The empirical pipeline estimates features by SVD and covariance from scores. A low-rank fit alone does not identify the nuisance subspace: a common-mode judge bias can be confounded with quality, and right-centering can remove that direction. Allocation stability under consistent, well-conditioned calibration does not supply a finite-δ certificate with estimated nuisance quantities. Such a certificate needs calibration confidence sets and corresponding threshold adjustments.

What the human-anchor test establishes. Disjoint vote splits test differential anchor distortion along the estimated judge-bias directions. Non-rejection is not proof of unbiasedness; the power curves and one-sided amplitude bounds must be read alongside the p-values. The MT-Bench test is underpowered near the tested degradation threshold. A preference shared by all voters enters $\boldsymbol{\mu}$ by construction and cannot be seen by this or any such test. Human vote blocks also provide

an empirical channel rather than exact Gaussian observations; reuse beyond the finite vote pool is explicitly marked as extrapolation.

Stationarity and asynchronous operation. The guarantees apply to a fixed roster within an evaluation run. Provider changes between runs require fresh calibration; undetected mid-run changes invalidate the stationary likelihood. A change aligned with the gauge can be invisible to judge-only monitoring. In practice, record model versions, monitor residuals, and restart on a fresh segment after an alarm. For asynchronous labels, issue anchor batches, update on returned observations, and exclude pending labels from the likelihood and stopping statistic. Correctness requires predictable requests and non-informative return delays. Limiting pending requests bounds unused paid labels, but does not by itself bound extra judge queries or wall-clock delay. Delayed optimality remains open.

Beyond Gaussian noise and broader impact. The gauge argument extends to identifiable additive location families, whereas the quadratic KL, closed-form leverage, and matching stopping analysis require additional distributional structure. The reported Student-t sweep is a robustness observation, not a uniform heavy-tail guarantee. Efficient anchoring can reduce annotation costs, but biased targets, missing feature directions, or version drift can produce misleading leaderboards. Report the target population, anchor provenance, uncertainty, and failures alongside cost savings. Larger controlled within-instance scaling and cross-fitted calibration diagnostics require further evidence.

Acknowledgment

The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China under Grant No. 62595734 for the research presented in this paper.

Disclosure. The authors declare no competing financial interests related to this work.

References

[1] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

[2] Lijie Chen, Jian Li, and Mingda Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics (AISTATS)*, 2017.

[3] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR, 2024. URL `https://proceedings.mlr.press/v235/chiang24b.html`.

[4] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28, 1979.

[5] Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[6] Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning (ICML)*, 2020.

[7] Yann Dubois, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024. URL `https://openreview.net/forum?id=CybBmzWBX0`.

[8] Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952.

[9] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027. PMLR, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

[10] Samarth Gupta, Shreyas Chaudhari, Gauri Joshi, and Osman Yağan. Multi-armed bandits with correlated arms. *IEEE Transactions on Information Theory*, 67(10):6711–6732, 2021. doi: 10.1109/TIT.2021.3081508. URL https://arxiv.org/abs/1911.03959.

[11] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 655–662. Omnipress, 2012. URL https://icml.cc/2012/papers/359.pdf.

[12] Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabas Poczos. The multi-fidelity multi-armed bandit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[13] Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL https://jmlr.org/papers/v22/18-798.html.

[14] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1): 1–42, 2016.

[15] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34209–34231. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25h.html. Dataset: https://huggingface.co/datasets/lmarena-ai/arena-hard-auto.

[16] Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[17] Friedrich Pukelsheim. *Optimal Design of Experiments*. SIAM, 2006.

[18] Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=5Nn2BLV7SB.

[19] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Proof of the Lower Bound

A.1 Gauge and sort-identifiability

Equality of all judge means for two instances implies $\Delta\mu_i \mathbf{1} + \Delta \boldsymbol{W} \boldsymbol{\phi}_i = 0$ for all i. Subtracting any two judge rows and using full column rank of $\boldsymbol{\Phi}$ shows that all rows of $\Delta \boldsymbol{W}$ are equal. Consequently the difference is a gauge transformation (3). Equality of the human means on A additionally imposes $\boldsymbol{\Phi}_A \boldsymbol{v} = 0$. The full quality vector is invariant under these differences exactly when $\ker(\boldsymbol{\Phi}_A) = \{0\}$, proving the first part of Theorem 2.

A contrast changes along the residual gauge by $(\boldsymbol{\phi}_p - \boldsymbol{\phi}_q)^\top \boldsymbol{v}$. It is invariant for all $\boldsymbol{v} \in \ker(\boldsymbol{\Phi}_A)$ exactly when its feature difference belongs to $\mathrm{range}(\boldsymbol{\Phi}_A^\top)$. If this fails, a gauge shift of either sign and sufficiently large magnitude reverses the boundary inequality without changing any available law. Two instances with different top-k sets and the same observation law cannot both have error probability below $1/2$: their disjoint correct-output events have probabilities summing to at most one. This proves the remaining claims and Theorem 3.

A.2 Fisher information and the cost change of measure

Vectorize $\boldsymbol{W}$ row by row and put $d = N + Mr$. Let $H_i \in \mathbb{R}^{M \times d}$ be the linear map $H_i(\Delta\boldsymbol{\mu}, \mathrm{vec}\, \Delta \boldsymbol{W}) = \Delta\mu_i \mathbf{1} + \Delta \boldsymbol{W} \boldsymbol{\phi}_i$, and let $\boldsymbol{h}_i = (\boldsymbol{e}_i, \mathbf{0}_{Mr})$. The per-action information and weighted information are

$$I_{J_i} = H_i^\top \boldsymbol{\Sigma}^{-1} H_i, \qquad I_{h_i} = \boldsymbol{h}_i \boldsymbol{h}_i^\top / \sigma_h^2, \qquad \mathcal{I}(\boldsymbol{\omega}) = \sum_a \omega_a I_a. \tag{10}$$

For a parameter difference $\boldsymbol{u}$, the weighted Gaussian KL is $\frac{1}{2} \boldsymbol{u}^\top \mathcal{I}(\boldsymbol{\omega}) \boldsymbol{u}$. All I_a are fixed positive semidefinite matrices. Moreover $I_0 := \sum_a I_a \succ 0$: the human actions identify every quality and, with these fixed, all judge actions and full-rank $\boldsymbol{\Phi}$ identify $\boldsymbol{W}$.

For a procedure with finite expected cost, the usual stopped change-of-measure inequality [9] gives

$$\sum_a \mathbb{E}_{\boldsymbol{\theta}}[N_a(\tau_\delta)] \mathrm{KL}(\nu_a^{\boldsymbol{\theta}} \| \nu_a^{\boldsymbol{\lambda}}) \geq \mathrm{kl}(1-\delta, \delta), \qquad \boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\theta}). \tag{11}$$

Normalize expected counts by expected cost. The resulting allocation belongs to Ω_c, since costs are deterministic and strictly positive. The closure of the alternative set is a union of boundary half-spaces. For $\widetilde{\boldsymbol{e}}_{pq} = (\boldsymbol{e}_p - \boldsymbol{e}_q, \mathbf{0}_{Mr})$, minimization for one pair therefore reduces to

$$\phi_{pq}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \inf_{\widetilde{\boldsymbol{e}}_{pq}^\top \boldsymbol{u} = -\Delta_{pq}} \frac{1}{2} \boldsymbol{u}^\top \mathcal{I}(\boldsymbol{\omega}) \boldsymbol{u} = \frac{\Delta_{pq}^2}{2\mathcal{V}(\mathcal{I}(\boldsymbol{\omega}), \widetilde{\boldsymbol{e}}_{pq})}. \tag{12}$$

The reciprocal is zero if the extended variance is infinite. When the contrast is estimable, the equality follows from constrained quadratic minimization; otherwise a null-space vector can satisfy the constraint at zero KL. Optimizing (11) over allocations gives $T^* \mathrm{kl}(1-\delta, \delta)$ and hence (6), using $\mathrm{kl}(1-\delta, \delta) \geq \log(1/(2.4\delta))$.

A.3 Profiling the nuisance and the general closed form

Let $\tau = \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}$, $\boldsymbol{a} = (\Delta \boldsymbol{W})^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}$, and $\Psi_{\boldsymbol{\omega}} = \sum_i \omega_{J_i} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$. Set $C = \boldsymbol{\Sigma}^{-1/2} \Delta \boldsymbol{W}$ and $\boldsymbol{e} = \boldsymbol{\Sigma}^{-1/2} \mathbf{1}$. Every feasible C has the decomposition $C = \boldsymbol{e} \boldsymbol{a}^\top / \tau + C_\perp$, $C_\perp^\top \boldsymbol{e} = 0$. Thus

$$\mathrm{tr}(C^\top C \Psi_{\boldsymbol{\omega}}) = \boldsymbol{a}^\top \Psi_{\boldsymbol{\omega}} \boldsymbol{a} / \tau + \mathrm{tr}(C_\perp^\top C_\perp \Psi_{\boldsymbol{\omega}}) \geq \boldsymbol{a}^\top \Psi_{\boldsymbol{\omega}} \boldsymbol{a} / \tau.$$

Equality is attained by $\Delta \boldsymbol{W} = \mathbf{1} \boldsymbol{a}^\top / \tau$. This proof does not require $\Psi_{\boldsymbol{\omega}}$ to be invertible.

