SURROGATE-AUGMENTED LEE–VALIANT ESTIMATION FOR ADAPTIVE BENCHMARK EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating modern models on large benchmarks is costly. We introduce the *Adaptive Surrogate-Augmented Lee–Valiant estimator* (ASA-LV), which combines historical-response prediction, adaptive querying, propensity-corrected residual contributions, and robust Lee–Valiant aggregation. We establish finite-sample error bounds under adaptive querying and reuse of queried answers for surrogate fitting. The bounds account for surrogate-class complexity and covering approximation error, yielding a $\widetilde{O}(k^{-1/2})$ estimation-error rate under the stated assumptions. We evaluate ASA-LV on a collection of 19 benchmarks and in the FAQ evaluation setting comprising MMLU-Pro and a composite benchmark. At query budgets of 5% and 10%, ASA-LV achieves the lowest root-mean-square error (RMSE) and mean absolute error (MAE), together with the highest Kendall and Spearman model-ranking correlations, among the compared estimators in both settings. Relative to Factorized Active Querying (FAQ), the RMSE reductions are 10.517%/6.468% on the 19-benchmark collection and 14.927%/10.781% on the FAQ data, respectively. On the FAQ data, ASA-LV also achieves lower RMSE than FAQ, two augmented inverse-probability-weighted (AIPW) variants and random sampling when 5% or 10% of target answers are flipped before querying, under both scoring targets and both query budgets.

1 INTRODUCTION

Large benchmark evaluations require many expensive model queries. A historical score matrix can predict an unevaluated model's answers, guide which questions to query, and support correction of prediction errors using the queried sample. Such historical information creates an opportunity to estimate benchmark accuracy accurately from only a small fraction of questions.

The statistical problem is that adaptive querying can assign small probability to some questions, so an observed prediction residual may be greatly amplified by propensity correction even though every answer is bounded. Moreover, the queried answers are valuable for updating the surrogate itself. Reusing them to fit a terminal surrogate makes the final contributions data dependent, so the usual predictable martingale argument cannot be applied directly. A useful evaluator must therefore combine prediction, adaptive sampling and robust estimation without losing finite-sample control.

Active statistical inference and Factorized Active Querying (FAQ) provide practical instances of prediction-guided sampling and correction (Zrnic & Candes, 2024; Wu et al., 2026). Building on this strategy, we study robust aggregation of surrogate-augmented contributions while allowing the queried answers to update the final surrogate. In this paper, we introduce the *Adaptive Surrogate-Augmented Lee–Valiant estimator* (ASA-LV). Historical responses provide an initial surrogate and adaptive query policy; the revealed answers can update the surrogate; final predictions and logged propensities then form augmented full-bank contributions; and Lee–Valiant (Lee & Valiant, 2022) aggregation limits the influence of propensity-amplified contributions. The construction accommodates fixed historical means, item-response theory (IRT), released factor surrogates, and a logistic-regression surrogate learned from historical features and queried labels.

Our contributions connect estimator design, finite-sample guarantees, and empirical improvements in accuracy estimation and model ranking. First, we develop ASA-LV to combine adaptive querying and surrogate updating with robust aggregation of inverse-probability-corrected contributions. Sec-

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

ond, we establish explicit finite-sample error bounds for ASA-LV under adaptive querying, including when the queried answers are reused to fit the final surrogate. The analysis controls the resulting data dependence without requiring a separate sample for surrogate fitting and quantifies the costs of surrogate-class complexity and covering approximation error. For the finite-dimensional surrogate classes considered here, the bounds yield a $\widetilde{O}(k^{-1/2})$ estimation-error rate, under the stated assumptions, where k is the query budget. Third, we evaluate ASA-LV in two settings: a collection of 19 benchmarks covering diverse tasks, and the FAQ evaluation setting, comprising MMLU-Pro and a composite benchmark with over 2,200 target models each. In both settings, ASA-LV achieves the lowest RMSE and MAE and the highest Kendall and Spearman model-ranking correlations among the compared estimators at both 5% and 10% query budgets. On the FAQ data, it also achieves lower RMSE than FAQ, both evaluated AIPW variants and random sampling across both answer-corruption rates, both scoring targets and both query budgets.

2 Related work

Active Testing studies label-efficient model evaluation through adaptive acquisition and importance-weighted correction (Kossen et al., 2021). Model-assisted generalized-difference estimation fits predictors on sampled observations (Dagdoug et al., 2023), while adaptive importance sampling provides martingale-based analyses (Imberg et al., 2025). Adaptive importance sampling with control variates (AISCV) admits nonasymptotic guarantees for weighted least-squares control-variate estimation (Leluc et al., 2022). Our analysis covers bounded surrogate classes beyond linear control-variate regression and nonlinear Lee–Valiant aggregation, including when queried answers are reused to fit the final surrogate. Prediction-powered inference (PPI++) combines predictions with observed-label corrections to improve statistical efficiency (Angelopoulos et al., 2023), while Hadad et al. (2021) develop adaptively weighted estimators with asymptotically normal test statistics for adaptive experiments. Historical item-response theory (IRT) appears in tinyBenchmarks and Fluid Benchmarking (Polo et al., 2024; Hofmann et al., 2025). FAQ combines prediction-guided querying with Proactive Active Inference (PAI) to estimate accuracy on a fixed benchmark (Wu et al., 2026). Its confidence intervals rely on asymptotic normality, whereas our analysis provides explicit nonasymptotic error bounds for ASA-LV, including when queried answers are reused to fit the final surrogate. TinyBenchmarks and Fluid provide surrogate-design precedents through their IRT and computerized-adaptive-testing pipelines, with their own fits, banks and acquisition protocols.

Lee & Valiant (2022) construct a finite-variance estimator for independent and identically distributed (iid) observations using a pilot and calibrated truncation; Lee et al. (2025) establish further iid robustness and moment properties. We retain their aggregation and supply the bridge needed for adaptive augmented contributions. Martingale Catoni methods give known-moment confidence sequences (Wang & Ramdas, 2023); Whitehouse et al. (2024) analyze centered truncation under martingale dependence; and Wagenmaker et al. (2022) develop cumulative-variance and covering tools for adaptive data-dependent functions. Those tools do not directly control LV's dependent chronological pilot or the change in its global calibration after terminal refitting. Our new results combine reset-supermartingale control of that pilot with a uniform terminal-class event, deterministic calibration mesh and Lipschitz stability of the calibrated aggregate; the concentration, covering machinery and LV estimator themselves remain prior tools. The variance-sensitive result uses an almost-sure cumulative-variance envelope and is a fixed-budget oracle guarantee, not an observable interval.

Zhang et al. (2026) systematically compare benchmark-prediction methods across 19 benchmarks and identify random sampling combined with regression as a strong baseline. Related approaches address robustness through importance weighting and sampling design: Su et al. (2020) shrink importance weights to minimize a bound on mean-squared error in doubly robust off-policy evaluation, while Li et al. (2026) combine uniform and uncertainty-guided sampling to protect active statistical inference against inaccurate uncertainty estimates. ASA-LV combines prediction and residual correction with adaptive querying and Lee–Valiant aggregation, while allowing the queried answers to update the final surrogate.

3 PROBLEM FORMULATION

We use uppercase letters for random objects and lowercase letters for fixed quantities or realized values. In particular, let $y_1, \ldots, y_n \in \{0, 1\}$ denote a target model's fixed correctness indicators on a benchmark of n questions: $y_i = 1$ if the model answers question i correctly and $y_i = 0$ otherwise. Its full-benchmark accuracy is $\mu = n^{-1} \sum_{i=1}^{n} y_i$. Our goal is to estimate μ using a fixed budget of k answer queries, typically with $k \ll n$, with historical model responses available to guide prediction and adaptive sampling. Here k is chosen before querying begins.

At each round $t = 1, \ldots, k$, the evaluator chooses a question, observes the target model's answer correctness, and may use this observation to update its predictions and subsequent query choices. Let $\mathcal{H}_0$ denote the information available before querying, including historical data, initialization, and fixed algorithm settings. Let $\mathcal{H}_{t-1}$ denote the information available immediately before query t, including all previous query indices and revealed answers, together with any algorithmic randomness used to construct the next query distribution.

Let $U_t \subseteq \{1, \ldots, n\}$ be the set of questions eligible for selection at round t. For each question i, let $Q_t(i)$ be its probability of being selected at that round. Let $F_{t-1}(i) \in [0, 1]$ be the current surrogate prediction of the target model's probability of answering question i correctly. The eligible set U_t, query distribution Q_t, and predictor F_{t-1} are determined from the information available before the draw; formally, they are $\mathcal{H}_{t-1}$-measurable. Let I_t denote the index of the question selected at round t. We then have

$$Q_t(i) = 0 \quad \text{for } i \notin U_t, \qquad \sum_{i \in U_t} Q_t(i) = 1, \qquad \Pr(I_t = i \mid \mathcal{H}_{t-1}) = Q_t(i).$$

The observed correctness is the random variable $Y_t := y_{I_t}$. We require $Q_t(i) \geq \frac{\tau}{n}$ for every $i \in U_t$, where $\tau \in (0, 1]$ is a fixed exploration parameter. This condition ensures that every eligible question retains a positive chance of being queried.

The query budget k denotes the total number of draws, indexed by $t = 1, \ldots, k$. With replacement (WR), every question remains eligible, so $U_t = \{1, \ldots, n\}$, and repeated queries count against the budget. Without replacement (WOR), each question can be queried at most once, so $U_t = \{1, \ldots, n\} \setminus \{I_1, \ldots, I_{t-1}\}$. Lowercase i_t, q_t, u_t denote realized values of I_t, Q_t, U_t, respectively.

The initial predictor F_0 is constructed from historical data. After observing (I_t, Y_t), the procedure updates the predictor to F_t according to the specified fitting schedule; if no update is scheduled, $F_t = F_{t-1}$. The next query distribution is constructed from the updated history.

4 OUR ASA-LV ESTIMATOR

ASA-LV combines surrogate predictions with propensity-corrected query observations and applies Lee–Valiant aggregation to the resulting contributions.

Surrogate-augmented contributions. For a predictor $f : \{1, \ldots, n\} \to [0, 1]$, where $f(i)$ predicts the target model's probability of answering question i correctly, define the baseline $B_t(f)$ and the augmented contribution $Z_t(f)$ at query round t,

$$B_t(f) = \frac{1}{n} \left[\sum_{i \notin U_t} y_i + \sum_{i \in U_t} f(i) \right], \qquad Z_t(f) = B_t(f) + \frac{Y_t - f(I_t)}{n Q_t(I_t)}. \tag{1}$$

With replacement, $B_t(f)$ is the mean prediction over the entire benchmark. Without replacement, $B_t(f)$ combines the known answers to previously queried questions with predictions for the remaining questions. The second term in $Z_t(f)$ corrects the prediction error on the selected question, weighted by the inverse of its query probability.

For every $t = 1, \ldots, k$, the inverse-probability correction ensures that $\mathbb{E}[Z_t(f) \mid \mathcal{H}_{t-1}] = \mu$ for a fixed predictor f. Because the query-time predictor F_{t-1} is determined by the information in $\mathcal{H}_{t-1}$, it also satisfies $\mathbb{E}[Z_t(F_{t-1}) \mid \mathcal{H}_{t-1}] = \mu$.

After all k queries, let F denote the final surrogate obtained under the specified fitting rule. This rule may retain the last query-time predictor or fit a new predictor using the accumulated queried answers.

Algorithm 1 Adaptive Surrogate-Augmented Lee–Valiant estimator (ASA-LV)

Require: Historical data, budget k, block count $1 \le b \le k$, exploration floor $0 < \tau \le 1$, and calibration mass $0 \le c < k$.

1: Initialize the query-time predictor F_0 from historical data.
2: **for** $t = 1, \dots, k$ **do**
3: Construct Q_t from the available history using the chosen query policy, satisfying the eligibility and exploration conditions in Section 3.
4: Draw $I_t \sim Q_t$, observe $Y_t = y_{I_t}$, and log the realized query index i_t, query distribution q_t, eligible set u_t, and observed correctness y_{i_t}.
5: Update the predictor to F_t according to the specified fitting schedule; otherwise retain $F_t = F_{t-1}$.
6: **end for**
7: Obtain the realized final surrogate f under the specified fitting rule.
8: Compute z_t for every $t = 1, \dots, k$ using Eq. (1), f, and the logged observations, eligible sets, and query probabilities.
9: Partition $z_1, \dots, z_k$ in query order into b consecutive blocks whose sizes differ by at most one. Let p be the median of their means.
10: Set $d_t(p) = \min\{\gamma(z_t - p)^2, 1\}$, choosing $\gamma \ge 0$ so that $\sum_{t=1}^{k} d_t(p) = c$. If calibration is unattainable, set $d_t(p) = 1$ for $z_t \ne p$ and $d_t(p) = 0$ otherwise.
11: **return** $T_c(z, p)$ from Eq. (2).

We use this single final surrogate to reconstruct all k contributions, $Z_t = Z_t(F)$ for $t = 1, \dots, k$, retaining the original logged query probabilities Q_t. Let $Z = (Z_1, \dots, Z_k)$ and $\bar{Z} = k^{-1} \sum_{t=1}^{k} Z_t$, and write $f, z_t, \bar{z}$ for their realized values. Because F can depend on Y_t and later queried answers, the terminal contributions need not satisfy $\mathbb{E}[Z_t \mid \mathcal{H}_{t-1}] = \mu$.

Lee–Valiant aggregation. ASA-LV applies the pilot–calibration construction of Lee & Valiant (2022) to these surrogate-augmented contributions. Given the realized values $z = (z_1, \dots, z_k)$ of the final surrogate-augmented contributions $Z_t = Z_t(F)$ and an integer $1 \le b \le k$, partition $z_1, \dots, z_k$ in query order into b consecutive blocks whose sizes differ by at most one. Let p be the median of the block means. This preliminary estimate p serves as the center for robust aggregation of the full contributions z_t.

Fix a calibration mass $0 \le c < k$, which controls the total amount of downweighting. Define $d_t(p) = \min\{\gamma(z_t - p)^2, 1\}$, where $\gamma \ge 0$ is chosen so that $\sum_{t=1}^{k} d_t(p) = c$. If this equality is unattainable, set $d_t(p) = 1$ when $z_t \ne p$ and $d_t(p) = 0$ otherwise; the resulting sum is at most c. ASA-LV returns

$$T_c(z, p) = p + \frac{1}{k} \sum_{t=1}^{k} (z_t - p)\{1 - d_t(p)\}. \tag{2}$$

The factor $1 - d_t(p)$ downweights contributions that deviate strongly from the preliminary estimate p. The complete algorithm is given in Algorithm 1.

5 Theoretical guarantees

We establish finite-sample error bounds for ASA-LV under adaptive querying and reuse of queried answers for surrogate fitting. Theorem 1 directly covers the fixed calibration mass used in our experiments, with explicit surrogate-class complexity and covering approximation terms. Theorem 2 analyzes the same aggregation rule with the calibration mass and block count chosen as functions of a prescribed failure probability α. All probability statements are conditional on the pre-query historical data and initialization.

Theorem 1. *Fix $0 < \alpha < 1$. Let the random final surrogate F belong to a predetermined class $\mathcal{F} \subset [0,1]^n$, and let $\mathcal{G} \subset \mathcal{F}$ be an ϵ-net in the sup norm, chosen before querying, with $j = |\mathcal{G}| \ge 1$. The class and covering net may depend on the fixed query budget k. Write $Z_t = Z_t(F)$ and $Z = (Z_1, \dots, Z_k)$. Query probabilities are predictable and satisfy $Q_t(i) \ge \tau/n$ for every $i \in U_t$, where $0 < \tau \le 1$.*

Let $P \in [\min_t Z_t, \max_t Z_t]$ be a random pilot and $0 \le c < k$ a fixed calibration mass. In exact arithmetic, let $D_t(P) = \min\{\Gamma(P)(Z_t - P)^2, 1\}$ for a random $\Gamma(P) \ge 0$, with $\sum_{t=1}^k D_t(P) \le c$. Set $m = 1 + \lceil \log_2 \max(1, k/\tau) \rceil$, $\ell = \log\{2m(j+1)/\alpha\}$. Let G_F be a nearest element of $\mathcal{G}$ to F, with ties resolved by a fixed rule. With probability at least $1 - \alpha$,

$$|T_c(Z, P) - \mu| \le \frac{2\sqrt{\max(V_{G_F}, 1)\ell}}{k} + \frac{4\ell}{3\tau k} + \left(1 + \frac{1}{\tau}\right)\epsilon + \frac{cr_\tau}{k}. \tag{3}$$

For each fixed $g \in \mathcal{G}$, the cumulative conditional variance is $V_g = \sum_{t=1}^k \mathrm{Var}(Z_t(g) \mid \mathcal{H}_{t-1}) \le \frac{k}{\tau}$. The term V_{G_F} evaluates this quantity at the selected net element. The deterministic range bound is $r_\tau = 2/\tau$ for WR and $r_\tau = 1 + 2/\tau$ for WOR.

A second bound avoids the explicit variance term. Let $\ell_0 = \log\{2(j+1)/\alpha\}$. With probability at least $1 - \alpha$,

$$|T_c(Z, P) - \mu| \le 2\sqrt{\frac{\ell_0}{\tau k}} + \frac{4\ell_0}{3\tau k} + \left(1 + \frac{1}{\tau}\right)\epsilon + \frac{cr_\tau}{k}. \tag{4}$$

Consequently, for fixed α, if τ is bounded away from zero, c and j are fixed, and $\epsilon = O(k^{-1/2})$, then $|T_c(Z, P) - \mu| = O(k^{-1/2})$ with probability at least $1 - \alpha$. More generally, when $j = j_k$ and $\epsilon = \epsilon_k$, the bound is $O(\sqrt{\log(j_k + 1)/k} + \log(j_k + 1)/k + \epsilon_k)$ under the same conditions on α, τ, and c.

The proof first bounds the error of the contribution mean $\bar{Z} = k^{-1}\sum_{t=1}^k Z_t$ uniformly over the surrogate class, then controls the change introduced by LV aggregation. For a realized contribution vector z and a pilot $p \in [\min_t z_t, \max_t z_t]$,

$$T_c(z, p) - \bar{z} = -\frac{1}{k}\sum_{t=1}^k (z_t - p) d_t(p), \qquad |T_c(z, p) - \bar{z}| \le \frac{c}{k}\,\mathrm{range}(z), \tag{5}$$

where $\mathrm{range}(z) = \max_t z_t - \min_t z_t$. Combining this deterministic bound with the uniform concentration bound for $|\bar{Z} - \mu|$ gives the stated result. The b-block pilot lies in this range, so the argument requires no separate concentration bound for the pilot and covers both saturation and unattainable calibration. At fixed c and τ, the displacement contributes an additional $O(k^{-1})$ term to the error bound. Appendix B.1 gives the proof, and Appendix B.4 addresses numerical calibration and arithmetic errors.

Application to the experimental surrogates. The surrogate classes used in our experiments satisfy the covering assumptions under the conditions detailed in Appendix B.2. The ridge-updated IRT predictor remains in a bounded parameter class with $d = 5$, while the released factor update admits a deterministic parameter bound with $d = 64$. Choosing $\epsilon = 1/k$ gives $\log j_k = O(d \log k)$ for these classes. At fixed confidence, exploration floor, and calibration mass, Theorem 1 therefore yields an error bound of order $O(\sqrt{d \log k/k} + d \log k/k + 1/k)$. Logistic regression with fixed historical features also admits a finite-bank cover, and a fixed historical mean corresponds to a singleton class with $j = 1$ and zero covering approximation error.

How the nonlinear correction can help. For the experimental choice $c = \log(10)/3 < 1$ and a nonconstant contribution vector, exact calibration gives $d_t(p) \le c < 1$ for every t, so $\gamma = c/\sum_{t=1}^k (z_t - p)^2$. Consequently, $T_c(z, p) - \bar{z} = -\frac{c}{k}\frac{\sum_{t=1}^k (z_t - p)^3}{\sum_{t=1}^k (z_t - p)^2}$. If all contributions are equal, ASA-LV returns their common value. The numerator measures the imbalance between positive and negative deviations from the pilot p: ASA-LV shifts the contribution mean downward when this quantity is positive and upward when it is negative. This correction can help when propensity-amplified residuals pull the contribution mean away from μ while the pilot remains close to μ. The identity describes the direction of the correction; whether it reduces estimation error depends on the contribution distribution and the accuracy of the pilot.

The next theorem gives an asymptotically sharp error bound when the calibration mass and block count depend on the prescribed failure probability. We first state the result for adapted contributions with a common conditional mean; the extension to a fitted surrogate follows below.

Theorem 2. *Let $W_1, \dots, W_k$ be adapted to $(\mathcal{H}_t)_{t=0}^k$, with $\mathbb{E}[W_t \mid \mathcal{H}_{t-1}] = \mu$ and $\sum_{t=1}^k \mathrm{Var}(W_t \mid \mathcal{H}_{t-1}) \le k\sigma^2$ almost surely for a deterministic $\sigma > 0$. For a failure probability $\alpha = \alpha_k \in (0,1)$, write $\ell = \log(2/\alpha)$ and choose $c = \ell/3$ and $b = \lceil \ell \rceil$. Assume $b \le k$ and $c < k$. Let P_W be the median of the means of b consecutive blocks whose sizes differ by at most one, averaging the two middle values when b is even. Apply the ASA-LV aggregation rule to $W = (W_1, \dots, W_k)$ with this pilot and exact calibration, including the stated fallback when calibration is unattainable. If $\ell \to \infty$ and $\ell/k \to 0$ as $k \to \infty$, then*

$$\Pr\left\{ |T_c(W, P_W) - \mu| \le (1 + o(1))\sigma \sqrt{\frac{2\log(2/\alpha)}{k}} \right\} \ge 1 - \alpha. \tag{6}$$

The $o(1)$ term is deterministic and uniform over processes satisfying these assumptions. If $\sigma = 0$, the estimator equals μ almost surely.

The proof first bounds aggregation centered at μ by $\sigma\sqrt{2\ell/k}\{1 + O((\log \ell)/\ell)\}$, then shows that replacing μ with the block-median pilot adds $O(\sigma\ell/k)$. Both remainder terms are negligible relative to $\sigma\sqrt{\ell/k}$ under the stated conditions. Appendix C proves this result and provides an explicit finite-sample bound.

The same uniform-cover argument used in Theorem 1 extends this result to the fitted surrogate F. Retain its notation $\mathcal{G}$, j, ϵ and G_F, and assume that each $g \in \mathcal{G}$ satisfies $\sum_{t=1}^k \mathrm{Var}(Z_t(g) \mid \mathcal{H}_{t-1}) \le k\sigma_g^2$ almost surely for a deterministic $\sigma_g \ge 0$. Choose $\ell = \log(2j/\alpha)$, $c = \ell/3$ and $b = \lceil \ell \rceil$, with a common deterministic block partition. If $\ell \to \infty$ and $\ell/k \to 0$, applying Theorem 2 at failure probability α/j and controlling the net-approximation error gives $|T_c(Z(F), P_F) - \mu| \le (1 + o(1))\sigma_{G_F}\sqrt{2\ell/k} + (1 + o(1))(1 + 1/\tau)\epsilon$ with probability at least $1 - \alpha$, where P_F is the block-median pilot for $Z(F)$. The leading constant is retained when $(1 + 1/\tau)\epsilon = o(\sigma_{G_F}\sqrt{\ell/k})$.

6 Experiments

Data and prediction task. Each benchmark is a binary model-by-question response matrix. We estimate each held-out model's full-benchmark accuracy using k queries. Historical-model responses are available for surrogate construction, while unqueried target answers remain hidden. We evaluate two public benchmark collections with substantially different historical-data and target-cohort sizes. The collection of Zhang et al. (2026) contains 19 benchmarks spanning reasoning, knowledge, mathematics, vision and language classification, with 51–327 historical models and 19–112 target models per benchmark. The FAQ evaluation data released by Wu et al. (2026) contain MMLU-Pro and a composite benchmark, each with over 2,200 historical models and over 2,200 target models. The collections share six benchmark families: MMLU-Pro, BBH, GPQA, IFEval, MATH and MuSR; the latter five are pooled into FAQ's composite benchmark but evaluated separately in the 19-benchmark collection. The two collections differ substantially in historical-data and target-cohort sizes. We use locally fitted IRT surrogates for the 19-benchmark collection and the authors' released 64-dimensional factors for the FAQ evaluation data, reporting the results separately. Appendix A.1 gives per-benchmark counts, and Appendix A.2 describes historical-data preparation.

Predictors, sampling and budgets. The evaluation uses twenty query repetitions; the released-factor stress tests in Table 4 use ten. We evaluate 5% and 10% query budgets. The local IRT surrogate has four latent coordinates and an intercept, and updates after every four new distinct target answers. Its terminal predictor is the last scheduled iterate. The released 64-dimensional factor model from FAQ uses a per-query Laplace update. A historical-mean surrogate stays fixed; the regression variant fits a new terminal surrogate to the queried answers.

All primary ASA-LV runs use sampling with replacement (WR), so repeated questions count against the budget. Query probabilities are constructed from the current surrogate predictions using the uncertainty score

$$S_t(i) = \sqrt{F_{t-1}(i)\{1 - F_{t-1}(i)\}}, \qquad Q_t(i) = \frac{\tau}{n} + (1 - \tau)\frac{S_t(i)}{\sum_{j=1}^n S_t(j)}. \tag{7}$$

If all scores are zero, we use uniform sampling. The probabilities are recomputed after each scheduled surrogate update. We use $\tau = .01$ for the 19-benchmark collection, .05 for MMLU-Pro, and

Table 1: Clean accuracy-estimation results using all target models and twenty query repetitions. Each entry reports RMSE/MAE, computed within each benchmark and then averaged equally across benchmarks; lower values are better for both metrics. Budgets are percentages of benchmark size.

	19 benchmarks (IRT)		FAQ data (released factors)	
Method	5%	10%	5%	10%
ASA-LV (ours)	**.040810/.030875**	**.028906/.021864**	**.011437/.008848**	**.007694/.005931**
FAQ	.045606/.035212	.030905/.023723	.013443/.010417	.008624/.006675
AIPW-B	.049525/.037421	.031983/.024012	.014720/.011508	.009434/.007359
AIPW-U	.047724/.037413	.031339/.024232	.014894/.011761	.009521/.007500
AISCV regression	.048672/.036759	.031073/.023799	.017135/.013595	.010957/.008669
RS mean	.057601/.045464	.040115/.031699	.019646/.015583	.013536/.010753

.25 for the composite benchmark. For the FAQ evaluation data, FAQ and ASA-LV use the query parameters released by Wu et al. (2026). Random sampling (RS) uses k distinct uniformly sampled questions. Our categorical policies make exactly k draws; Bernoulli baselines target k expected queries.