After this profiling, the quadratic KL is

$$\frac{1}{2} \sum_i A_i \Delta\mu_i^2 + \boldsymbol{a}^\top \sum_i \omega_{J_i} \Delta\mu_i \boldsymbol{\phi}_i + \frac{1}{2\tau} \boldsymbol{a}^\top \Psi_{\boldsymbol{\omega}} \boldsymbol{a}. \tag{13}$$

Put $D = \mathrm{diag}(A_i)$ and let the ith row of B be $\omega_{J_i} \boldsymbol{\phi}_i^\top$. Its information matrix is

$$\widetilde{I} = \begin{pmatrix} D & B \\ B^\top & \Psi_{\boldsymbol{\omega}}/\tau \end{pmatrix}, \qquad \Psi_{\boldsymbol{\omega}}/\tau - B^\top D^\dagger B = G(\boldsymbol{\omega}).$$

The last identity follows term by term from

$$\frac{\omega_{J_i}}{\tau} - \frac{\omega_{J_i}^2}{A_i} = \frac{1}{\tau}\frac{\omega_{J_i}\omega_{h_i}/\sigma_h^2}{A_i}. \tag{14}$$

Zero rows of D also have zero rows of B and contribute zero to this identity.

To compute a contrast variance, complete the square in the conjugate representation for $\widetilde{I}$. With $\boldsymbol{z}$ its nuisance coordinate and $\boldsymbol{x}$ its quality coordinate, set $\boldsymbol{v} = \boldsymbol{x} + D^\dagger B\boldsymbol{z}$. For $\boldsymbol{c} = \boldsymbol{e}_p - \boldsymbol{e}_q$ and $A_p, A_q > 0$, the supremum splits into

$$\sup_{\boldsymbol{v}}\{2\boldsymbol{c}^\top\boldsymbol{v} - \boldsymbol{v}^\top D\boldsymbol{v}\} + \sup_{\boldsymbol{z}}\{-2\boldsymbol{b}_{pq}^\top\boldsymbol{z} - \boldsymbol{z}^\top G\boldsymbol{z}\}.$$

The first value is $1/A_p + 1/A_q$; the second is the extended leverage in (5). Therefore

$$\mathcal{V}(\mathcal{I}(\boldsymbol{\omega}), \widetilde{\boldsymbol{e}}_{pq}) = \frac{1}{A_p} + \frac{1}{A_q} + \mathcal{V}(G(\boldsymbol{\omega}), \boldsymbol{b}_{pq}). \tag{15}$$

If an endpoint has $A_i = 0$, its quality is unobserved and the variance is infinite. This proves the general rate, including pure-human allocations and anchored boundary endpoints, and completes Theorem 5.

B Proof of the Oracle Upper Bound

B.1 A conservative stopping threshold

Time-uniform deviation bounds under adaptive sampling are central to sequential stopping [13]. For the joint Gaussian nuisance model here, a direct finite-count argument gives the following conservative threshold. Let $d = N + Mr$ and let $t = \sum_a N_a(t)$ count all action queries, including the pilot. Define

$$x(t,\delta) = \log\frac{\pi^2}{6\delta} + (K+2)\log(t+1), \tag{16}$$

$$\beta(t,\delta) = \frac{d}{2} + \sqrt{d\,x(t,\delta)} + x(t,\delta). \tag{17}$$

This is the threshold for the known-input oracle correctness statement. The reported stopping simulations use the practical stitched threshold $\beta_{\mathrm{prac}}(t,\delta) = \log(1/\delta) + 2\log(\log(t+2)+1)$. Their results concern that practical rule, rather than empirical validation of (17) or a finite-δ guarantee for the practical rule.

Finite-confidence correctness. Couple the observations to an independent infinite sample sequence for each action. For a fixed count vector $\boldsymbol{n} \in \mathbb{N}_0^K$, take the first n_a observations from each sequence and whiten them by their known channel covariances. This is a Gaussian linear model with a deterministic design. If its design rank is $r_{\boldsymbol{n}} \le d$, its maximized log-likelihood ratio against the true parameter satisfies

$$Q_{\boldsymbol{n}} := \sup_{\boldsymbol{\lambda}} \log\frac{p_{\boldsymbol{\lambda}}(\text{prefix data})}{p_{\boldsymbol{\theta}}(\text{prefix data})} \overset{\text{law}}{=} \tfrac{1}{2}\chi^2_{r_{\boldsymbol{n}}}.$$

Indeed, the maximized likelihood gain is one half of the squared norm of standard Gaussian noise projected onto the whitened design's column space. For $Y \sim \chi^2_d$ and $0 < \lambda < 1/2$,

$$\log \mathbb{E}e^{\lambda(Y-d)} = -d\lambda - \frac{d}{2}\log(1-2\lambda) \le \frac{d\lambda^2}{1-2\lambda}.$$

Chernoff's inequality with $\lambda = \sqrt{x}/(\sqrt{d} + 2\sqrt{x})$ gives $\mathbb{P}(Y \ge d + 2\sqrt{dx} + 2x) \le e^{-x}$. The same upper bound holds for any smaller rank. At total query count t, there are at most $(t+1)^K$ possible count vectors. A union bound over these deterministic prefixes and all times therefore gives

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}\big(\exists t \ge 0 : Q_{\boldsymbol{N}(t)} \ge \beta(t,\delta)\big) &\le \sum_{t\ge 0}(t+1)^K e^{-x(t,\delta)} \\ &= \frac{6\delta}{\pi^2}\sum_{t\ge 0}(t+1)^{-2} = \delta.\end{aligned}$$

Every adaptively collected sample is one of these prefixes; the argument does not condition on the adaptive counts. On an incorrect stopping event, the true parameter belongs to $\mathrm{Alt}(\hat{\boldsymbol{\theta}}_t)$, and the Gaussian MLE identity then yields

$$Z(t) \le \sum_a N_a(t)\mathrm{KL}(\nu_a^{\hat{\boldsymbol{\theta}}_t} \| \nu_a^{\boldsymbol{\theta}}) = Q_{\boldsymbol{N}(t)}.$$

Thus the probability of an incorrect output under (17) is at most δ.

Almost-sure termination under Algorithm 1 follows independently from the exploration floor (21) and the MLE consistency proved in Appendix B.3. Here $I_t = \sum_a N_a(t) I_a$ is the accumulated Fisher information. Put $\Delta_{\min} = \min_{(p,q)\in \mathrm{bdry}(\boldsymbol{\mu})} \Delta_{pq} > 0$. Eventually the estimated boundary is correct and every estimated boundary gap is at least $\Delta_{\min}/2$. Since $\|\widetilde{e}_{pq}\|^2 = 2$, the profiled GLR then satisfies

$$Z(t) \ge \lambda_{\min}(I_t)\Delta_{\min}^2/16 \ge \kappa_{\mathrm{floor}}\Delta_{\min}^2\sqrt{C(t)}/16.$$

For fixed δ, the threshold is $O_\delta(\log(t+1))$, whereas $t \le C(t)/c_{\min}$. It is therefore crossed at a finite time almost surely. This proves the finite-confidence correctness component independently of the tracking argument needed for the matching cost constant.

B.2 Exploration and tracking statements

Put $c_{\min} = \min_a c_a$, $c_{\max} = \max_a c_a$. The finite pilot makes the initial information positive definite. The submitted tracking lemma states that all actions are pulled infinitely often, forced exploration has $o(C)$ total cost, and $N(t)/C(t) \to \boldsymbol{\omega}^*(\boldsymbol{\theta})$ almost surely. The exact placement of the query price in the submitted trigger is $N_a < \sqrt{C/(Kc_a)}$. The following bounds retain that trigger and its minimum-cost-count selection rule, while making the cost constants explicit. Set $S_c = \sum_a c_a$ and $\alpha_c = \sqrt{c_{\min}/(2K)}$. Whenever action a is selected by forced exploration, its cost count is no larger than the cost count of some deficient action, and hence no larger than $g(C) = \sqrt{c_{\max}C/K}$. Its number of forced pulls therefore satisfies

$$F_a(t) \le \frac{\sqrt{c_{\max}C(t)/K}}{c_a} + 1, \qquad \sum_a c_a F_a(t) \le \sqrt{Kc_{\max}C(t)} + S_c. \tag{18}$$

Thus forced exploration has $O(\sqrt{C})$ total cost for the exact algorithm.