The experiments fix $b = 3$ and $c = \log(10)/3$, as covered by Theorem 1. Table 6 specifies budget rounding and calibration settings; Appendix A.4 gives surrogate-fitting and sampling details.

Compared methods. FAQ is Factorized Active Querying (Wu et al., 2026). AIPW-B and AIPW-U (Robins et al., 1994) use augmented inverse-probability weighting with Bernoulli sampling: B allocates according to fixed historical uncertainty, whereas U uses uniform allocation. Both correct surrogate predictions using inverse-probability-weighted residuals from queried answers (Appendix A.4). The AISCV-style feature-regression adapter (AISCV regression) fits weighted control-variate regression to fixed historical item features on the ASA-LV query trajectories (Leluc et al., 2022). RS reports the unadjusted mean of its fixed-size uniform sample. These comparisons use the same nominal answer-query budgets.

Evaluation metrics. We compute root-mean-square error (RMSE) and mean absolute error (MAE) over target models and query repetitions within each benchmark, then average these errors equally across benchmarks. To assess model ranking, we compute Kendall τ_b and Spearman ρ between estimated and true model accuracies within each benchmark and repetition, then average over repetitions and equally across benchmarks. Lower RMSE and MAE and higher ranking correlations indicate better performance. We report all four metrics for both clean evaluation settings at the 5% and 10% query budgets.

6.1 Experimental Results

Clean accuracy estimation. Table 1 shows that ASA-LV achieves the lowest RMSE and MAE among the compared estimators in both evaluation settings at both query budgets. Relative to FAQ, ASA-LV reduces RMSE by 10.517% and 6.468% on the 19-benchmark collection and by 14.927% and 10.781% on the FAQ evaluation data at the 5% and 10% budgets, respectively. At the 5% budget, ASA-LV also has lower RMSE than FAQ, AISCV regression and RS mean in every one of the 19 benchmarks (Table 9). ASA-LV also achieves the lowest MAE: relative to FAQ, ASA-LV reduces MAE by 12.3% and 7.8% on the 19-benchmark collection and by 15.1% and 11.2% on the FAQ evaluation data at the 5% and 10% budgets, respectively. The improvements therefore extend to average absolute error, rather than appearing only under the greater penalty that RMSE assigns to large errors. Appendix Table 10 reports RMSE differences and marginal 95% bootstrap intervals; all twenty intervals lie above zero, supporting lower average RMSE for ASA-LV than each of the five external baselines in both settings at both budgets.

Model ranking. Table 2 shows that ASA-LV achieves the highest Kendall τ_b and Spearman ρ scores among the compared estimators in both settings at both query budgets. Its improvements therefore extend from estimating individual model accuracies to recovering their relative ordering.

Comparison with other robust estimators. We replace the final LV aggregation step in ASA-LV with four alternatives: median-of-means (MOM), Winsorized mean (Dixon, 1960), Huber (Huber,

Table 2: Model-ranking correlations using all target models and twenty query repetitions. Kendall τ_b accounts for ties, and Spearman ρ is the correlation of average ranks. Results are averaged over repetitions and equally across benchmarks; higher values are better.

	19 benchmarks (IRT)				FAQ data (released factors)			
	Kendall τ_b		Spearman ρ		Kendall τ_b		Spearman ρ	
Method	5%	10%	5%	10%	5%	10%	5%	10%
ASA-LV	**.768523**	**.821281**	**.890252**	**.926258**	**.922370**	**.945174**	**.991489**	**.995612**
FAQ	.751374	.810262	.875330	.917639	.910969	.939249	.988900	.994664
AIPW-B	.744479	.809590	.870402	.916789	.903525	.934035	.987149	.993804
AIPW-U	.741974	.809117	.869754	.917469	.901242	.932366	.986540	.993479
AISCV regression	.752105	.816354	.877318	.922766	.902076	.929493	.986999	.993046
RS mean	.715675	.777396	.840217	.889446	.876838	.908522	.978593	.987980

Table 3: Comparison with other robust estimators. Entries report equal-benchmark mean RMSE at 5%/10% query budgets; lower values are better.

Method	19 benchmarks	FAQ data
ASA-LV	**.040810/.028906**	**.011437/.007694**
MOM	.046262/.033148	.012997/.008816
Winsor	.041084/.029129	.011602/.007748
Huber	.041084/.029130	.011602/.007748
Catoni	.041084/.029129	.011613/.007753

Table 4: Answer-record corruption on the FAQ benchmarks. Entries report equal-benchmark mean RMSE at 5%/10% query budgets; lower values are better. "Clean" and "altered" refer to the original and corrupted answer means.

	5% corruption		10% corruption	
Method	Clean	Altered	Clean	Altered
ASA-LV	**.023077/.020416**	**.015634/.010701**	**.039201/.037169**	**.018404/.012521**
FAQ	.025478/.021366	.018055/.011780	.041200/.037837	.021018/.013656
AIPW-B	.026331/.021937	.018987/.012477	.041923/.038216	.021819/.014363
AIPW-U	.025129/.021137	.017571/.011296	.040336/.037406	.019405/.012526
RS mean	.026800/.022537	.019974/.013741	.040648/.037884	.020243/.013940

1992) and Catoni (Catoni, 2012) estimators. All methods use the same predictions, queried answers and query probabilities, so only the final aggregation step changes. MOM uses three consecutive blocks. Each of the other methods uses a tuning parameter selected on development query runs and shared across evaluation settings and budgets. Appendix D.2 describes the tuning procedure, including its overlap with evaluation targets.

Table 3 shows that LV aggregation gives the lowest RMSE in both evaluation settings at both budgets. On the 19-benchmark collection, it reduces RMSE relative to Catoni, the best-performing alternative, by 0.668% and 0.763% at the 5% and 10% budgets, respectively. On the FAQ evaluation data, the corresponding reductions relative to Winsorized mean are 1.426% and 0.696%. Thus, replacing LV with these alternatives does not improve accuracy in the evaluated settings. Appendix D.2 provides uncertainty estimates and tuning-sensitivity results.

Accuracy under answer-record corruption. We test robustness on the FAQ evaluation data by changing only the held-out target models' answer records. Before querying, we uniformly select a fixed number of questions corresponding to a rounded 5% or 10% corruption rate for each target and repetition, and replace their answers by $1 - y_i$. Historical responses, released item factors and the prior remain unchanged. A queried corrupted item returns its flipped answer, which is also used to update the surrogate; unqueried answers remain hidden. We score each estimate against both the original answer mean ("clean") and the corrupted answer mean ("altered"). The altered mean is the estimation target corresponding to the available answers, whereas clean-target scoring assesses recovery of the original accuracy under corruption, which is outside our estimation guarantees. The experiment uses released factors, all 4,425 target model–benchmark pairs, and ten query repetitions.

Table 4 shows that ASA-LV has lower RMSE than FAQ, AIPW-B, AIPW-U and RS mean at both corruption rates, under both scoring targets and at both query budgets. Its RMSE reductions range from 1.8% to 13.4% relative to FAQ and from 0.04% to 11.0% relative to AIPW-U, the strongest external comparator in each comparison. The advantage therefore extends beyond clean answers, although its magnitude varies across conditions. Increasing the query budget reduces ASA-LV's RMSE more substantially against the altered mean than against the original clean mean. More queries improve estimation of the corrupted records' average, but do not by themselves remove the difference between that average and the original accuracy.

Alternative surrogates. Beyond the IRT and released-factor surrogates used in the main experiments, we evaluate ASA-LV with historical-mean and logistic-regression surrogates. The historical-mean surrogate predicts each question's correctness using the average response of historical models and remains fixed during querying. To compare more fairly with AISCV regression, which learns from queried target answers, we also fit a logistic-regression surrogate using the same historical features and queried answers, and apply ASA-LV to its surrogate-augmented contributions. This variant, Logistic Regression–ASA-LV, and AISCV regression are evaluated on identical query trajectories generated using the historical-mean surrogate. Logistic Regression–ASA-LV achieves lower RMSE and MAE and higher Kendall τ_b and Spearman ρ than AISCV regression in both evaluation settings at both budgets. Appendix D.3 reports the historical-mean and regression-based comparisons separately.

Alternative query policy. We also evaluate covariance-aware querying on the 19-benchmark collection, keeping surrogate fitting and ASA-LV aggregation unchanged. Across all 974 target model–benchmark pairs and twenty repetitions, this modification reduces equal-benchmark mean RMSE by 1.11% and 2.86%, and MAE by 1.79% and 2.65%, at the 5% and 10% budgets, respectively. Thus the query-policy modification provides additional improvements without changing the estimator. Appendix D.4 defines the policy and reports the comparison.

Query counts. ASA-LV executes exactly k query draws. Because it samples with replacement, the mean number of distinct queried items is 97.32% and 94.73% of the draw count on the 19-benchmark collection, and 96.81% and 93.66% on the FAQ evaluation data, at the 5% and 10% budgets, respectively. AIPW-B and AIPW-U have random realized query counts, whose mean deviations from the nominal budgets are below 0.05% in both settings at both budgets. Table 7 reports query calls and distinct queried items separately.

7 Discussion and limitations

ASA-LV is intended for benchmark evaluation when querying every answer is expensive and historical responses provide useful but imperfect predictions. Its empirical improvements in accuracy estimation and model ranking support using surrogate-assisted adaptive evaluation under limited query budgets. Robust aggregation is particularly relevant when small query probabilities amplify occasional prediction errors into influential contributions.

The benefit of robust correction depends on the resulting contribution distribution. If the surrogate predicts every answer exactly, all augmented contributions equal the true benchmark accuracy and the LV correction vanishes. When an accurate surrogate already produces well-concentrated contributions, the additional benefit of nonlinear aggregation may be small. Average predictive accuracy alone does not determine this benefit, since a few residual errors on rarely queried questions can still produce large contributions. Surrogate quality, query allocation and aggregation therefore need to be considered together.

Our theory establishes finite-sample error control under adaptive querying and reuse of queried answers for surrogate fitting. These results control estimation error; comparative RMSE performance is assessed empirically. The class-complexity bounds can be conservative for rich surrogate families. Moreover, our experiments concern fixed benchmark matrices and model cohorts. Under answer corruption, the available observations directly identify the altered-record accuracy; recovering the original clean accuracy requires additional assumptions.

An important direction for future work is to develop stronger surrogates, particularly ones that remain accurate on questions that are assigned small query probabilities. Jointly improving surrogate learning and query allocation could reduce the variability of the augmented contributions before aggregation. Another direction is to choose the calibration mass c from the observed contributions, rather than fixing it in advance, while retaining finite-sample guarantees. Such a rule could apply little correction when prediction errors are small and increase downweighting when inverse-probability weighting amplifies occasional residuals into unusually large contributions.

REPRODUCIBILITY STATEMENT

Section 4 specifies the ASA-LV estimator, and Section 6 describes the experimental settings and evaluation metrics. Appendix A documents dataset sizes, historical-data preparation, surrogate fitting, sampling policies, and implementation settings. The proofs and surrogate-class assumptions are provided in Appendices B.1, B.2, and C. Appendix D reports per-benchmark results and additional experiments, including the definitions and tuning procedures for the alternative robust estimators.

AI USE STATEMENT

We used generative AI tools to assist with proof writing and checking, implementation of experimental code, and language editing. The authors take responsibility for the final content of this work, including all text, claims, code, and analyses.

REFERENCES

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pp. 1148–1185, 2012.

Mehdi Dagdoug, Camelia Goga, and David Haziza. Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542): 1234–1251, 2023.

Wilfrid J Dixon. Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, pp. 385–391, 1960.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*, 2025.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Henrik Imberg, Xiaomi Yang, Carol Flannagan, and Jonas Bärgman. Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples. *Technometrics*, 67(1):46–57, 2025.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pp. 5753–5763. PMLR, 2021.

Jasper CH Lee and Paul Valiant. Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 672–683. IEEE, 2022.

Jasper CH Lee, Walter McKelvie, Maoyuan Song, and Paul Valiant. All-purpose mean estimation over r: Optimal sub-gaussianity with outlier robustness and low moments performance. In *Forty-second International Conference on Machine Learning*, 2025.

Rémi Leluc, François Portier, Johan Segers, and Aigerim Zhuman. A quadrature rule combining control variates and adaptive importance sampling. *Advances in neural information processing systems*, 35:11842–11853, 2022.

Puheng Li, Tijana Zrnic, and Emmanuel Candes. Robust sampling for active statistical inference. *Advances in Neural Information Processing Systems*, 38:68727–68756, 2026.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pp. 9167–9176. PMLR, 2020.

Joel Tropp. Freedman's inequality for matrix martingales. 2011.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022.

Hongjian Wang and Aaditya Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and Their Applications*, 163:168–202, 2023.

Justin Whitehouse, Ben Chugg, Diego Martinez-Taboada, and Aaditya Ramdas. Mean estimation in banach spaces under infinite variance and martingale dependence. *arXiv preprint arXiv:2411.11271*, 2024.

Skyler Wu, Yash Nair, and Emmanuel J Candès. Efficient evaluation of llm performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026.

Guanhua Zhang, Florian E Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. *Advances in Neural Information Processing Systems*, 38:135638–135678, 2026.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

A Data and experimental details

A.1 Data sources and benchmark sizes

Table 5 reports the dimensions of each response matrix used in our experiments. Questions correspond to columns, and historical and target models correspond to rows. For the 19-benchmark collection, historical-model counts reflect the preparation described in Appendix A.2; all target models are retained. Counts summed across benchmarks represent model–benchmark pairs, since a model may appear in more than one benchmark. For Composite, the target accuracy is the mean correctness over all 9,574 pooled questions, so component benchmarks contribute in proportion to their question counts. We compute evaluation metrics separately for MMLU-Pro and Composite and then average the two benchmarks equally.