A uniform lower floor also holds at times when the forced branch is active. For a terminal cost $x = C(t)$, choose the first round s with $C(s) \ge x/2$; its cost is at most $x/2 + c_{\max}$. Suppose an action has terminal cost count $L = c_a N_a(t) < \sqrt{c_{\min}x/(2K)}$. For every $u \in \{s, \dots, t-1\}$, its current count then obeys

$$c_a N_a(u) \le L < \sqrt{c_{\min}x/(2K)} \le \sqrt{c_a C(u)/K}.$$

It remains deficient throughout this interval, so every intervening pull is forced. The selected action has pre-pull cost count at most L. For each action b, its total cost increment on the interval is consequently at most $L + c_b$, and summing gives

$$x/2 - c_{\max} \le C(t) - C(s) \le KL + S_c < \sqrt{Kc_{\min}x/2} + S_c.$$

This is impossible for $x \ge C_E := 8Kc_{\min} + 4(S_c + c_{\max}) + 1$. Therefore, uniformly over all action sequences generated by the algorithm,

$$N_a(t) \ge \frac{\alpha_c\sqrt{C(t)}}{c_a}, \qquad C(t) \ge C_E. \tag{19}$$

This includes the human actions that identify the gauge. It supplies the $\sqrt{C}$ information floor used in the rebuttal without assuming its simplified price normalization.

The quantitative tracking stability property used in the final posted argument is the following: if all plug-in targets after cost C_1 are within η of $\boldsymbol{\omega} \in \Omega_c$, then constants independent of C_1, η, δ give

$$\|N(t)/C(t) - \boldsymbol{\omega}\|_\infty \le L_0\eta + K_1/\sqrt{C(t)} + K_2(1 + C_1)/C(t). \tag{20}$$

This property is the tracking input in the rebuttal's last-exit proof. It is stated for the submitted one-step cost-aware tracking rule; no alternative tracking algorithm is introduced here.

B.3 Cost-indexed Gaussian concentration

Lemma 10 (Cost-indexed tail; rebuttal Lemma 22′)**.** *Under the exploration floor* (19)*, for any $\xi > 0$ there are constants $B, b_{\boldsymbol{\theta}} > 0$ and $C_0(\xi) < \infty$ such that*

$$\mathbb{P}\Big(\exists t : C(t) \geq x,\ \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\| > \xi\Big) \leq B \exp(-b_{\boldsymbol{\theta}} \xi^2 \sqrt{x}), \quad x \geq C_0(\xi).$$

The last cost at which the estimate is outside this neighborhood has finite expectation.

Proof. Let $I_t = \sum_a N_a(t) I_a$ and let M_t be the Gaussian score at the true parameter. After the full-rank pilot, I_t is invertible and $\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta} = I_t^{-1} M_t$. Set $\lambda_0 = \lambda_{\min}(I_0) > 0$ and $\kappa_{\max} = \max_a \|I_a\|_{\mathrm{op}}/c_a$. Define $\kappa_{\mathrm{floor}} = \lambda_0 \alpha_c / c_{\max} > 0$. By (19), for $C(t) \geq C_E$,

$$\lambda_{\min}(I_t) \geq \kappa_{\mathrm{floor}} \sqrt{C(t)}, \qquad \lambda_{\max}(I_t) \leq \kappa_{\max} C(t). \tag{21}$$

For each fixed $\boldsymbol{u}$, $\exp(\boldsymbol{u}^\top M_t - \boldsymbol{u}^\top I_t \boldsymbol{u}/2)$ is a nonnegative mean-one martingale. Integrating against a standard Gaussian prior and applying Ville's inequality gives

$$\mathbb{P}\Big(\exists t : \|M_t\|^2_{(I_t + I_d)^{-1}} > 2\log(1/\eta) + \log\det(I_d + I_t)\Big) \leq \eta. \tag{22}$$

Here I_d is the identity matrix, not Fisher information. If $\lambda_{\min}(I_t) \geq L \geq 1$, eigenvalue comparison shows $\|I_t^{-1} M_t\|^2 \leq 2\|M_t\|^2_{(I_t+I_d)^{-1}}/L$.

On the dyadic cost block $C(t) \in [2^j, 2^{j+1})$, put $L_j = \kappa_{\mathrm{floor}} 2^{j/2}$ and $\eta_j = \exp(-\xi^2 L_j/8)$. For sufficiently large j, $L_j \geq 1$ and $d\log(1 + 2\kappa_{\max} 2^j) \leq \xi^2 L_j/4$. Using (22) then bounds the probability of any $\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\| > \xi$ on that block by $\exp[-\xi^2 \kappa_{\mathrm{floor}} 2^{j/2}/8]$. For sufficiently large $C_0(\xi)$, summing the dyadic blocks gives $B = 3$ and $b_{\boldsymbol{\theta}} = \kappa_{\mathrm{floor}}/(8\sqrt{2})$, with

$$\mathbb{P}\Big(\exists t : C(t) \geq x,\ \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\| > \xi\Big) \leq B \exp(-b_{\boldsymbol{\theta}} \xi^2 \sqrt{x}), \qquad x \geq C_0(\xi). \tag{23}$$

For the corresponding last bad cost T_ξ,

$$\mathbb{E}[T_\xi] \leq C_0(\xi) + \frac{2B}{(b_{\boldsymbol{\theta}} \xi^2)^2} < \infty.$$

This is the cost-indexed tail bound supplied in the rebuttal. Almost-sure consistency alone would not justify an expected stopping-cost statement. □

B.4 Continuity and passage to expected cost

For fixed contrast $\boldsymbol{c}$, the extended variance is lower semicontinuous. It is continuous at every allocation where it is finite: if $\boldsymbol{\omega}_n \to \boldsymbol{\omega}$ then $\mathcal{I}(\boldsymbol{\omega}_n) \succeq (1 - \epsilon_n)\mathcal{I}(\boldsymbol{\omega})$ for some $\epsilon_n \to 0$, by retaining the finitely many actions positive at $\boldsymbol{\omega}$. The conjugate representation therefore bounds its variance above by $\mathcal{V}(\mathcal{I}(\boldsymbol{\omega}), \boldsymbol{c})/(1 - \epsilon_n)$; lower semicontinuity gives the reverse limit inequality. At a non-estimable allocation the variance tends to infinity, again by lower semicontinuity. Consequently its reciprocal, with value zero there, is continuous throughout Ω_c.

In a neighborhood with the same top-k set, the rate $F(\boldsymbol{\omega}, \boldsymbol{\theta}) = \min_{pq} \phi_{pq}(\boldsymbol{\omega}, \boldsymbol{\theta})$ is jointly continuous and Ω_c is compact. Berge's theorem and uniqueness of the oracle allocation give continuity of $\boldsymbol{\omega}^*(\boldsymbol{\theta})$. The concentration bound and the submitted tracking convergence imply $\hat{\boldsymbol{\theta}}_t \to \boldsymbol{\theta}$ and $N(t)/C(t) \to \boldsymbol{\omega}^*(\boldsymbol{\theta})$ almost surely. Since a strictly positive design has finite pair costs, the optimal rate is positive. Its contrasts are estimable in $\mathcal{I}(\boldsymbol{\omega}^*)$; for positive endpoint information this is the condition $\boldsymbol{b}_{pq}(\boldsymbol{\omega}^*) \in \mathrm{range}(G(\boldsymbol{\omega}^*))$.

Last-exit step; rebuttal Lemma 22″. The following step uses the tracking stability property (20), as in the posted final proof. Fix $\epsilon \in (0, 1)$. Choose ξ smaller than one quarter of the minimum boundary gap and sufficiently small that all subsequent target allocations are within a prescribed η of $\boldsymbol{\omega}^*$. Continuity of F and (20) give a random $T'_\epsilon \leq K_\epsilon(T_\xi \vee C_\epsilon) + K_\epsilon$ such that

$$Z(t) \geq \frac{(1 - \epsilon)C(t)}{T^*(\boldsymbol{\theta})}, \qquad C(t) \geq T'_\epsilon. \tag{24}$$

The constants are independent of δ. In particular T'_ϵ has a tail of the form $B_\epsilon e^{-b_\epsilon\sqrt{x}}$ and finite expectation.

Let b_δ be a deterministic cost large enough that, for every $x \geq b_\delta$,

$$\frac{(1-\epsilon)x}{T^*(\boldsymbol{\theta})} \geq \beta(\lceil x/c_{\min}\rceil + 1, \delta).$$

Such a value exists because the threshold grows sublinearly in x. For (17), write $L = \log(1/\delta)$ and extend the threshold to nonnegative real query counts. The continuous envelope $\bar{\beta}_\delta(x) = \beta(x/c_{\min} + 2, \delta)$ upper-bounds $\beta(\lceil x/c_{\min}\rceil + 1, \delta)$. Its ratio to x decreases to zero. To see the monotonicity, set $v = L + \log(\pi^2/6) + (K+2)\log(x/c_{\min} + 3) > K + 2$; then

$$x\bar{\beta}'_\delta(x) \leq (K+2)\left(1 + \frac{\sqrt{d}}{2\sqrt{v}}\right) \leq v + \sqrt{dv} + d/2 = \bar{\beta}_\delta(x).$$

For every fixed $s > 0$, $\bar{\beta}_\delta(sL)/L \to 1$ as $\delta \to 0$, since the extra terms are $O(\sqrt{L} + \log L)$, and $\bar{\beta}_\delta(x) \geq L$. Choosing b_δ as the crossing of this envelope with $(1-\epsilon)x/T^*(\boldsymbol{\theta})$ therefore gives $b_\delta/L \to T^*(\boldsymbol{\theta})/(1-\epsilon)$. This deterministic threshold-growth step does not supply the separate tracking stability property used to obtain (24). At the first round exceeding $\max(T'_\epsilon, b_\delta)$, the stopping condition holds by (24); the overshoot is at most $c_{\max}$. Thus

$$\mathbb{E}[C(\tau_\delta)] \leq b_\delta + c_{\max} + \int_{b_\delta}^{\infty} \mathbb{P}(T'_\epsilon > x)\, dx.$$

The integral tends to zero by the explicit tail. Divide by $\log(1/\delta)$ and let $\epsilon \downarrow 0$ to prove Theorem 7.

B.5 Pilot and estimated-input scope

The r pilot rows exist because $\boldsymbol{\Phi}$ has full column rank. With one judge query per arm, any null direction of the initial design must be a gauge direction. The r independent human pilot rows remove it, so the initial information is positive definite. Pilot cost is $Nc_J + rc_h$. One label per pilot anchor identifies the design but does not by itself give consistent estimates; the continuing exploration floor is essential.

A boundary-span pilot can be useful with a gauge-robust warm start, but the formal theorem uses the full-rank pilot. The population span criterion does not by itself prove finite-time consistency from finitely many labels.

B.6 Conditional perturbation of calibrated allocations

An invertible coordinate change $\boldsymbol{\Phi} \mapsto \boldsymbol{\Phi}R$ transforms $G \mapsto R^\top G R$ and $\boldsymbol{b}_{pq} \mapsto R^\top \boldsymbol{b}_{pq}$. The conjugate variance, including its infinite value outside the information range, is invariant under this simultaneous change.

Suppose an independent calibration supplies aligned features and covariance within ϵ of the true inputs. On a compact neighborhood of feasible designs where all relevant contrasts remain estimable and their restricted information eigenvalues have a common positive lower bound, the pair costs are locally Lipschitz in these inputs. If $\widehat{f}$ and f denote the calibrated and true maximum pair-cost objectives and $\sup_{\boldsymbol{\omega}} |\widehat{f}(\boldsymbol{\omega}) - f(\boldsymbol{\omega})| \leq L\epsilon$ on that neighborhood, then a calibrated minimizer $\hat{\boldsymbol{\omega}}$ satisfies

$$f(\hat{\boldsymbol{\omega}}) \leq \widehat{f}(\hat{\boldsymbol{\omega}}) + L\epsilon \leq \widehat{f}(\boldsymbol{\omega}^*) + L\epsilon \leq f(\boldsymbol{\omega}^*) + 2L\epsilon.$$

This bound concerns the maximum objective, rather than every individual pair's change relative to its value at the oracle optimum. Obtaining such a calibration event from SVD requires an identifiable signal, a singular-value gap, and appropriate coordinate normalization. Extending the oracle stopping guarantee additionally requires control of calibration failure and stopping-cost tails. These conditions are not established by a descriptive low-rank fit or an O_P estimation rate alone.

C Convexity and Allocation Design Details

C.1 Global convexity and lower semicontinuity

For $X \succeq 0$, define

$$\mathcal{V}(X, \boldsymbol{c}) = \sup_{\boldsymbol{y} \in \mathbb{R}^d} \{2\boldsymbol{c}^\top \boldsymbol{y} - \boldsymbol{y}^\top X \boldsymbol{y}\} = \begin{cases} \boldsymbol{c}^\top X^\dagger \boldsymbol{c}, & \boldsymbol{c} \in \operatorname{range}(X), \\ +\infty, & \boldsymbol{c} \notin \operatorname{range}(X). \end{cases} \tag{25}$$

If $\boldsymbol{c} \in \operatorname{range}(X)$, complete the square about $X^\dagger \boldsymbol{c}$. Otherwise let $\boldsymbol{z}$ be the nonzero projection of $\boldsymbol{c}$ onto $\ker(X)$; the expression at $\boldsymbol{y} = t\boldsymbol{z}$ equals $2t\|\boldsymbol{z}\|^2$ and diverges. For fixed $\boldsymbol{y}$, the conjugate expression is affine and continuous in $\boldsymbol{\omega}$ when $X = \mathcal{I}(\boldsymbol{\omega})$. Its supremum is therefore convex and lower semicontinuous on the entire cost simplex, proving Lemma 6 across support changes. The finite maximum over boundary pairs preserves both properties. Ω_c is compact, and a strictly positive design has positive-definite information and finite pair costs, so the finite optimum is attained.

C.2 Calibration split and KKT interpretation

Let $u = \tau\omega_J$, $v = \omega_h/\sigma_h^2$ and $g(u, v) = uv/(u + v)$. Then $G = \tau^{-2} \sum_i g(u_i, v_i)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$. The derivatives $g_u = v^2/(u + v)^2$, $g_v = u^2/(u + v)^2$ and the Hessian

$$\nabla^2 g = \frac{2}{(u+v)^3} \begin{pmatrix} -v^2 & uv \\ uv & -u^2 \end{pmatrix} \preceq 0$$

give concavity of the anchor information contribution.

For a calibration-only arm and fixed contrast, the objective derivatives through G have a common quadratic factor and ratio

$$\frac{\partial f/\partial \omega_J}{\partial f/\partial \omega_h} = \sigma_h^2 \tau \left(\frac{v}{u}\right)^2.$$

The same ratio holds for a weighted sum of active contrasts. Interior KKT stationarity equates it to c_J/c_h, yielding (8). This argument requires that the arm affect the active objective only through the calibration Gram. It omits the direct and $\boldsymbol{b}_{pq}$ derivatives of an active boundary endpoint, so it cannot be applied to every anchor in the general program.

At a minimizer with $\mathcal{I}(\boldsymbol{\omega}) \succ 0$, the KKT rules can be written using $\boldsymbol{v}_{pq} = \mathcal{I}(\boldsymbol{\omega})^{-1}\widetilde{\boldsymbol{e}}_{pq}$. For active-pair multipliers $z_{pq} \geq 0$, $\sum_{pq} z_{pq} = 1$, the marginal benefit of action a is $\sum_{pq} 2z_{pq}\boldsymbol{v}_{pq}^\top I_a \boldsymbol{v}_{pq}/\Delta_{pq}^2$. At a positive allocation this benefit equals λc_a; at a zero allocation it is at most λc_a. Positive multipliers occur only for pairs attaining the maximum cost. This includes direct human information on boundary arms. Singular designs are covered by the extended convex formulation (25); the explicit derivative rule above is not asserted there. Replacing the inverse by a Moore–Penrose inverse need not give a valid derivative when the information rank changes.

C.3 Finite-contrast anchor design

For a fixed calibration split and fixed target directions, write $M(\boldsymbol{\alpha}) = \sum_i \alpha_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$. The criterion in (9) is a convex finite-contrast optimal-design problem. At a nonsingular optimum on the relevant design space, multipliers on its active contrasts give the equivalence conditions

$$\sum_{pq} z_{pq} (\boldsymbol{\phi}_i^\top M^{-1} \boldsymbol{d}_{pq})^2 \begin{cases} = \lambda, & \alpha_i > 0, \\ \leq \lambda, & \alpha_i = 0. \end{cases}$$

These conditions describe the restricted design problem without claiming that a general min–max contrast design is given by the single-contrast Elfving construction. Coverage by available candidate features and good conditioning in the tested directions remain the operational requirements.

D Experimental Details and Additional Results

D.1 Real human-anchor construction and bias test

The real-vote instances split votes at the prompt/session level into three disjoint subsets. The calibration subset fixes opponent strengths and the score normalization; the target subset defines $\boldsymbol{\mu}$

by Bradley–Terry–Elo; the anchor subset supplies labels to the algorithms. No vote is reused across these three roles. Judges score the original responses seen by voters. GPT-4 votes in the MT-Bench human-judgment dataset are excluded from the judge channel. An anchor query for arm i consists of $g = 20$ votes and fits a one-parameter Bradley–Terry model with calibration opponent strengths held fixed. A whole block costs $c_h = 30c_J$, rather than each individual vote costing $30c_J$. The resulting fitted block labels are empirical observations, not raw independent Gaussian labels; their use does not automatically satisfy the Gaussian theorem's assumptions.

The MT-Bench instance has $N = 6, k = 3$, with 338/676/675 calibration/target/anchor sessions and 3,355 human votes in total. It adds 4,800 judge scores ($6 \times 10 \times 80$). The Chatbot Arena instance has $N = 20, k = 8$, with 5,000/9,600/26,000 votes in the three subsets and 39,200 new judge scores ($1{,}960 \times 2 \times 10$ response scores). The factor subspace is estimated from judge scores with the anchor channel held out; each judge's distortion is evaluated leave-one-out. The target fixes the quality–bias gauge for this comparison. Per-arm anchor-bias intervals relative to the held-out target cover zero, and the maximum reported absolute bias is 0.04.