Table 5: Response-matrix dimensions for the evaluated benchmarks. Historical models provide responses for surrogate construction; target models are evaluated using limited answer queries. Question counts refer to the processed matrices used in our experiments.

Benchmark	Questions	Historical models	Target models
19-benchmark collection (Zhang et al., 2026)			
ARC-Challenge	1,172	162	54
BBH	5,761	326	111
Commonsense	500	63	20
GPQA	1,192	324	111
GSM	1,000	63	20
IFEval	541	327	112
ImageNet	50,000	83	27
LegalBench	2,047	60	19
MATH	437	63	20
MedQA	1,000	63	20
MMLU	567	63	20
MMLU-Pro	12,032	327	112
MNLI	9,815	65	21
MRPC	408	66	21
MuSR	756	327	112
OpenLLM MATH	894	286	111
QNLI	5,463	57	21
RTE	277	51	21
SST-2	872	61	21
FAQ evaluation data (Wu et al., 2026)			
MMLU-Pro	12,032	2,212	2,213
Composite	9,574	2,211	2,212

A.2 Historical-data preparation

The primary 19-benchmark results use historical data after removing potential overlap with target models. All 974 target rows are retained. In the original split, 52 target rows in the seven benchmarks with aligned model identifiers shared a name and an identical answer vector with a historical row. Twelve benchmarks lack verified row-level identifiers. The removal procedure below screens historical rows using both available identifiers and full response vectors, then refits the historical surrogates. Earlier results using the original histories and a separate target-filtered sensitivity are retained only in the experiment archive; they are not the primary results reported here.

Overlap screening and refitting. The historical-data preparation used for the primary results retains all 974 target rows and removes historical rows with a verified target-name match or an exactly matching full response vector. There are 52 target rows with named overlap and 87 with response-vector overlap; these counts overlap and are not added. The union of the removal criteria excludes 130 of 2967 historical rows, leaving 2837. Seven benchmarks have aligned model identifiers; twelve remain identity-unverified. Twenty vector-matched target rows occur in those twelve benchmarks.

Vector equality is a conservative possible-duplicate criterion, not proof of identical model weights. Full answer matrices are accessed for offline dataset curation, not for prediction updates or estimator selection.

All nineteen historical IRT models are refitted from scratch with seed 923, 600 steps, and four latent coordinates; historical means and second moments are recomputed on the retained histories. Independent checks verify the retained training lists, unchanged target lists, recomputed moments, and numerical agreement of fits in benchmarks with no removed rows. The primary clean evaluation uses these frozen fits and twenty query repetitions on all 974 targets. Earlier ten-repetition checks and target-filtered analyses are archived separately and are not pooled with the primary results.

A.3 Compared estimators

The following definitions specify the compared estimators and ASA-LV variants. Uppercase symbols denote random objects; lowercase symbols denote fixed quantities or realized values.

FAQ. Factorized Active Querying is the evaluation pipeline of Wu et al. (2026). The reported accuracy estimate is its Proactive Active Inference (PAI) endpoint.

AIPW-B/U. These are the retained native Bernoulli augmented inverse-probability-weighted estimators in the FAQ evaluation protocol. AIPW-B uses uncertainty allocation and AIPW-U substitutes uniform allocation. Their inclusion probabilities and expected-budget construction are given below; they are not categorical-path ablations.

ASA-LV. The Adaptive Surrogate-Augmented Lee–Valiant estimator is $T_c(z, p)$ in Eq. (2), applied to the adaptive surrogate-augmented contributions $z_t = Z_t(f)$.

AISCV regression. This comparator fits the weighted control-variate regression of adaptive importance sampling with control variates (AISCV) to fixed historical item features on our logged categorical trajectories. It does not reproduce AISCV's native proposal adaptation. The fitted intercept is the reported estimate.

Logistic Regression–ASA-LV. This admissible ASA-LV variant fits an inverse-propensity-weighted logistic ridge model to the queried answers and fixed historical item features. Its sigmoid predictions lie in $[0, 1]$. ASA-LV is applied to the resulting augmented contributions. Appendix B.2 uses this class to state the terminal-fit complexity bound.

RS mean. Random sampling (RS) reports the unadjusted mean of a fixed-size uniform sample of answers.

A.4 Surrogate fitting and query policies

Surrogate fitting. The local historical IRT fit has logits $a_m + b_i + u_m^\top v_i$, with four latent coordinates. All historical parameters start from independent $\mathcal{N}(0, .1^2)$ draws using seed 923. We run 600 block-scaled gradient steps with step size .5 and ridge coefficient .01: model gradients average over items and item gradients average over historical models. The resulting fixed item feature is $\phi_i = (1, v_i^\top)^\top$. Target parameters start from $\mathcal{N}(0, .1^2)$ using the paired study-specific seed. The update uses all previously observed distinct answers, forty steps, ridge .05, and the learning rate specified in the IRT parameter-bound appendix; it is not inverse-probability-weighted fitting. Logits are clipped to $[-30, 30]$. Updates occur after each four new answers, so the available terminal fit can have pending observations. The released 64-dimensional factor model instead retains its historical prior and Laplace update after each query.

Regression implementation. Historical item features are fixed before target querying. For the local IRT runs and the historical-mean regression comparisons in Appendix D.3, the input columns are the historical mean correctness, fitted item intercept, and four local IRT item coordinates. For AISCV regression in the released-factor main comparison, they are the historical mean and 64 released item-factor coordinates. We center columns over the full benchmark, remove constant or linearly dependent directions using a singular-value decomposition with relative threshold 10^{-12}, and rescale the retained basis to have identity empirical second-moment matrix. A separate intercept is included in each regression.

AISCV regression fits weighted least squares on the queried answers with weights proportional to $1/\{nq_t(i_t)\}$ and reports the fitted intercept; the population means of the control features are zero. Logistic Regression–ASA-LV uses the same control features and query observations, fits weighted logistic regression with weights normalized to average one, and applies ridge penalty 0.01 to the non-intercept coefficients. The implementation uses at most 60 damped Newton steps with backtracking and clips logits to $[-30, 30]$. The resulting full-benchmark predictions define the augmented contributions, which are aggregated with $b = 3$ and $c = \log(10)/3$. Repeated draws retain their individual logged probabilities in both regressions.

Categorical allocation. For the uncertainty policy, form $s_i = \sqrt{\pi_i(1-\pi_i)}$ from the current predicted correctness probability π_i and set $q_i = \tau/n + (1-\tau)s_i/\sum_j s_j$, with a uniform fallback if every score is zero. The WOR implementation removes previously queried items from this full-benchmark mixture and renormalizes the remaining probabilities. This still ensures $q_t(i) \geq \tau/n$ on the remaining set, but is not defined by simply substituting the remaining-set size for n in each mixture component.

Native Bernoulli controls. B/U refer to the Bernoulli AIPW variants used here. A random item order is scanned once. B allocates using fixed historical uncertainty, whereas U uses constant allocation scores. With t previously scanned items, cumulative expected calls E_t, current score s_i and mean previous score $\bar{s}_t$, the inclusion probability is

$$g_i = \min\left\{1, \max\left\{.01k/n, \frac{\max(0, k - E_t)}{n - t}\frac{s_i}{\bar{s}_t}\right\}\right\}.$$

The first item uses its own score as the reference mean; nonuniform scores are floored at 10^{-12}. Thus the Bernoulli inclusion floor is $.01k/n$ even when the categorical floor differs. This online rule targets k expected queries but need not attain that expectation exactly; Table 7 reports the observed mean query counts. B/U return $n^{-1}\sum_{i=1}^{n}\{\pi_i + S_i(y_i - \pi_i)/g_i\}$, where π_i is the prediction immediately before item i's inclusion draw and S_i is its Bernoulli inclusion indicator. Each estimator uses the common surrogate implementation fitted on its own previously observed answers. Native allocation designs and realized costs differ, and the same-trajectory controls isolate estimator changes.

ASA-LV calibration implementation. The chronological pilot takes the median of three contiguous block means, with block lengths differing by at most one. Calibration sorts squared pilot deviations and solves the piecewise-linear deletion-mass equation. Constant vectors return that constant; unattainable mass deletes every nonzero deviation and is recorded separately. The ASA-LV output is a direct calibrate-and-average calculation and has no iterative center solver. Source hashes and diagnostics record the calibration outcomes.

A.5 Budgets, computation and reproducibility

Budgets and calibration. All ASA-LV estimates in the main clean evaluation use unit round weights, $b = 3$, and $c = \log(10)/3$. The general b-block definition requires $1 \leq b \leq k$; hence the primary choice $b = 3$ requires $k \geq 3$. All reported budgets satisfy this condition, and the implementation uses the stated fixed block count rather than an implicit $\min(k, b)$ convention.

Table 6: Primary categorical-allocation and ASA-LV calibration settings. Here $\phi \in \{.05, .10\}$ is the query-budget fraction.

Corpus/predictor	Categorical floor τ	Budget k	Mass c
Nineteen benchmarks, mean/IRT	.01	$\lceil \phi n \rceil$	$\log(10)/3$
FAQ MMLU-Pro, mean/IRT/factor	.05	$\lfloor \phi n \rfloor$	$\log(10)/3$
FAQ composite, mean/IRT/factor	.25	$\lfloor \phi n \rfloor$	$\log(10)/3$

At the 5% and 10% budgets, the authors' released query parameters give the score $\sqrt{F_{t-1}(i)\{1 - F_{t-1}(i)\}}$, with no additional covariance term in the sampling score. The posterior covariance is still used in the per-query surrogate update. Categorical sampling uses exactly k

calls, including repeated items under WR; native Bernoulli controls use an expected-budget design. Actual and distinct calls are retained.

Computation versus query cost. The query budget is the number of target-answer calls: categorical runs make exactly k calls (including repeated WR items), while native Bernoulli methods target an expected budget and retain their realized counts. Online allocation requires one full-benchmark score refresh whenever the predictor is updated; local IRT performs forty five-parameter gradient steps after every four new distinct labels, and the released factor model performs its specified 64-dimensional Laplace update after each query. Terminal reconstruction is one full-benchmark prediction plus $O(k)$ contribution formation. The generic ASA-LV implementation uses sorted calibration and therefore adds $O(k \log k)$ time and $O(k)$ storage; in the deployed $c < 1$ case, closed-form unsaturated calibration also permits an $O(k)$ implementation. Wall-clock latency was not benchmarked across heterogeneous machines, so the paper makes query-cost, not deployment-latency, claims.

Query calls and distinct reveals. Table 7 separates calls from distinct queried items. Categorical ASA-LV always uses exactly the nominal number of draws; repeated indices make the number of distinct reveals smaller. Bernoulli paths never repeat an item but have random realized counts. Across the four primary comparisons, the mean Bernoulli count differs from nominal by at most .043%.

Table 7: Query counts and distinct questions. "Distinct" is the mean percentage of ASA-LV draws that reveal a distinct item. The last two columns are percentage deviations of mean realized Bernoulli calls from the nominal budget.

Setting	Budget	ASA-LV distinct	AIPW-B Δ calls	AIPW-U Δ calls
19-benchmark IRT	5%	97.32%	-0.042%	$+0.021\%$
19-benchmark IRT	10%	94.73%	-0.002%	-0.001%
FAQ factors	5%	96.81%	-0.020%	-0.027%
FAQ factors	10%	93.66%	-0.011%	-0.011%

Model-ranking computation. For every benchmark and query repetition, Kendall τ_b is computed between the estimated and full-data model rankings with tie adjustment; Spearman ρ is the Pearson correlation of their average ranks. Correlations are averaged over twenty repetitions within benchmark and then equally over benchmarks. The complete primary values are reported in Table 2; ASA-LV has the largest displayed average for both metrics, both budgets, and both surrogate regimes. Ranking was added after inspecting RMSE and is therefore treated as exploratory rather than confirmatory. Tables 12 and 13 report paired differences and marginal 95% intervals based on the twenty complete query repetitions; the calculation and its conditional scope are described alongside those tables.

Evaluation runs. Table 8 summarizes the evaluation runs and their repetition counts.

Table 8: Evaluation runs and their roles.

Study	Targets	Repetitions	Purpose
Clean evaluation	974 local; 4425 FAQ	20	Main clean comparison
Answer corruption	4425 FAQ	10	Table 4
Regression surrogates	974 local; 4425 FAQ	20	Appendix D.3
Robust aggregation	974 local; 4425 FAQ	20	Table 3
Covariance-aware querying	974 local	20	Appendix D.4

Development and evaluation samples. The main clean evaluation uses twenty query repetitions on the fixed target models. Preliminary checks used ten repetitions on the same 974 targets in the 19-benchmark collection. Robust-estimator tuning uses 190 of these target model–benchmark pairs with separate query runs, so the development and evaluation samples overlap at the target level. No FAQ evaluation runs are used to select the tuning multipliers. Different query seeds measure sampling variability on the evaluated targets rather than variation across new models.

Reconstructing results and rerunning experiments. Reconstructing a table from saved estimates or squared errors repeats the aggregation and statistical analysis but does not rerun adaptive queries or surrogate fitting. A full experimental rerun additionally requires the response matrices and historical splits, predictor fitting procedures, query policies, random seeds, budgets and aggregation settings described above. Repetition counts and target cohorts refer to the experimental runs, not to the number of times their saved outputs were analyzed. The reported uncertainty intervals are computed from those saved repeated-run outputs.