To avoid dependence on the choice of factor coordinates, bias amplitude uses the fitted distortion in arm space,

$$\zeta = \frac{\|\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{w}}_{\text{anchor}}\|}{M^{-1}\sum_{j=1}^{M}\|\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{w}}_j\|}.$$

This ratio is invariant under an invertible change of feature coordinates. On Arena, the human amplitude is 0.12, the null 95th percentile is 0.26, and $p = 0.57$. On MT-Bench the corresponding values are 0.13, 0.31, and $p = 0.52$. Judge leave-one-out amplitudes span 0.48–1.61 on Arena and 0.45–1.58 on MT-Bench, all with $p < 0.001$ against the respective null thresholds. These comparisons detect larger judge distortion but do not certify zero anchor bias. MT-Bench boundary accuracy is 74.1% for Independent-proxy and 96.4% for P-T&S.

Power and detectable scope. Table 4 reports random-direction bias injection into the *anchor split only*, with 200 repetitions per amplitude. Linear interpolation places 80% power near $\zeta = 0.34$ on Arena and 0.45 on MT-Bench. At $\zeta = 0.35$, power is 86% and 60.5%, respectively; MT-Bench has limited power at that amplitude. The observed amplitudes have reported one-sided 95% upper bounds 0.24 (Arena) and 0.30 (MT-Bench). The test addresses differential bias along the estimated judge-bias directions. A preference shift shared by calibration, target, and anchor voters enters the definition of $\boldsymbol{\mu}$ and is invisible to this test. Neither non-rejection nor these upper bounds establishes absolute unbiasedness of human preference.

Table 4: Bias-test power: rejection counts out of 200 anchor-only injections.

Injected ζ	0.10	0.20	0.30	0.35	0.40	0.50
Chatbot Arena	18	42	113	172	175	198
MT-Bench	35	59	96	121	146	177

D.2 Full frozen-judge recovery comparison

This fixed-budget comparison uses the frozen $N = 12, M = 10$ Arena-Hard instance, top-5, $c_h/c_J = 30$, $B = 5{,}000$, and 200 seeds. Judge scores are real; the human channel is simulated around the benchmark-derived target. Static methods use the allocation adaptations in Appendix D.6 and oracle allocations, fixed within a run. Adaptive P-T&S uses plug-in estimates and a budget truncation. Table 5 reproduces all reported methods and intervals. The separate budget sweep targets 95% *average boundary-pair* accuracy; it is neither a 95% exact recovery threshold nor a fixed-confidence stopping guarantee.

Hardest tercile. The 12 smallest of the 35 target boundary gaps lie in $[0.10, 0.98]$. Their results appear in Table 3. The other two terciles have 99.5–99.9% accuracy, motivating this harder subset. Paired budget ratios to adaptive P-T&S at the 90% hardest-pair threshold are 1.38 $[1.01, 1.90]$ (Pure-human), 1.30 $[0.81, 1.64]$ (Human-only oracle), and 1.07 $[0.85, 1.43]$ (static P-T&S). Paired accuracy differences favor adaptive P-T&S by 2.1 percentage points $[0.0, 4.2]$ over Pure-human and

Table 5: Full R1 recovery comparison. Percentages and reported 95% intervals.

Method	Exact top-5	Set overlap	Boundary accuracy	Budget to 95%
P-T&S adaptive	47.5 [40.0,53.6]	89.5 [88.2,90.9]	96.8 [96.2,97.3]	2,860 [2,480,3,310]
P-T&S static	42.5 [35.0,48.5]	87.2 [85.4,88.7]	96.5 [95.8,97.0]	3,270 [2,970,3,780]
Human-only oracle	34.5 [27.8,41.0]	86.1 [84.5,87.5]	95.9 [95.2,96.4]	3,130 [2,980,3,470]
Pure-human	38.0 [31.0,44.0]	87.4 [86.0,88.7]	96.2 [95.6,96.7]	3,420 [2,950,3,960]
Uniform	26.5 [20.0,32.0]	83.5 [81.6,85.0]	93.6 [92.7,94.3]	7,030 [6,260,8,440]
DS-Active	29.0 [22.5,35.0]	84.5 [82.9,85.9]	94.1 [93.2,94.8]	7,050 [5,830,8,270]
Estimate-and-correct	53.5 [46.5,60.0]	79.3 [74.7,83.4]	85.4 [81.1,89.2]	9,840 [7,450,13,170]
PPI-inspired	11.0 [6.5,15.5]	75.7 [73.8,77.4]	90.0 [89.0,91.0]	11,130 [9,620,12,690]
MF-BAI adaptation	21.5 [16.0,27.5]	53.8 [48.8,58.6]	65.4 [61.0,69.5]	non-monotone
Heuristic anchor	47.5 [40.5,54.0]	63.0 [57.0,68.7]	67.2 [61.4,72.7]	not reached
Independent-proxy	0.0*	56.7 [53.3,59.7]	70.5 [66.7,73.8]	not reached

*No exact successes in 200 runs. The source reports $[0, 0]$; this is not a population-probability upper bound.

2.0 $[-0.2, 4.1]$ over the oracle. The full-pair 95% and hardest-pair 90% thresholds differ, so their ratios do not establish a controlled trend with pair difficulty.

D.3 Full real-vote comparison on the larger roster

The Chatbot Arena comparison uses the real-vote block construction above, $N = 20$, $M = 10$, $k = 8$, $c_h/c_J = 30$, $B = 20{,}000$, and 200 seeds. Static baselines and the static/adaptive P-T&S variants use the same protocol. Some threshold budgets exceed the finite available vote pool: daggered values extrapolate under reuse of vote blocks, and do not represent collection of fresh independent votes at that cost.

Table 6: R2 on real Chatbot Arena vote blocks. Percentages and reported 95% intervals. †: threshold extrapolated with block reuse beyond the finite vote pool.

Method	Exact top-8	Set overlap	Boundary accuracy	Budget to 95%
P-T&S adaptive	71.5 [65.6,77.0]	97.0 [96.3,97.7]	96.5 [95.8,97.2]	13,400 [12,300,14,700]
P-T&S static	67.0 [60.9,72.7]	96.5 [95.7,97.2]	95.9 [95.1,96.6]	15,100 [13,800,16,600]
DS-Active	45.5 [39.1,52.0]	93.4 [92.4,94.4]	94.4 [93.4,95.3]	23,100 [21,000,25,600]
PPI-inspired	37.0 [30.9,43.5]	92.0 [90.8,93.1]	93.5 [92.4,94.5]	28,600 [25,800,31,900]
Human-only oracle	34.5 [28.5,41.0]	91.5 [90.3,92.6]	93.1 [91.9,94.2]	26,800 [24,200, 29,800]†
MF-BAI adaptation	33.5 [27.6,39.9]	91.2 [90.0,92.4]	93.0 [91.8,94.1]	30,200 [27,100,33,900]
Estimate-and-correct	27.0 [21.5,33.1]	89.6 [88.3,90.9]	91.9 [90.6,93.1]	36,700 [32,600,41,600]
Pure-human	24.5 [19.2,30.4]	88.9 [87.5,90.3]	91.4 [90.0,92.7]	38,400 [34,000, 43,600]†
Heuristic anchor	16.5 [12.1,21.8]	85.6 [84.0,87.2]	88.9 [87.4,90.4]	44,100 [38,800, 50,400]†
Uniform	10.5 [7.0,15.0]	82.9 [81.1,84.7]	87.2 [85.5,88.8]	53,900 [47,000, 62,300]†
Independent-proxy	0.5 [0.0,1.9]	74.2 [72.2,76.2]	71.9 [70.1,73.7]	not reached

The analytical DS-Active/P-T&S ratio is 1.52 on this instance versus 1.12 on the smaller roster. Roster and prompt differences prevent interpreting this as a scaling law. A within-instance N/k sweep would be needed to isolate scale; no such result is claimed here.

D.4 Anchor-bias sensitivity

Bias is injected into the simulated anchor channel of the frozen R1 instance, with static allocations and an estimator unaware of the injection. Within each seed, judge draws are shared across ζ values. We compare the worst direction and the median of 20 random directions. The normalization is the arm-space amplitude defined above. Table 7 reports boundary accuracy and the difference from the best competitor. The reported standard errors for the relative gap range from 0.3 to 0.8 percentage points.

At zero bias, the relative gap is only $+0.3$ percentage points. Worst-direction absolute accuracy falls by 18.1 points by $\zeta = 1$; the -1.0-point relative gap at that amplitude is a different quantity. The random-direction median remains nearly flat (96.5 to 96.7%), reflecting the direction dependence of boundary damage. Methods using the anchor channel degrade together; Pure-human stays within one percentage point of P-T&S, rather than remaining close to its own zero-bias accuracy. Independent-proxy remains at 70.5%.

Table 7: Bias injection under the static R1 protocol. Accuracies are percentages; relative gaps are in percentage points.