B Proof of Theorem 1 and surrogate-class bounds

FAQ constructs each contribution using the predictor available before the corresponding query (Wu et al., 2026). In contrast, we reconstruct all contributions using a terminal surrogate fitted with the queried answers, while retaining the original logged query propensities. This reconstruction motivates an analysis that permits same-sample terminal fitting: although the reconstructed contributions need not retain the conditional-mean property of their predictable counterparts, the uniform surrogate-class argument below provides finite-sample error control under the stated exploration and complexity conditions.

B.1 Proof of Theorem 1

Here ϵ denotes the sup-norm approximation error of the surrogate-class covering net, not the predictor's error relative to the benchmark answers. The latter affects the residuals and their conditional variance. Theorem 1 controls estimation error for the fixed-mass configuration; it does not compare its expected error with another aggregation rule.

Event and filtration. Condition on the pre-query historical data and initialization. The answer bank is fixed; conditional on this initial information, randomness comes from adaptive querying and any subsequent algorithmic randomization. The finite net $\mathcal{G} \subset \mathcal{F} \subset [0,1]^n$ is predetermined. For each fixed $g \in \mathcal{G}$, concentration is under the original adaptive law with predictable random propensities. In the deterministic comparison of predictor outputs on one realized trajectory, reuse its logged probabilities in Eq. (1); do not rerun the policy under a counterfactual predictor. The eligible set is the full bank in WR and the remaining items in WOR. Its known-answer augmentation gives $\mathbb{E}[Z_t(g) \mid \mathcal{H}_{t-1}] = \mu$ in both cases. Indeed, conditionally on $\mathcal{H}_{t-1}$, the predictor g, eligible set, and query probabilities are fixed, so $\mathbb{E}[(Y_t - g(I_t))/(nQ_t(I_t)) \mid \mathcal{H}_{t-1}] = n^{-1}\sum_{i\in U_t}(y_i - g(i))$. Adding $B_t(g)$ gives $n^{-1}\sum_i y_i = \mu$. Write

$$A_g = \sum_{t=1}^{k}(Z_t(g) - \mu), \qquad V_g = \sum_{t=1}^{k}\mathrm{Var}(Z_t(g) \mid \mathcal{H}_{t-1}).$$

All probability statements below are conditional on this initial information; the finite net and the horizon are fixed under that conditioning. The centered increment is bounded by $1 + 1/\tau \le 2/\tau$ in absolute value. With $D_U = \sum_t |U_t|/n \le k$,

$$V_g = \sum_t \left\{ \frac{1}{n^2}\sum_{i\in U_t}\frac{(y_i - g(i))^2}{Q_t(i)} - \left[\frac{1}{n}\sum_{i\in U_t}(y_i - g(i))\right]^2 \right\} \le D_U/\tau \le k/\tau.$$

Freedman's inequality and uniform concentration. For a fixed g, let $X_t(g) = Z_t(g) - \mu$. These martingale differences satisfy $|X_t(g)| \le R := 2/\tau$. The scalar Freedman inequality (Tropp, 2011, Theorem 1.1), applied to both signs, gives $\Pr\{|A_g| \ge \sqrt{2vx} + 2Rx/3,\ V_g \le v\} \le 2e^{-x}$ for deterministic $v, x > 0$. We apply this inequality on deterministic variance ranges and take a union bound over the net predictors. Partition the range of V_g into the first layer $V_g \le 1$ and successive dyadic layers up to k/τ. There are at most $m_{\mathrm{peel}} = 1 + \lceil \log_2 \max(1, k/\tau)\rceil$ layers. On each layer its upper endpoint is at most $2\max(V_g, 1)$. Scalar Freedman (Tropp, 2011, Theorem 1.1), on both signs and all $j = |\mathcal{G}|$ net points, gives with probability at least $1 - \alpha$,

$$|A_g| \le 2\sqrt{\max(V_g, 1)\ell} + \frac{4\ell}{3\tau} \quad \text{for all } g \in \mathcal{G}, \qquad \ell = \log\frac{2m_{\mathrm{peel}}(j+1)}{\alpha}.$$

The union failure probability is at most $2jm_{\mathrm{peel}}e^{-\ell} = \alpha j/(j+1) \le \alpha$. Here m_{peel} and ℓ are exactly m and ℓ in Theorem 1. For $\|f-g\|_\infty \le \epsilon$,

$$|B_t(f) - B_t(g)| \le \epsilon |U_t|/n \le \epsilon, \qquad |Z_t(f) - Z_t(g)| \le (1+1/\tau)\epsilon.$$

Hence, for a nearest net element G_F to the data-dependent terminal fit F (with a fixed tie-breaking rule),

$$|\bar{Z} - \mu| \le \frac{2\sqrt{\max(V_{G_F}, 1)\ell}}{k} + \frac{4\ell}{3\tau k} + (1+1/\tau)\epsilon.$$

Choosing G_F after seeing the data is allowed because the event is uniform. Conditioning on the final fitted predictor itself would not justify this step.

For the variance-insensitive alternative in Eq. (4), use the deterministic envelope $V_g \le k/\tau$ established above in scalar Freedman before the union bound, rather than peeling the realized V_g. Applying both signs and all j net points with $\ell_0 = \log\{2(j+1)/\alpha\}$ gives

$$|A_g| \le 2\sqrt{k\ell_0/\tau} + \frac{4\ell_0}{3\tau} \qquad \text{for every } g \in \mathcal{G}$$

with probability at least $1-\alpha$. Freedman actually gives $\sqrt{2k\ell_0/\tau}$ in the square-root term; replacing $\sqrt{2}$ by 2 yields the displayed bound. Dividing by k and applying the same net approximation gives

$$|\bar{Z} - \mu| \le 2\sqrt{\frac{\ell_0}{\tau k}} + \frac{4\ell_0}{3\tau k} + (1+1/\tau)\epsilon.$$

This is a separate event with probability at least $1-\alpha$; no simultaneous $1-\alpha$ claim for the two bounds is needed. This argument explains why a log-free $O(k^{-1/2})$ statement requires fixed j (or another bounded-complexity argument); it is not obtained by suppressing a growing covering number.

Pathwise transfer to the experimental estimator. For the pilot p and attained deletion mass $\widetilde{c} = \sum_t d_t(p) \le c$, the exact identity is

$$T_c(z,p) - \bar{z} = -\frac{1}{k}\sum_t (z_t - p) d_t(p).$$

Every chronological block mean and its median belong to $[\min_t z_t, \max_t z_t]$. Therefore

$$|T_c(z,p) - \bar{z}| \le c\{\max_t z_t - \min_t z_t\}/k.$$

In WR the realized terminal baseline $B_t(f) = \bar{f}$ is constant across queries and the residual term lies in $[-1/\tau, 1/\tau]$, so the range is at most $2/\tau$. In WOR the baseline may vary in $[0,1]$, giving $1 + 2/\tau$. These range bounds hold for every realized bounded predictor, including the data-dependent final predictor. The same displacement argument holds for any pilot in the contribution range, as allowed in Theorem 1.

Conclusion of Theorem 1. On the uniform peeled event, the triangle inequality now gives

$$\begin{aligned} |T_c(Z,P) - \mu| &\le |\bar{Z} - \mu| + |T_c(Z,P) - \bar{Z}| \\ &\le \frac{2\sqrt{\max(V_{G_F}, 1)\ell}}{k} + \frac{4\ell}{3\tau k} + \left(1 + \frac{1}{\tau}\right)\epsilon + \frac{c r_\tau}{k}. \end{aligned}$$

This is precisely Eq. (3), with probability at least $1-\alpha$. On the separate deterministic-variance event, the same displacement bound gives

$$|T_c(Z,P) - \mu| \le 2\sqrt{\frac{\ell_0}{\tau k}} + \frac{4\ell_0}{3\tau k} + \left(1 + \frac{1}{\tau}\right)\epsilon + \frac{c r_\tau}{k},$$

which is Eq. (4), also with probability at least $1-\alpha$. For fixed α, c, j, with τ bounded away from zero and $\epsilon = O(k^{-1/2})$, the latter is $O(k^{-1/2})$. More generally it is $O(\sqrt{\log(j_k+1)/k} + \log(j_k+1)/k + \epsilon_k)$ under the same conditions on α, c, τ. Unattainable calibration with $\widetilde{c} < c$, saturation, and any hull-valued pilot are covered. The result does not need concentration of the pilot.

For any data-selected $A \in \{0,1\}$, the estimator $AT_c(z,p) + (1-A)\bar{z}$ has no larger displacement bound. A common projection onto $[0,1]$ cannot increase its absolute error to μ. Thus arbitrary zero-gamma switching retains the rate; it need not retain unbiasedness or improve expected squared error. Fixed c and a fixed positive τ give an $O(k^{-1})$ correction. If they vary, their displayed dependence must remain.

B.2 SURROGATE-CLASS BOUNDS

The following class bounds concern the stated update recursions in exact arithmetic, conditional on historical features and initialization. Fixed-horizon classes need not contain a converged optimizer: every scheduled intermediate iterate is covered. Floating-point spot checks are implementation diagnostics, not proofs of parameter containment or covariance positivity on every path. Write $N_\infty(\epsilon, \mathcal{F})$ for the minimum cardinality of an ϵ-net of $\mathcal{F}$ in the supremum norm over the benchmark questions, with net points in $\mathcal{F}$.

Fixed historical mean. Conditioning on the historical data fixes the mean predictor, so its class is a singleton ($j = 1$), with no terminal-fitting complexity cost.

Logistic and clipped regression with fixed historical features. Let $\phi_1, \dots, \phi_n \in \mathbb{R}^p$ be fixed before target querying (with a constant coordinate included when the fitted model has an unpenalized intercept) and let $\psi : \mathbb{R} \to [0, 1]$ be either clipping to $[0, 1]$ or a logistic link with fixed logit clipping. Consider

$$\mathcal{F}_{\mathrm{reg}} = \{i \mapsto \psi(\phi_i^\top \beta) : \beta \in \mathbb{R}^p\}.$$

This class includes every intermediate and terminal output of the bounded regression variant because only the coefficients use queried labels; the historical feature vectors and link are fixed in advance.

Lemma 1 (Finite-bank cover without a coefficient bound)**.** *For $0 < \epsilon < 1$, let $q = \lceil 1/\epsilon \rceil$. Then*

$$N_\infty(\epsilon, \mathcal{F}_{\mathrm{reg}}) \le \sum_{r=0}^{\min(p,L)} 2^r \binom{L}{r}, \qquad L \le n(q-1),$$

where L is the number of effective affine-threshold hyperplanes. Consequently, for fixed p, $\log N_\infty(\epsilon, \mathcal{F}_{\mathrm{reg}}) = O\{p \log(1 + n/\epsilon)\}$.

Proof. Partition $[0, 1]$ into q intervals of length at most ϵ, assigning endpoints by a fixed rule. For each item and each effective internal output boundary, membership on either side is determined by at most one affine threshold in coefficient space; constant conditions, including clipped-link plateaus and zero feature rows, are omitted. Thus there are at most $L \le n(q-1)$ effective threshold hyperplanes. Retain threshold equalities and faces of every dimension. On every face of their arrangement, each item prediction remains in one output interval, so two predictions from the same face differ by at most ϵ in sup norm. Choose one coefficient vector from every nonempty face. The total number of faces of L hyperplanes in p dimensions is at most $\sum_{r=0}^{\min(p,L)} 2^r \binom{L}{r}$: adding one hyperplane gives the recursion $F(L, p) \le F(L-1, p) + 2F(L-1, p-1)$. Degenerate or coincident hyperplanes only reduce this count. The chosen representatives belong to the original class, which proves the cover without bounding $\|\beta\|$. □

If several feature families are selected after observing target labels, their union must be included in the predetermined class and its cover sizes added. At $\epsilon = 1/k$, fixed p, fixed sampling fraction k/n, and fixed τ, c, α, substitution in Theorem 1 gives $O\{\sqrt{p \log k/k} + p \log k/k + 1/k\}$, not a log-free $O(k^{-1/2})$ statement. This class argument does not assume optimization convergence or a penalized intercept.

Local IRT parameter bound. Write $\mathrm{sigmoid}(x) = (1 + e^{-x})^{-1}$ for the logistic link used in the IRT and released-factor updates below.

Let $\phi_i = (1, v_i^\top)^\top$ be the fixed historical item feature and $r_\phi = \max_i \|\phi_i\|_2$. At a realized update, let $\mathcal{J}$ be the set of distinct queried questions used in that update and $m = |\mathcal{J}|$. The average $\overline{\|\phi\|^2} = m^{-1} \sum_{i \in \mathcal{J}} \|\phi_i\|^2$ uses each of these questions once. Write $f_\theta(i) = \mathrm{sigmoid}(\mathrm{clip}_{[-30,30]}(b_i + \phi_i^\top \theta))$, where $\mathrm{clip}_{[-30,30]}(x) = \min\{30, \max\{-30, x\}\}$, matching the logit clipping used in the implementation. The queried-item update uses ridge strength $\lambda = .05$ and learning rate $\eta = .8/(.25\,\overline{\|\phi\|^2} + .05)$, with forty gradient steps after each four newly observed distinct labels. Each step has the form

$$\theta^+ = (1 - \eta\lambda)\theta - \eta \frac{1}{m} \sum_{i \in \mathcal{J}} \phi_i \{f_\theta(i) - y_i\}.$$

Here $0 \le \eta\lambda \le .8$ and the norm of the averaged data-dependent term is at most r_ϕ. Consequently, the ball of radius $r_\theta = \max(\|\theta_0\|, r_\phi/\lambda)$ is invariant. Because the clipped logistic link is $1/4$-Lipschitz, $\|f_\theta - f_{\theta'}\|_\infty \le (r_\phi/4)\|\theta - \theta'\|$. A Euclidean ball cover therefore yields $N_\infty(\epsilon, \mathcal{F}) \le (1 + r_\theta r_\phi/(2\epsilon))^d$. This establishes a rate statement, not a guarantee that the resulting numerical confidence radius is useful.

Released-factor parameter bound. Let M and Σ denote the random target latent posterior mean and covariance, v and b_i the fixed historical item vector and intercept, and $\Pi_i = \mathrm{sigmoid}(b_i + v^\top M)$ its random current prediction. For observed answer y, the released Laplace update has

$$\Sigma^+ = \Sigma - \frac{w\Sigma v v^\top \Sigma}{1 + w v^\top \Sigma v}, \qquad M^+ = M + \Sigma^+ v(y - \Pi_i), \qquad w = \Pi_i(1 - \Pi_i) \ge 0.$$

Starting with $\Sigma_0 \succ 0$, the matrix inversion lemma yields $0 \prec \Sigma^+ = (\Sigma^{-1} + wvv^\top)^{-1} \preceq \Sigma$. Thus, for $\|v\| \le r$ and $y, \Pi_i \in [0,1]$, $\|M^+ - M\| \le \|\Sigma_0\| r$ and $\|M_k\| \le m_k := \|M_0\| + k\|\Sigma_0\| r$. Conditioning on history and initialization fixes this class in advance. A cover of the 64-dimensional ball gives $N_\infty(\epsilon, \mathcal{F}) \le (1 + m_k r/(2\epsilon))^{64}$. With historical feature norms, prior bounds, and dimension fixed, $m_k = O(k)$, so the log-cover contribution remains $O(d \log k)$ at $\epsilon = 1/k$. Their dependence must be retained if these quantities vary with k. The prior matrices in both released banks were numerically checked to be positive definite; covariance and mean-update checks on the first, middle, and last targets agree with the exact-arithmetic inequalities. These checks do not supply a global roundoff certificate. The numerical radii remain conservative.

B.3 Sampling without replacement

For a realized terminal predictor f, let $\bar{b}_f = k^{-1} \sum_t B_t(f)$ be the average baseline in Eq. (1).

For WOR, previously queried answers are known. Replacing their predictions by these known answers gives

$$\bar{b}_f = \frac{1}{n} \sum_i f(i) + \frac{1}{nk} \sum_{t=1}^{k} (k - t)\{y_{I_t} - f(I_t)\}.$$

The coefficient counts how many later rounds include each queried item in the known set. This augments the plain unbiased risk estimator (PURE) of Farquhar et al. (2021) with surrogate predictions. Using only the full-bank prediction mean with WOR per-draw probabilities would omit this correction.

B.4 Numerical implementation

The statements above describe exact arithmetic. More generally, suppose the computed vector and hull pilot satisfy $\|\widetilde{z} - z\|_\infty \le \eta_z$ and $\widetilde{p} \in [\min_t \widetilde{z}_t, \max_t \widetilde{z}_t]$. For returned weights $0 \le \widetilde{d}_t \le 1$, write $\widetilde{c} = \sum_t \widetilde{d}_t \le c + \eta_{\mathrm{cal}}$, and let η_{out} bound the error in evaluating the final calibrate-and-average expression on these computed inputs. Pathwise,

$$|\widetilde{T} - \bar{z}| \le \eta_z + \frac{(c + \eta_{\mathrm{cal}})(r_\tau + 2\eta_z)}{k} + \eta_{\mathrm{out}}.$$

Thus this expression replaces cr_τ/k in Eq. (3). The code accepts a reported attainable-mass residual of at most $10^{-10} \max(1, c)$; it does not enforce one-sided under-calibration. Its residual checks use floating-point arithmetic, so this threshold alone is not a certified bound on all input and arithmetic errors. One may instead use the actual deletion sum when available. Neither pilot LV nor this displacement argument requires an iterative center solve.

Scope of the sharper result. Theorem 1 applies to the fixed experimental $c = \log(10)/3$ and the general b-block construction (evaluated with $b = 3$) at every confidence level through the linear event; its radius is conservative. The sharper cumulative-variance LV argument in Appendix C instead chooses mass and block count as functions of confidence, and its terminal-class extension pays a net-size adjustment. Those are different estimator configurations. Neither theorem establishes a lower experimental MSE, estimates the unknown variance envelope from final residuals, or proves a general MSE dominance claim.

C Proof of Theorem 2

We prove Eq. (6) in three steps: bound the block-median pilot under the cumulative conditional-variance assumption, control aggregation centered at the true mean, and bound the change caused by substituting the pilot. We then extend the result uniformly to the fitted surrogate. The pointwise inequality imported from Lemma 7 and its proof in the arXiv version (arXiv:2011.08384v1) of Lee & Valiant (2022) is deterministic; all probabilistic steps below are derived under the present adaptive law. The resulting variance-dependent bound does not require a variance input to the estimator, but is not an observable confidence interval unless a justified variance envelope is available.

C.1 Setup and pilot concentration

Let $W = (W_1, \dots, W_k)$ be adapted, $\mathbb{E}[W_t \mid \mathcal{H}_{t-1}] = \mu$, and $X_t = W_t - \mu$. This auxiliary sequence is the one in Theorem 2; terminal contributions $Z_t(F)$ are handled by the uniform-class argument below. Suppose a deterministic $v_\star$ satisfies $\sum_{t=1}^{k} \mathbb{E}[X_t^2 \mid \mathcal{H}_{t-1}] \le v_\star$ almost surely. An expected variance bound alone is insufficient for this assumption. Here $k \ge 1$ is a fixed integer and $v_\star < \infty$; all constructions and bounds in this section initially refer to exact arithmetic. For total failure $\alpha \in (0,1)$, let $\delta = \alpha/2$, $\ell = \log(1/\delta)$, and $c = \ell/3 < k$. Set $b = \min(k, \lceil \ell \rceil)$, with $b \ge 1$. Partition the observations chronologically into b blocks of sizes differing by at most one. Let P_W be the median of their means, averaging the two middle values when b is even. Calibrate $\Gamma_W \ge 0$ so that $D_t^W = \min\{\Gamma_W (W_t - P_W)^2, 1\}$ satisfies $\sum_t D_t^W = c$, and return

$$T_c(W, P_W) = P_W + \frac{1}{k} \sum_t (W_t - P_W)(1 - D_t^W).$$

If there are fewer than c nonzero deviations, delete all of them. The estimator uses no variance input. A fixed horizon and deterministic partition are part of the result.

Lemma 2 (Dependent-block pilot)**.** *For any deterministic partition into b blocks of minimum size s, any $u > 0$, and any $\lambda > 0$,*

$$\Pr\{|P_W - \mu| > u/s\} \le \exp\{\lambda v_\star / u^2 - (b/2)\log(1+\lambda)\}.$$

If $v_\star > 0$ and $0 < \rho < 1$, choose $u^2 = a_\rho v_\star / b$ and $\lambda = a_\rho/2 - 1$, where $a_\rho > 2$ uniquely solves $\frac{1}{2}\log(a_\rho/2) - \frac{1}{2} + 1/a_\rho = \log(1/\rho)/b$. Then the failure probability is at most ρ.

Proof. Within a block, write S for its partial centered sum and $\psi(S) = 1 + \lambda S^2/u^2$. For $v_t = \mathbb{E}[X_t^2 \mid \mathcal{H}_{t-1}]$,

$$\mathbb{E}[\psi(S + X_t) \mid \mathcal{H}_{t-1}] = \psi(S) + \lambda v_t / u^2 \le \psi(S) e^{\lambda v_t / u^2}.$$

At a completed block j, let $A_j = \mathbf{1}\{|S_j| > u\}$. Replacing its potential by $1 + \lambda A_j$ only decreases it. Hence the product of completed-block factors, the current partial-block potential, and $\exp(-\lambda \sum_{t \text{ observed}} v_t/u^2)$ is a nonnegative supermartingale starting at one. At the terminal time, $\sum_j A_j \ge b/2$ makes it at least $(1+\lambda)^{b/2} e^{-\lambda v_\star/u^2}$. Markov's inequality proves the bound. A median error greater than u/s forces at least $b/2$ block means to exceed that threshold in absolute value, so their block sums exceed u. This also handles even b and unequal sizes. No independence of blocks or separate block variance budgets was used. For the stated choice of u and λ, the exponent equals $-bh(a_\rho)$, where $h(a) = \frac{1}{2}\log(a/2) - \frac{1}{2} + 1/a$. Since $h(2) = 0$, $h'(a) = (a-2)/(2a^2) > 0$ for $a > 2$, and $h(a) \to \infty$, the defining equation has a unique solution and the failure bound equals ρ. □

C.2 Aggregation centered at the true mean

For $v_\star = k\sigma^2 > 0$, write $X_t = (W_t - \mu)/\sigma$ and $v_t = \mathbb{E}[X_t^2 \mid \mathcal{H}_{t-1}]$, so $\mathbb{E}[X_t \mid \mathcal{H}_{t-1}] = 0$ and $\sum_t v_t \le k$ almost surely. Set $D_\gamma = \sum_t \min(\gamma X_t^2, 1)$ and $H_\gamma = \sum_t X_t[1 - \min(\gamma X_t^2, 1)]$.

A deterministic inequality and its conditional application. Lemma 7 of the same arXiv version and the parameter choices in its proof give the following pointwise inequality. Substitute $y = \sqrt{\gamma} x$

in that lemma and write η for its parameter named b. For $v \in [.05, 55.5]$ and every real x,

$$e^{\lambda x[1-\min(\gamma x^2,1)]-\eta\min(\gamma x^2,1)} \le 1+\lambda x+\beta x^2,$$
$$\gamma = \frac{c}{kv}, \qquad \lambda = a\sqrt{\gamma}, \qquad \beta = \gamma v[-3+a\sqrt{6/v}-\eta].$$

At $v = .05$, take $a = .75, \eta = \sqrt{3}$; otherwise take $a = 24\sqrt{v}/(\sqrt{6}+\sqrt{6+48v})$ and $\eta = 3-a^2/2$. The latter a is the positive solution of $\sqrt{v}(a^2-12)+\sqrt{6}a = 0$ used in the cited proof. The special endpoint choice need not be continuous with the other choices. These choices have $\beta > 0$: in the general case $\beta/\gamma = v(6-a^2/2) > 0$, and direct substitution gives the same sign at $v = .05$.

For each deterministic choice of these parameters, conditional mean zero gives

$$\mathbb{E}[e^{\lambda X_t(1-\min(\gamma X_t^2,1))-\eta\min(\gamma X_t^2,1)} \mid \mathcal{H}_{t-1}] \le 1+\beta v_t \le e^{\beta v_t}.$$

Thus $M_s = \exp\{\sum_{t=1}^{s}[\lambda X_t(1-\min(\gamma X_t^2,1))-\eta\min(\gamma X_t^2,1)-\beta v_t]\}$ is a nonnegative supermartingale with $M_0 = 1$. Since $\beta \ge 0$ and $\sum_t v_t \le k$, Markov's inequality yields $\Pr\{\lambda H_\gamma - \eta D_\gamma > \beta k + L\} \le e^{-L}$ for $L > 0$. For $L \le 0$ the same bound is trivially valid because its right side is at least one. This step requires neither independence nor identical conditional distributions. A random calibrated γ_* cannot be substituted into the conditional expectation; the next step handles it by deterministic comparisons.

Covering all calibrated values. Take $v_j = .05(55.5/.05)^{j/m_{\text{grid}}}$, $0 \le j \le m_{\text{grid}}$, where $m_{\text{grid}} = \lceil \log(55.5/.05)\max(1,\ell)\rceil$. Put $\gamma_j = c/(kv_j)$, $e_0 = 0$, $e_j = v_j/v_{j-1}-1$ for $j \ge 1$, $s_j = a_j/\sqrt{v_j}$ and $p_j = e_j(a_j+|\eta_j|)$. Here γ_* is any exact calibration on the normalized vector at the true mean, so $D_{\gamma_*} = c$.

For an interior interval $\gamma_j \le \gamma_* \le \gamma_{j-1}$, $1 \le j \le m_{\text{grid}}$, and any $q \ge 0$,

$$0 \le \min(\gamma_* q, 1) - \min(\gamma_j q, 1) \le e_j \min(\gamma_j q, 1).$$

Nonzero differences require $|X_t| < 1/\sqrt{\gamma_j}$, and $D_{\gamma_j} \le D_{\gamma_*} = c$. Summing therefore gives $|H_{\gamma_*} - H_{\gamma_j}| \le e_j c/\sqrt{\gamma_j}$ and $|c - D_{\gamma_j}| \le e_j c$. Consequently $H_{\gamma_*} > \sqrt{kc}\,r$ implies

$$\begin{aligned}\lambda_j H_{\gamma_j} - \eta_j D_{\gamma_j} - \beta_j k &> \lambda_j\sqrt{kc}\,r - e_j a_j c - \eta_j c - e_j|\eta_j|c - \beta_j k \\ &= c\{3+s_j(r-\sqrt{6})-p_j\}.\end{aligned}$$

The equality follows from $\lambda_j\sqrt{kc} = cs_j$ and $\beta_j k = c(-3+s_j\sqrt{6}-\eta_j)$. Thus the probability of this event on the interior interval is at most $\exp[-c\{3+s_j(r-\sqrt{6})-p_j\}]$.