ζ	0	0.15	0.35	0.50	0.75	1.00
Worst-direction accuracy	96.5	95.2	93.3	91.4	86.4	78.4
Median random-direction accuracy	96.5	96.6	96.6	96.7	96.7	96.7
Gap to best competitor	$+0.3$	-0.0	-0.2	-0.0	-0.9	-1.0

D.5 Bootstrap uncertainty and additional audits

Paired prompt bootstrap of the analytical objective. The bootstrap draws 1,000 resamples of the 200 prompts and reestimates features, covariance, target, and every allocation within each replicate. Table 8 gives ratios of analytical T^* to the P-T&S allocation at $c_h/c_J = 30$. The intervals for DS-Active, Uniform, PPI-inspired, and Estimate-and-correct include one. Thus the fitted objective does not establish a robust constant-factor advantage over these competitors on this instance.

Table 8: Paired prompt bootstrap of analytical T^* ratios to P-T&S. The final column is the fraction of replicates favoring P-T&S.

Method	Median ratio	95% interval	Favoring P-T&S
DS-Active	1.12	[0.57,3.54]	72%
Uniform	1.28	[0.63,4.12]	85%
PPI-inspired	1.94	[0.86,4.96]	96%
Estimate-and-correct	7.31	[0.93,27.3]	97%
MF-BAI adaptation	7.76	[2.32,69.1]	100%
Heuristic anchor	18.8	[7.7,28.7]	100%
Pure-human	150	[55,1251]	100%

Query-level covariance. The query-level residual common shock accounts for 62.5% of residual variance; mean off-diagonal correlation is $+0.60$. After per-judge standardization, estimated effective information is $\hat{\tau} = 1.7$ for 10 judges. The original prompt-averaged, row-centered figure removes this common shock and represents a different covariance; its correlations must not be substituted into the query-level information formula. At the original $c_h/c_J = 30$ allocation, analytical $T^* = 947.3$; anchor leverage accounts for 95.4% of the maximizing pair's cost term and 83–99% across pairs.

Distribution of analytical anchor-design gaps. Across 30 draws from the synthetic $N = 8, M = 6, r = 2, k = 3$ family, the adversarial-to-boundary-aware analytical T^* ratio has median 236, with interquartile range $[81, 4{,}440]$. The random-to-boundary-aware ratio is 3.9; the earlier single-instance 74-fold extreme is near the 25th percentile of the adversarial distribution. These are analytical allocation comparisons, not measured recovery speedups.

Heavy-tail sensitivity. The synthetic rate instance is rerun with variance-normalized Student t_ν noise, the unchanged Gaussian algorithm, $\delta = 0.1$, and 400 runs per cell. Stopping-cost ratios to Gaussian noise are 0.97 $[0.88, 1.06]$ for t_{10}, 0.94 $[0.85, 1.03]$ for t_5, and 0.99 $[0.90, 1.08]$ for t_3; each cell has zero errors in 400 runs. At the calibrated budget, boundary accuracy is $95.8/95.6/96.0/96.3\%$ for Gaussian/t_{10}/t_5/t_3 noise; at twice that budget it is $97.9/97.9/97.8/98.0\%$, with standard errors approximately 0.2 points. This sensitivity check does not establish uniform δ-correctness of the Gaussian stopping rule under heavy-tailed noise.

Synthetic fixed-allocation audit over the confidence grid. The corrected instance is $N = 5, M = 4, r = 2$, with $\delta \in \{0.2, 0.1, 0.05, 0.02, 0.01, 0.005, 0.002\}$ and 50 runs per confidence level, hence 350 runs per row. Table 9 reports the posted aggregate quantities. Its normalized costs use the reported practical threshold $\beta_{\mathrm{prac}}(t, \delta) = \log(1/\delta) + 2\log(\log(t + 2) + 1)$, not the conservative oracle threshold (17). Pooled errors across this grid must not be compared against the smallest endpoint δ. DS-Active errors at cost ratio 1 occur at $\delta = 0.1$ (two) and 0.005 (one); at ratio 10, one occurs at each of $0.2, 0.1, 0.002$. Additional 500-run checks give 2/500 and 0/500 errors at cost ratio 1 for $\delta = 0.005, 0.002$, and 1/500 at each level for cost ratio 10. No reported per-level binomial test

rejects, but the original 50-run cells have limited power. On 10 other independent instances (1,800 runs), the reported ordering holds on all 10, the median runtime normalization is 0.98, and there are 2/1,800 errors.

Table 9: Synthetic fixed-allocation confidence-grid audit. Error counts pool seven different confidence levels. The normalized-cost columns reproduce the reported aggregate summaries.

Allocation	c_h/c_J	T^*	$\mathbb{E}C/\log(1/\delta)$	$\mathbb{E}C/(T^*\beta_{\mathrm{prac}})$	Pooled errors
Uniform	1	119.8	213.5	1.06	0/350
Uniform	10	658.9	936.7	0.85	0/350
DS-Active	1	1,328.1	2,216.2	0.95	3/350
DS-Active	10	2,442.9	3,402.0	0.79	3/350
Heuristic anchor	1	103.6	209.8	1.21	1/350
Heuristic anchor	10	365.9	670.8	1.08	0/350

Frozen-instance certification audit. The $N = 12, M = 10$ audit uses cost ratios 30 and 100, $\delta \in \{0.1, 0.02\}$, and 100 runs per cell (2,000 total). Table 10 reports measured cost ratios to P-T&S at ratio 30; pooled errors and timeouts span all four cost/confidence cells. At cost ratio 100, analytical ratios are 2.50 (Estimate-and-correct), 8.2 (Heuristic anchor), and 173.5 (Pure-human), while measured ratios at $\delta = 0.1/0.02$ are $2.40/1.65$, $18.4/16.3$, and $154.9/161.0$, respectively. The analytical objective understates Heuristic-anchor cost and overstates Estimate-and-correct cost at tighter confidence; both departures limit a literal finite-sample reading of T^* ratios.

Table 10: Frozen-instance audit at $c_h/c_J = 30$. Errors and timeouts pool four cells (two cost ratios and two confidence levels).

Allocation	T^*	Measured ratio $(0.1/0.02)$	Analytical ratio	Errors	Timeouts
P-T&S	947.3	$1/1$	1	0/400	0
DS-Active	1,041.8	$1.05/1.07$	1.10	0/400	0
Estimate-and-correct	5,170.9	$5.44/4.58$	5.46	5/400	15
Heuristic anchor	20,394.1	$41.4/39.3$	21.5	2/400	88
Pure-human	140,612.8	$151.1/156.6$	148.4	0/400	0

Heuristic-anchor cells have 17–29% timeouts at a cap of 150,000 pulls. The terminal estimate and capped cost are included, including in the error counts; the observed cost ratios are lower bounds on uncapped certification cost. Aggregate counts do not establish per-cell finite-confidence correctness. This audit tests reported fixed allocations, separately from the adaptive fixed-budget comparisons above.

D.6 Baseline allocation adaptations

The submitted analytical comparisons normalize prescribed pull-rate vectors to $\sum_a c_a \omega_a = 1$. Uniform uses equal pull rates on both channels, $\omega_{J_i} = \omega_{h_i} = 1/[N(c_J + c_h)]$; this gives equal rates, not equal cost, when channel costs differ. Pure-human uses $\omega_{J_i} = 0, \omega_{h_i} = 1/(Nc_h)$. Independent-proxy uses boundary judge weights without human anchors. Estimate-and-correct selects top-k candidates as anchors, assigning 30% of cost to human labels and 70% uniformly to judges. Heuristic anchor uses the largest feature norms rather than boundary-aware coverage, together with boundary judge weights and the geometric-mean split. The MF-BAI adaptation treats judges as cheap bounded-bias proxies and places human effort on top candidates. DS-Active is a reliability-based allocation inspired by crowdsourcing [4]: inverse marginal judge variances weight boundary observations, with anchors in top-k∪bottom-k. PPI-inspired uses residual-variance-weighted boundary allocation after the rank-r factor fit [1]. These allocation adaptations do not inherit the original methods' guarantees in the correlated unknown-bias model. Human-only oracle denotes the separate oracle human-channel comparator reported in the recovery experiments. Its detailed per-arm implementation is not inferred from the original Uniform/Pure-human definitions.

D.7 Frozen multi-judge benchmark

The Arena-Hard-v2.0 judge-score collection contains $N = 12$ arms, $M = 10$ judges, and $P = 200$ prompts. The ten judge aliases are Claude Haiku 4.5, Claude Sonnet 4.5, GPT-5-mini, GPT-5.4-mini, Qwen3-8B, Qwen3-14B, DeepSeek-V3.1-Terminus, GLM-4.6, Moonshot Kimi-K2, and Tencent Hunyuan-A13B. They score arm–prompt pairs on a 1–10 rubric. The collection contains $23{,}545$ scores, or 98.1% of the $24{,}000$ planned entries, with a reported parsing rate of at least 99.5%.