For the upper exterior $\gamma_* \ge \gamma_0$, every unsaturated summand of $\lambda_0 H_\gamma - \eta_0 D_\gamma$ has derivative $-(\lambda_0 X_t + \eta_0)X_t^2 \le 0$ for $\gamma \ge \gamma_0$, since $|\lambda_0 X_t| \le a_0 < \eta_0$. Saturated summands are constant and the expression is continuous at their boundaries. Hence its value at γ_0 is at least its value at γ_*, and the same tail argument applies with $j = 0$ and $p_0 = 0$.

For the lower exterior $\gamma_* \le \gamma_{m_{\text{grid}}}$, calibration implies $D_{\gamma_{m_{\text{grid}}}} \ge c$. For $s > 0$, convexity on $[0,1]$ gives $\mathbb{E}[e^{s\min(\gamma_{m_{\text{grid}}}X_t^2,1)} \mid \mathcal{H}_{t-1}] \le e^{(e^s-1)\gamma_{m_{\text{grid}}}v_t}$. The compensated-product argument and Markov's inequality, with $s = \log 55.5$, bound the probability by

$$\exp\{-c\log 55.5 + (55.5-1)\gamma_{m_{\text{grid}}}k\} = \delta^\kappa, \qquad \kappa = \frac{\log 55.5 - 1 + 1/55.5}{3} > 1.$$

This event covers both signs, since it depends only on squared observations. Applying the interior and upper-exterior bounds also to $-X$ gives a total upper bound $\delta^\kappa + 2\sum_{j=0}^{m_{\text{grid}}}\exp[-c\{3+s_j(r-\sqrt{6})-p_j\}]$ for the two-sided error on attainable calibrations. Define the deterministic radius parameter r_δ by

$$\delta^\kappa + 2\sum_{j=0}^{m_{\text{grid}}} e^{-c\{3+s_j(r_\delta-\sqrt{6})-p_j\}} = \delta. \tag{8}$$

The left side is continuous and strictly decreasing in r, exceeds δ at $r = \sqrt{6}$, and tends to $\delta^\kappa < \delta$. Thus $r_\delta > \sqrt{6}$ exists uniquely.

If calibration is unattainable, deleting all nonzero deviations at the center μ returns μ exactly, so this case cannot cause a positive error. On attainable calibrations, rescaling gives $T_c(W,\mu)-\mu = \sigma H_{\gamma_*}/k$. We have therefore proved the oracle-center bound

$$\Pr\{|T_c(W,\mu)-\mu| > r_k^{\mathrm{or}}(\sigma,\alpha)\} \le \alpha/2, \qquad r_k^{\mathrm{or}}(\sigma,\alpha) = \sigma\sqrt{c/k}\, r_{\alpha/2}.$$

When $\sigma = 0$, all centered increments vanish almost surely and the output is exact.

Expansion of the oracle-center radius. Dividing Eq. (8) by $\delta = e^{-3c}$ gives $2\sum_j e^{-c\{s_j(r_\delta-\sqrt{6})-p_j\}} = 1-\delta^{\kappa-1}$. Let $s_{\min} = \min_j s_j$ and $p_{\max} = \max_j p_j$. Substitution in this equation yields

$$0 < r_\delta - \sqrt{6} \le \frac{p_{\max} + c^{-1}\log\{2(m_{\mathrm{grid}}+1)/(1-\delta^{\kappa-1})\}}{s_{\min}}.$$

Indeed, at the proposed upper bound each summand is at most $(1-\delta^{\kappa-1})/[2(m_{\mathrm{grid}}+1)]$. The explicit parameter formulas give a universal positive lower bound for $s_{\min}$ and universal upper bounds for a_j and $|\eta_j|$. Moreover, $m_{\mathrm{grid}} = O(\ell)$ and $e_j \le e^{1/\ell}-1 = O(1/\ell)$ as $\ell\to\infty$. Hence $p_{\max} = O(1/\ell)$, and

$$r_\delta = \sqrt{6} + O(\log\ell/\ell), \qquad r_k^{\mathrm{or}}(\sigma,\alpha) = \sigma\sqrt{\frac{2\ell}{k}}\{1+O(\log\ell/\ell)\}.$$

All constants here are independent of the adaptive process and its variance envelope.

C.3 Replacing the true mean with the pilot

Calibration and continuity at its boundaries. For a deviation vector $u\in\mathbb{R}^k$, let $N(u) = |\{i : u_i \ne 0\}|$ and $D_u(\gamma) = \sum_i \min(\gamma u_i^2, 1)$. If $0 < c < N(u)$, this continuous function increases strictly until it reaches $N(u)$, so $D_u(\gamma) = c$ has a unique finite solution. If $N(u)\le c$, all nonzero deviations receive weight one; this is exact calibration when $N(u) = c$ and the prescribed fallback when $N(u) < c$. Define $H_c(u) = \sum_i u_i(1-d_i(u))$, so $H_c(u) = 0$ in both of these cases. At $c = 0$ take $H_0(u) = \sum_i u_i$.

For completeness, H_c is continuous even at an unattainable-calibration boundary. If $N(u) > c$, select finite $\gamma_- < \gamma_+$ with $D_u(\gamma_-) < c < D_u(\gamma_+)$; nearby calibration solutions are bracketed by these values and converge to the unique solution at u. If $N(u)\le c$ and $u^{(r)}\to u$, any subsequence using fallback has output zero. On a subsequence with exact calibration, either the calibrated parameters have a finite convergent further subsequence or they tend to infinity. In the finite case, continuity of D_u forces $D_u(\gamma) = c$; this is impossible when $N(u) < c$, and when $N(u) = c$ it deletes every nonzero coordinate. In the infinite case every nonzero limiting coordinate is eventually saturated, while all other coordinates tend to zero. Thus $H_c(u^{(r)})\to 0$ in both cases, proving continuity.

Sensitivity to the center. Fix a realized vector w and write $u_i = w_i - a$, so $T_c(w,a) = a + H_c(u)/k$. On a smooth calibrated branch, let U be the unsaturated indices, n_{sat} the number of saturated indices, and $D = \sum_{i\in U} u_i^2 > 0$. Differentiating $\gamma D = c - n_{\mathrm{sat}}$ gives $\partial_a\gamma = 2\gamma(\sum_U u_i)/D$, and hence

$$k\partial_a T_c(w,a) = n_{\mathrm{sat}} + 3\gamma D - 2\gamma\frac{(\sum_U u_i)(\sum_U u_i^3)}{D}.$$

Here $n_{\mathrm{sat}} + 3\gamma D = 3c - 2n_{\mathrm{sat}} \in [0,3c]$, $|\sum_U u_i| \le \sqrt{kD}$, and $|\sum_U u_i^3| \le D/\sqrt{\gamma}$. Therefore $|\partial_a T_c(w,a)| \le L_0 := 3c/k + 2\sqrt{c/k}$.

To extend this bound beyond smooth branches, on any compact interval of centers the coordinates $w_i - a$ are affine and their squared-ordering and saturation comparisons are polynomial inequalities after clearing the positive denominator D. There are finitely many possible active sets and finitely many nonidentical polynomial roots; identically tied boundaries have the same continuous output. Calibration-fallback centers are among the finitely many values w_i, since $c < k$. Thus the interval can be split into finitely many smooth pieces, and continuity proved above allows the derivative bound to be integrated across their endpoints. Consequently

$$|T_c(w,a) - T_c(w,a')| \le L_0|a-a'| \quad \text{for every } a,a'\in\mathbb{R}.$$

For $c = 0$, $T_0(w,a) = \bar{w}$ is independent of a and the same bound holds with $L_0 = 0$.

Finite-sample bound with the pilot. For $v_\star = k\sigma^2$, use $\delta = \alpha/2$, $c = \ell/3 < k$, $b = \min(k, \lceil \ell \rceil)$ and $s = \lfloor k/b \rfloor$. Let a_ρ be the constant in Lemma 2 with $\rho = \alpha/2$, and set $r_k^{\mathrm{or}}(0, \alpha) = 0$. The oracle-center event and pilot event have respective failure probabilities at most $\alpha/2$ and $\alpha/2$. On their intersection, the deterministic center bound gives

$$|T_c(W, P_W) - \mu| \le |T_c(W, \mu) - \mu| + L_0 |P_W - \mu| \le r_k^{\mathrm{or}}(\sigma, \alpha) + L_0 \frac{\sqrt{a_\rho k \sigma^2 / b}}{\lfloor k/b \rfloor}.$$

No independence between these events is required. Taking a union bound proves

$$\Pr\left\{ |T_c(W, P_W) - \mu| > r_k^{\mathrm{or}}(\sigma, \alpha) + L_0 \frac{\sqrt{a_\rho k \sigma^2 / b}}{\lfloor k/b \rfloor} \right\} \le \alpha. \tag{9}$$

Conclusion of Theorem 2. Suppose $\ell = \log(2/\alpha) \to \infty$ and $\ell/k \to 0$. For all sufficiently large k, $b = \lceil \ell \rceil \le k$, $c < k$, and $\lfloor k/b \rfloor \ge k/(2b)$. Since $\log(1/\rho)/b = \ell/\lceil \ell \rceil \le 1$, the increasing function h in Lemma 2 gives $a_\rho \le a_*$, where $h(a_*) = 1$ is a universal constant. It follows that

$$\frac{\sqrt{a_\rho k \sigma^2 / b}}{\lfloor k/b \rfloor} \le 2\sigma \sqrt{\frac{a_* b}{k}} = O\left(\sigma \sqrt{\frac{\ell}{k}} \right), \qquad L_0 = \frac{\ell}{k} + 2\sqrt{\frac{\ell}{3k}}.$$

Their product is $O(\sigma \ell / k)$, since $\ell/k \to 0$. Combining this with the oracle-radius expansion in Eq. (9) gives

$$\Pr\left\{ |T_c(W, P_W) - \mu| \le \sigma \sqrt{\frac{2\ell}{k}} \left[1 + O\left(\frac{\log \ell}{\ell} \right) + O\left(\sqrt{\frac{\ell}{k}} \right) \right] \right\} \ge 1 - \alpha.$$

More explicitly, there are universal constants C_1, C_2 such that the bracket can be replaced, for all sufficiently large k, by $1 + \eta_k$, where $\eta_k = C_1 \log(\ell)/\ell + C_2 \sqrt{\ell/k} \to 0$. This remainder is deterministic and independent of the process. Substituting $\ell = \log(2/\alpha)$ proves exactly

$$\Pr\left\{ |T_c(W, P_W) - \mu| \le (1 + o(1)) \sigma \sqrt{\frac{2 \log(2/\alpha)}{k}} \right\} \ge 1 - \alpha,$$

which is Eq. (6), including its asserted uniformity. If $\sigma = 0$, every $W_t = \mu$ almost surely and the estimator is exact. The finite bound (9) is valid whenever $c < k$ with the stated truncated block count; the sharp expansion specifically requires the two limiting conditions above.

C.4 Extension to fitted surrogates

For a fixed partition and mass, write $\mathrm{LV}_c(z) = T_c(z, p(z))$. For deviations u from a fixed center, on each smooth branch,

$$\partial_i H_c = 1 - 3\gamma u_i^2 + 2\gamma u_i \frac{\sum_U u_j^3}{D} \quad (i \in U), \qquad \partial_i H_c = 0 \quad (i \notin U).$$

The triangle inequality, $\sum_U |u_i| \le \sqrt{kD}$, $|\sum_U u_i^3| \le D/\sqrt{\gamma}$ and $\gamma D \le c$ give $\|\nabla H_c\|_1 \le k + 3c + 2\sqrt{kc} = k(1 + L_0)$. For $c = 0$, the gradient is the all-ones vector and this bound still holds. Along any line segment in deviation space, there are finitely many algebraic active-set pieces, as in the center argument above. A piece lying entirely in $N(u) \le c$ has $H_c = 0$; on the other pieces the displayed derivatives apply to the unsaturated coordinates, with tied saturation boundaries sharing the same continuous value. Continuity at support degeneracies was proved above. Integrating along the segment yields $|H_c(u) - H_c(u')| \le k(1 + L_0)\|u - u'\|_\infty$.

Each block mean is 1-Lipschitz in the input sup norm. Order statistics preserve coordinatewise bounds, so their median, including the average of the middle two, is also 1-Lipschitz: $|p(z) - p(z')| \le \|z - z'\|_\infty$. Inserting $T_c(z, p(z'))$ between the two outputs bounds the change in center by $L_0\|z - z'\|_\infty$ and the change in input at that fixed center by $(1 + L_0)\|z - z'\|_\infty$. Thus

$$|\mathrm{LV}_c(z) - \mathrm{LV}_c(z')| \le (1 + 2L_0)\|z - z'\|_\infty. \tag{10}$$

For the contributions $Z_t(f)$ in Eq. (1), every fixed f gives common conditional mean μ under WR or WOR, so the adapted-sequence result applies with $W_t = Z_t(f)$. On the same logged trajectory, $\max_t |Z_t(f) - Z_t(g)| \le (1 + 1/\tau)\|f - g\|_\infty$. This is a deterministic comparison on one logged trajectory: it reuses that trajectory's propensities. The probability theorem itself remains under the original adaptive law; conditioning on the entire propensity path would be a different argument, and replaying acquisition would define a different experiment.