For the original benchmark-calibrated diagnostics, the quality target is benchmark-derived and the human channel is simulated as $Y_i \sim \mathcal{N}(\mu_i, 1)$. Thus the real component is the judge-score tensor; these diagnostics do not test the human-channel assumption. The submission's centered judge-score fit has rank-1 and rank-2 explained-variance fractions of 90.5% and 98.4%, respectively. After the rank-2 fit, mean residual correlation is 0.31 within the Claude, OpenAI, and Qwen families and -0.08 between families. These are descriptive fit statistics. A low-rank matrix cannot distinguish quality from a bias component shared by every judge, and therefore does not establish that the learned subspace is an identifiable nuisance subspace.

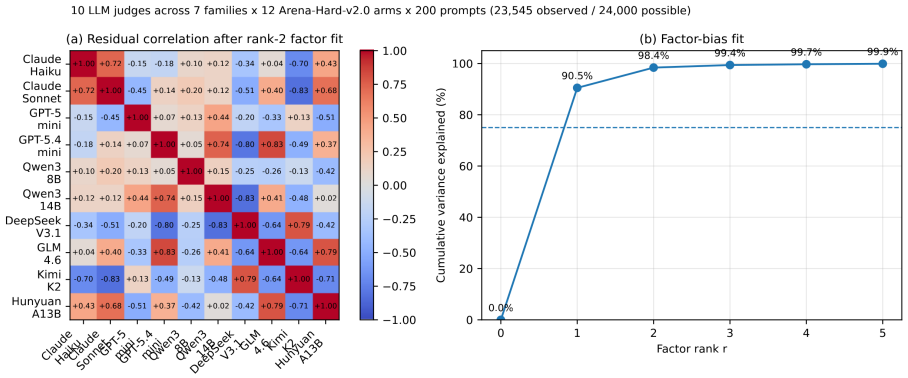

Figure 2: Descriptive residual fit on the frozen 10-judge score collection. Left: residual correlations after a rank-2 fit. Right: cumulative variance explained by SVD relative to the benchmark-derived target. Low rank supports a compact approximation; it does not resolve the quality–bias gauge.

D.8 Analytical allocation diagnostics

Table 11 evaluates each prescribed allocation inside the same fitted Gaussian $T^*(\boldsymbol{\omega})$ objective. It is an allocation diagnostic, not an independently measured recovery rate or stopping cost. The P-T&S allocation uses the boundary-aware design, the within-anchor split, and boundary judge weights; its favorable objective value is consistent with that construction. The difference from DS-Active is substantially smaller than the large gaps obtained on the synthetic instance in Table 12.

Table 11: Analytical $T^*(\boldsymbol{\omega})$ on the submitted frozen 10-judge instance ($N = 12$, top-5, simulated $\sigma_h = 1$). These are fitted-objective values rather than empirical stopping costs.

c_h/c_J	1	3	10	30	100
Uniform	77.8	155.2	423.9	1,171.8	3,590.9
Pure-human	4,687.1	14,061.3	46,870.9	140,612.8	468,709.4
Independent-proxy	∞	∞	∞	∞	∞
Estimate-and-correct	815.1	1,902.3	3,666.8	5,170.9	6,759.7
Heuristic anchor	7,917.7	12,266.6	17,206.6	20,394.1	22,170.9
MF-BAI adaptation	359.1	888.7	2,259.9	4,424.8	8,030.9
DS-Active	75.3	143.6	380.5	1,041.8	3,195.5
PPI adaptation	77.5	194.9	595.3	1,660.3	4,731.5
P-T&S allocation	60.2	134.9	363.1	947.3	2,701.1

D.9 Synthetic diagnostics from the submission

The submitted synthetic suite uses $N \in \{5, 8\}$, $M \in \{4, 6\}$, $r = 2$, $\sigma_h = 1$, and equicorrelated covariance $\mathbf{\Sigma} = (1 - \rho)\boldsymbol{I} + \rho\mathbf{1}\mathbf{1}^\top$. Arm means are equispaced with boundary gap $g \in \{0.4, 0.6\}$; features and judge weights are sampled from standard Gaussian distributions at fixed seeds. The rate illustration uses $N = 5$, $M = 4$, $\rho = 0.3$, 25 runs per condition, cost ratios $\{1, 10, 50\}$, and $\delta \in \{0.2, 0.1, 0.05, 0.02, 0.01\}$, with a cap of 15,000 rounds. Its practical implementation uses greedy boundary-aware anchors and a heuristic threshold $\log(1/\delta) + 2\log(\log(t + 2) + 1)$. The finite collection of runs and this heuristic threshold do not establish the finite-confidence guarantee of the theoretical algorithm.

The other diagnostics evaluate the closed-form objective directly. The leverage sweep enumerates all size-r anchor sets on an $N = 5, M = 4$ instance. The split sweep varies the human-to-judge ratio while holding other allocation components fixed, so it need not attain the unrestricted joint optimum. The anchor sweep uses $N = 8, M = 6, k = 3$ and compares a greedy boundary-aware design, 19 random selections, and an adversarial selection. This one-instance comparison gives objective values of approximately 106, 196 (random median), and 7,811, respectively. The 30-instance distribution reported above supplies a broader summary.

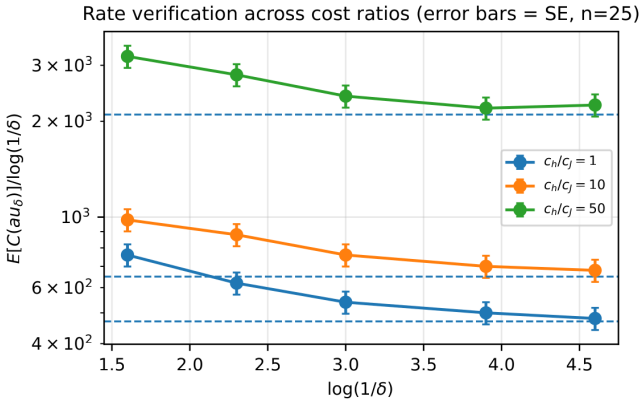

Figure 3: Synthetic rate illustration from the submitted paper: normalized cost across the reported confidence levels. Error bars are one standard error over 25 runs. The practical threshold and finite grid do not establish the formal finite-confidence theorem.

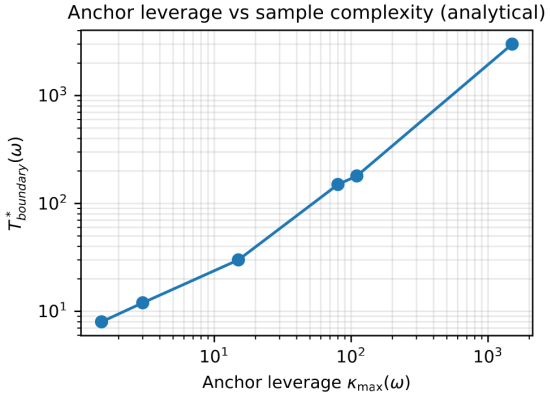

Figure 4: Submitted analytical anchor-leverage sweep across size-r anchor sets on a fixed $N = 5, M = 4, r = 2$ instance.

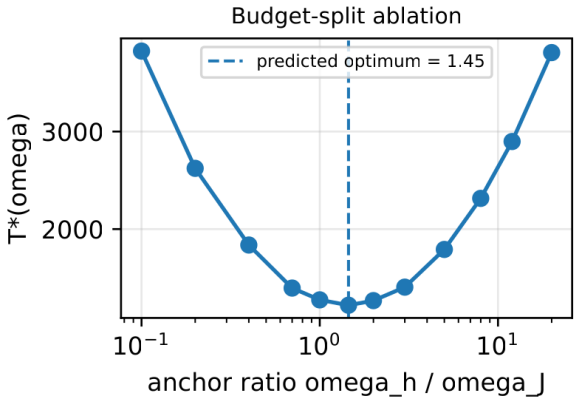

Figure 5: Submitted within-anchor split sweep, with other allocation components constrained. The dashed prediction concerns the calibration split, not an unrestricted closed-form global allocation.

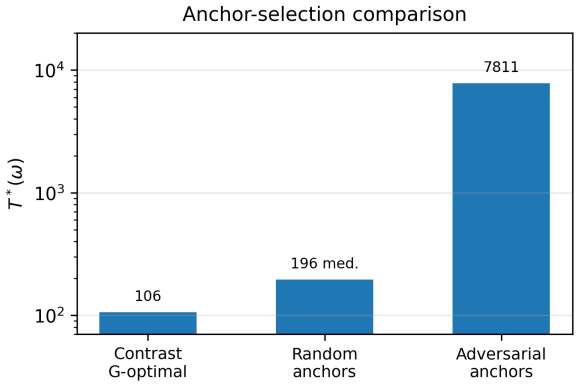

Figure 6: Submitted single-instance analytical anchor-selection comparison. The broader 30-instance distribution is reported in Appendix D.5; this chart is an illustrative instance rather than a typical empirical speedup.

Additional small-roster analytical comparison. The submitted appendix also fits a separate $N = 6$ MT-Bench instance using 5,755 pairwise votes and five proxies selected from 61 raters (60 human experts and GPT-4). Its BT–Elo target and rank-2 residual fit (76.4% explained variance) yield analytical allocation comparisons. At $c_h/c_J = 100$, the reported P-T&S objective is lower than Uniform by a factor of 7 and Pure-human by a factor of 240. These are analytical certification constants on that fitted instance, not observed reductions in human labels or fixed-budget recovery costs. This comparison is distinct from the disjoint-vote A1 construction above and does not test the unbiased-anchor premise.

Semi-synthetic scale illustration. The original Arena-inspired synthetic instance has $N = 15, M = 8, r = 3$, $\rho = 0.4, \sigma_h = 0.3$, and top-5. Its means, features, and human channel are synthetic; the result is an objective-level illustration calibrated to evaluation-like statistics, rather than a new human-vote experiment.

Feature sensitivity. The original feature perturbation sweep adds independent Gaussian noise to the features and evaluates the fitted-objective ratio to the oracle allocation. It illustrates sensitivity to feature perturbation on a fixed instance; it does not establish that SVD learns the correct nuisance coordinates, or that the finite-confidence theorem applies after estimation.

Table 12: Submitted synthetic-instance analytical $T^*(\boldsymbol{\omega})$ ($N = 8, M = 6, r = 2, \rho = 0.3, \sigma_h = 1$, top-3). ∞ denotes an uncovered boundary direction in this instance.

c_h/c_J	1	3	10	30	100
Uniform	75.0	149.9	412.3	1,161.8	3,785.3
Pure-human	88.9	266.7	888.9	2,666.7	8,888.9
Independent-proxy	∞	∞	∞	∞	∞
Estimate-and-correct	174.7	255.4	537.6	1,344.0	4,166.3
Heuristic anchor	120.8	191.0	362.9	737.9	1,816.1
MF-BAI adaptation	139.4	226.2	530.1	1,398.1	4,436.4
DS-Active	66.2	73.9	101.0	178.4	449.4
P-T&S allocation	29.5	35.3	46.8	66.3	132.8

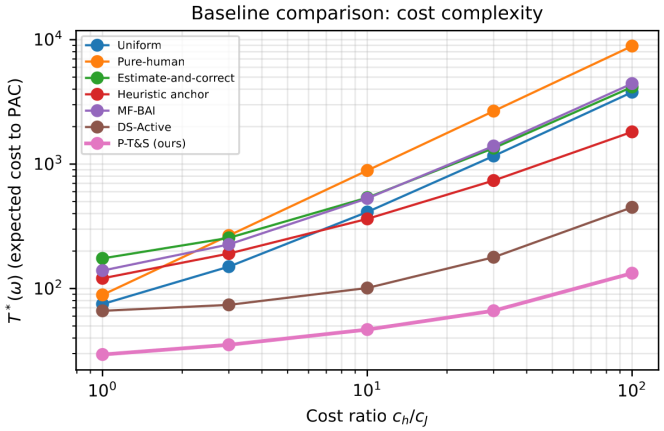

Figure 7: The submission's synthetic analytical allocation comparison (Table 12). The plotted objective is not an observed PAC stopping cost; its allocation rankings are a structural diagnostic.

D.10 Upstream asset provenance

The submitted study cites three upstream resources: Arena-Hard-Auto [15], released under the Apache License 2.0; MT-Bench human judgments [19], under Creative Commons Attribution 4.0 International (CC BY 4.0); and Arena Human Preference 55K [3], under the Apache License 2.0. The MT-Bench resource contains 3,355 human judgments and 2,400 GPT-4 pairwise judgments; the 5,755 total must not be interpreted as all-human annotations. The original submission uses Arena-Hard responses for the frozen judge instance, MT-Bench for the smaller analytical comparison, and Arena preferences to calibrate the semi-synthetic instance. The later real-vote construction and its disjoint roles are described above. The license labels refer to the upstream dataset metadata, without asserting a separate privacy audit of all response and conversation content.

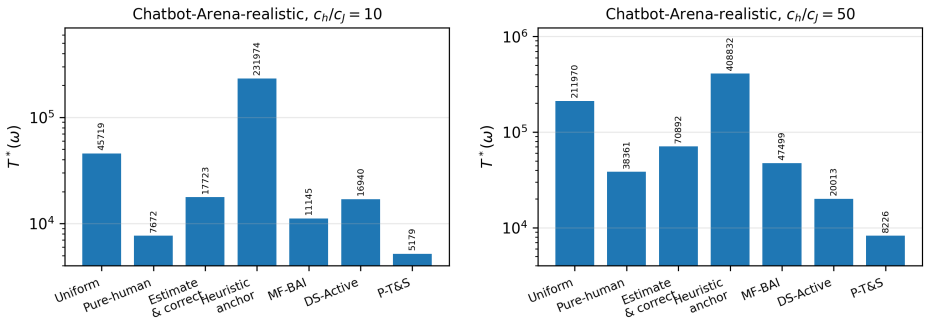

Figure 8: Submitted Arena-inspired semi-synthetic allocation diagnostic at cost ratios 10 and 50.

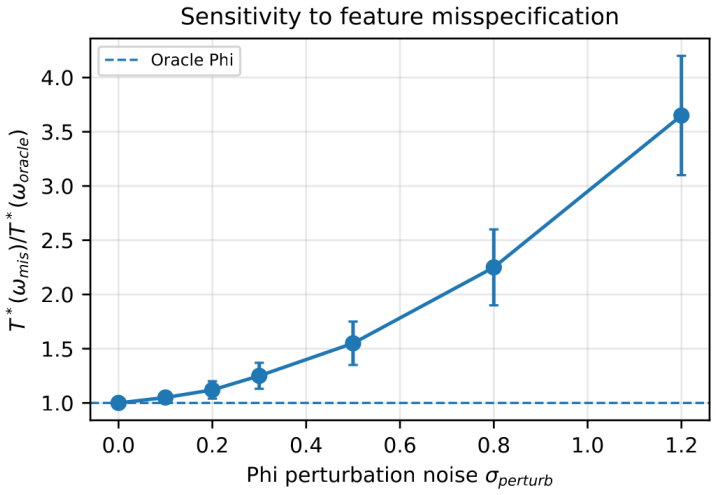

Figure 9: Submitted feature-perturbation sensitivity diagnostic. Error bars are one standard deviation across 20 random perturbations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract separates the oracle Gaussian guarantees, simulated-anchor evaluation, and real-vote experiments. Sections 8 and 9 report overlapping intervals, extrapolations, and calibration limits.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 9 discusses estimated nuisance quantities, common-mode bias, non-rejection and test power, stationary model versions, asynchronous labels, heavy tails, and finite vote pools.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: Sections 3–7 state the assumptions and Appendices A–C give the profiling, convexity, and cost-indexed concentration arguments. The conservative stopping threshold now has a self-contained Gaussian correctness proof. Author verification of the oracle-target tracking argument and unconditional plug-in expectation argument remains required for a complete-proof declaration.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The manuscript records the experimental designs and numerical results reported during review. A complete archive of the additional vote splits, acquisition records, inference settings, and adaptive implementation is needed for full reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: A final public repository and complete additional experimental artifacts have not been supplied for this revision. The public upstream benchmarks are cited, but that alone is insufficient to reproduce the added experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Section 8 and Appendix D specify available instance sizes, budgets, vote splits, and run counts. Exact additional split identifiers, inference settings, and experiment scripts remain to be provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Recovery tables report the 95% intervals provided in the author responses; analytical ratios have a paired prompt bootstrap. Bias tests include non-rejections, injected-bias power curves, and one-sided amplitude bounds. Intervals and extrapolation flags are retained without interpreting non-rejection as proof.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Complete compute and wall-clock records for the additional review-stage experiments must be supplied by the authors to document resource requirements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The evaluation uses cited public benchmarks and model-judge observations. No new participant recruitment is described. The intended release must respect upstream licenses and protect any personal information present in response corpora.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 9 discusses reduced annotation cost and the risks of biased targets, missing feature directions, undetected model drift, and misleading leaderboard decisions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work introduces an allocation method rather than a new foundation model. Dataset handling and release must nevertheless follow the upstream access and privacy conditions.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix D credits the upstream datasets and states their licenses: Arena-Hard-Auto and Arena Human Preference 55K use Apache License 2.0; MT-Bench Human Judgments uses CC BY 4.0. Redistribution of the final artifact archive must retain attribution and respect these upstream conditions.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The frozen judge-score tensor and additional analysis are research artifacts. Their complete documented release, including vote splits and acquisition metadata for the added experiments, remains to be supplied.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The reported human-anchor experiments reuse existing public votes and do not describe new crowdsourcing or participant recruitment. The original data collection is documented by the cited benchmark releases.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human-subject recruitment is described. The authors should verify the institutional requirements that apply to reuse of the released vote datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM judges generate the empirical score channel. The paper describes their evaluation role. LLM assistance was also used in preparing and checking this camera-ready revision; the authors remain responsible for verifying the mathematics, experimental provenance, and final claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.