Let $\mathcal{G}$ be an ϵ-net of size j fixed before the target observations. Suppose every $g \in \mathcal{G}$ has deterministic variance envelope $k\sigma_g^2$, with $\sigma_g \ge 0$. Apply the preceding construction at total failure α/j: use common $\delta = \alpha/(2j)$, $c = \log(2j/\alpha)/3 < k$, and $b = \min(k, \lceil \log(2j/\alpha) \rceil)$ with one common deterministic block partition. A union bound in Eq. (9) and Eq. (10) simultaneously give, for every $f \in \mathcal{F}$,

$$|\operatorname{LV}_c(Z(f)) - \mu| \le r_k^{\mathrm{or}}(\sigma_g, \alpha/j) + L_0 \frac{\sqrt{a_\rho k \sigma_g^2 / b}}{\lfloor k/b \rfloor} + (1 + 2L_0)(1 + 1/\tau)\epsilon,$$

where g is any net point within ϵ of f and a_ρ uses $\rho = \alpha/(2j)$. Evaluating this simultaneous bound at $f = F$ and $g = G_F$ covers the same-sample terminal fit. With $\ell = \log(2j/\alpha) \to \infty$ and $\ell/k \to 0$, the preceding radius expansion is uniform over the net points, and $L_0 = O(\sqrt{\ell/k} + \ell/k) = o(1)$. Therefore, on an event of probability at least $1 - \alpha$,

$$|T_c(Z(F), P_F) - \mu| \le (1 + o(1))\sigma_{G_F}\sqrt{\frac{2\ell}{k}} + (1 + o(1))(1 + 1/\tau)\epsilon.$$

This is the fitted-surrogate conclusion stated after Theorem 2. The selection of G_F is made only after establishing the simultaneous event; we never assume that $Z_t(F)$ has conditional mean μ. The bounded IRT/factor parameter balls established earlier provide such covers; at $\epsilon = 1/k$ the entropy cost is $O(d \log k)$. The sharp leading term uses $\ell = \log(2j/\alpha)$ and requires $\ell \to \infty$ and $\ell/k \to 0$. For the full terminal-class radius to have that same leading constant, the displayed net-approximation term must also be $o(\sigma_g\sqrt{\ell/k})$; this does not follow automatically when σ_g tends to zero.

C.5 Numerical implementation

Theorem 2 assumes exact calibration. To quantify an error in the attained calibration mass, fix the contribution vector z and pilot p. Suppose the computed weights satisfy $\widetilde{d}_t = \min\{\widetilde{\gamma}(z_t - p)^2, 1\}$ for a common $\widetilde{\gamma} \ge 0$, and let $\widetilde{c} = \sum_t \widetilde{d}_t$. If both c and $\widetilde{c}$ are attainable, monotonicity of the weights gives $|T_{\widetilde{c}}(z, p) - T_c(z, p)| \le k^{-1} \max_t |z_t - p|\,|\widetilde{c} - c|$. If the requested mass is unattainable, the algorithm uses the fallback defined in Section 4. This bound controls calibration-mass error with the contributions and pilot held fixed; errors in those inputs or in evaluating the final sum require separate bounds.

D Additional experimental results

D.1 Per-benchmark results at the 5% budget

Table 9 supports the main-text comparison at the 5% budget: ASA-LV has lower RMSE than FAQ, AISCV regression and RS mean on each of the nineteen benchmarks. These values use all 974 targets and twenty query repetitions from the same local IRT evaluation as Table 1. The final row averages benchmark RMSEs equally and reproduces the corresponding main-table values.

Uncertainty in average RMSE differences. Table 10 reports the difference between each baseline's equal-benchmark mean RMSE and that of ASA-LV. All twenty marginal 95% bootstrap intervals lie above zero, supporting lower average RMSE for ASA-LV against each listed baseline in both evaluation settings at both budgets.

We use 3,999 target-stratified bootstrap resamples of the twenty query repetitions. Within each target, the same resampled repetition indices are used across methods; each resample recomputes

Table 9: Per-benchmark RMSE at the 5% query budget with the local IRT surrogate. Each value is computed over all target models and twenty query repetitions within that benchmark. Lower values are better.

Benchmark	ASA-LV	FAQ	AISCV regression	RS mean
ARC-Challenge	**.038767**	.040294	.043624	.061370
BBH	**.022542**	.022933	.022844	.027965
Commonsense	**.058545**	.062636	.080449	.063958
GPQA	**.051245**	.052551	.052840	.057398
GSM	**.048147**	.048573	.052418	.058714
IFEval	**.060876**	.064034	.072880	.080623
ImageNet	**.004106**	.004129	.004149	.008208
LegalBench	**.032649**	.034054	.034761	.045078
MATH	**.071335**	.080228	.094018	.076231
MedQA	**.052928**	.055145	.057305	.068000
MMLU	**.065311**	.069299	.074292	.091006
MMLU-Pro	**.012782**	.012870	.012798	.017416
MNLI	**.013330**	.013820	.013442	.021614
MRPC	**.061412**	.074784	.076093	.098728
MuSR	**.057387**	.060615	.065509	.076911
OpenLLM MATH	**.024926**	.025743	.027344	.029621
QNLI	**.015319**	.015609	.015947	.028765
RTE	**.051498**	.091004	.087412	.124978
SST-2	**.032277**	.038193	.036649	.057837
Equal-benchmark mean	**.040810**	.045606	.048672	.057601

RMSE within each benchmark before averaging benchmarks equally. The 2.5th and 97.5th percentiles give the reported intervals. These approximate intervals describe query variability conditional on the fixed histories, benchmarks and target models. They are per-comparison intervals, not simultaneous confidence guarantees or intervals for performance on new models or benchmarks.

Table 10: Uncertainty in equal-benchmark mean RMSE differences. Each entry gives baseline RMSE minus ASA-LV RMSE, followed by its marginal 95% bootstrap interval. All values are in units of 10^{-3}; positive differences favor ASA-LV.

	19 benchmarks (IRT)		FAQ data (released factors)	
Baseline	5%	10%	5%	10%
FAQ	4.797 [4.360, 5.250]	1.999 [1.750, 2.253]	2.007 [1.960, 2.054]	0.930 [0.904, 0.954]
AIPW-B	8.715 [6.804, 10.874]	3.076 [1.980, 4.190]	3.284 [3.168, 3.398]	1.740 [1.666, 1.814]
AIPW-U	6.915 [5.832, 7.971]	2.433 [1.655, 3.198]	3.458 [3.366, 3.547]	1.827 [1.768, 1.888]
AISCV regression	7.863 [6.908, 8.847]	2.166 [1.845, 2.477]	5.699 [5.631, 5.768]	3.263 [3.221, 3.307]
RS mean	16.792 [15.606, 17.942]	11.209 [10.409, 12.031]	8.209 [8.100, 8.318]	5.842 [5.771, 5.914]

Uncertainty in MAE and model-ranking differences. Tables 11–13 supplement the RMSE comparison with MAE and model-ranking uncertainty. The estimates use the same complete target cohorts and twenty query repetitions as the main tables; reconstructed squared errors agree with the archived primary losses. For each repetition, we compute MAE or a ranking correlation within each benchmark and then average benchmarks equally. We form paired differences between ASA-LV and each baseline within that repetition. If these twenty differences are $D_1, \ldots, D_{20}$, the estimated standard error is $s_D/\sqrt{20}$ and the interval is $\bar{D} \pm t_{19,0.975} s_D/\sqrt{20}$, where s_D is their sample standard deviation and $t_{19,0.975}$ is the Student-t quantile. This repetition-level calculation retains the joint model ranking within each run; it does not resample model identities. Unlike the target-stratified

RMSE bootstrap above, it treats each complete query repetition as the unit of analysis and uses a Student-t approximation.

All sixty marginal intervals are above zero. Thus the observed advantages in MAE and both ranking metrics are supported by the variation across query repetitions in these evaluations. These approximate intervals condition on the fixed historical data, target models and benchmarks; they do not provide simultaneous coverage, guarantees for new target cohorts, or the finite-sample confidence bounds proved in the theoretical sections. Ranking remains an exploratory outcome.

Table 11: MAE differences and marginal 95% paired repetition-t intervals. Differences are baseline minus ASA-LV. All values are in units of 10^{-3}; positive differences favor ASA-LV.

	19 benchmarks (IRT)		FAQ data (released factors)	
Baseline	5%	10%	5%	10%
FAQ	4.337 [3.773, 4.902]	1.859 [1.618, 2.101]	1.569 [1.538, 1.601]	0.744 [0.723, 0.765]
AIPW-B	6.546 [5.583, 7.509]	2.148 [1.564, 2.732]	2.660 [2.598, 2.721]	1.428 [1.388, 1.469]
AIPW-U	6.538 [5.662, 7.414]	2.368 [1.732, 3.004]	2.913 [2.840, 2.987]	1.570 [1.521, 1.618]
AISCV regression	5.884 [5.304, 6.465]	1.935 [1.675, 2.195]	4.747 [4.702, 4.792]	2.739 [2.701, 2.776]
RS mean	14.589 [13.425, 15.754]	9.835 [9.306, 10.364]	6.735 [6.656, 6.814]	4.822 [4.733, 4.912]

Table 12: Kendall τ_b differences and marginal 95% paired repetition-t intervals. Differences are ASA-LV minus baseline. All values are in units of 10^{-3}; positive differences favor ASA-LV.

	19 benchmarks (IRT)		FAQ data (released factors)	
Baseline	5%	10%	5%	10%
FAQ	17.149 [14.318, 19.981]	11.019 [7.959, 14.078]	11.401 [11.096, 11.707]	5.926 [5.705, 6.146]
AIPW-B	24.045 [16.486, 31.603]	11.691 [5.510, 17.872]	18.846 [18.154, 19.537]	11.139 [10.703, 11.576]
AIPW-U	26.549 [18.352, 34.746]	12.164 [5.482, 18.846]	21.128 [20.450, 21.806]	12.808 [12.346, 13.270]
AISCV regression	16.418 [11.245, 21.591]	4.927 [1.363, 8.490]	20.294 [19.952, 20.637]	15.681 [15.337, 16.024]
RS mean	52.848 [43.233, 62.463]	43.885 [37.412, 50.357]	45.533 [44.884, 46.182]	36.652 [35.747, 37.558]

Sensitivity to benchmark composition. At the 5% budget, RTE contributes 43.35% of the summed absolute RMSE reduction relative to FAQ; RTE, MRPC, MATH and SST-2 together contribute 74.27%. Thus improvement directions agree across benchmarks, but improvement magnitudes are heterogeneous. Removing RTE leaves relative reductions in equal-benchmark mean RMSE of 6.66% and 3.99% at the 5% and 10% budgets. Removing all four leaves reductions of 4.03% and 2.74%. The corresponding absolute differences and marginal 95% bootstrap intervals are 0.002868 $[0.002491, 0.003235]$ and 0.001179 $[0.000957, 0.001396]$ after removing RTE, and 0.001563 $[0.001284, 0.001833]$ and 0.000737 $[0.000567, 0.000902]$ after removing all four. These sensitivity intervals use 3,999 target-stratified paired repetition resamples, with benchmarks retained and reweighted equally after exclusion (random seed 260926). Across all nineteen benchmarks, the median per-benchmark relative RMSE reduction is 4.02% at the 5% budget and 2.59% at the 10% budget. The exclusion analyses were motivated by inspection of the per-benchmark results and describe sensitivity rather than a separately held-out evaluation.

Table 13: Spearman ρ differences and marginal 95% paired repetition-t intervals. Differences are ASA-LV minus baseline. All values are in units of 10^{-3}; positive differences favor ASA-LV.

	19 benchmarks (IRT)		FAQ data (released factors)	
Baseline	5%	10%	5%	10%
FAQ	14.922 [12.484, 17.360]	8.618 [6.210, 11.026]	2.590 [2.500, 2.680]	0.947 [0.903, 0.992]
AIPW-B	19.850 [13.515, 26.185]	9.469 [4.430, 14.507]	4.340 [4.148, 4.532]	1.808 [1.709, 1.906]
AIPW-U	20.498 [13.366, 27.631]	8.789 [3.495, 14.083]	4.950 [4.743, 5.157]	2.132 [2.021, 2.244]
AISCV regression	12.934 [8.354, 17.514]	3.492 [0.916, 6.068]	4.490 [4.390, 4.590]	2.566 [2.493, 2.639]
RS mean	50.036 [41.477, 58.595]	36.812 [31.550, 42.073]	12.897 [12.625, 13.168]	7.631 [7.405, 7.857]

D.2 Comparison with other robust estimators

This comparison changes only the final aggregation of the surrogate-augmented contributions from the main clean evaluation. Predictor fits, query indices, propensities, budgets, targets and twenty repetitions are identical across methods. It retains all 974 19-benchmark targets and all 4425 released-factor target rows. For a contribution vector $z = (z_1, \ldots, z_k)$, let p be the median of the three consecutive block means and define

$$s = \begin{cases} 1.4826 \ \mathrm{median}_t \, |z_t - p|, & \text{if this is positive and finite,} \\ \{k^{-1} \sum_t (z_t - p)^2\}^{1/2}, & \text{otherwise,} \end{cases}$$

with machine epsilon used only when both expressions vanish. MOM returns p. At multiplier h, Winsor returns $p + k^{-1} \sum_t \mathrm{clip}(z_t - p, -hs, hs)$, and Huber is the root of $\sum_t \mathrm{clip}(z_t - r, -hs, hs) = 0$. The Catoni comparator is the root of $\sum_t \psi\{\eta_{\mathrm{C}}(z_t - r)\} = 0$, where $\eta_{\mathrm{C}} = h/(s\sqrt{k})$ and

$$\psi(u) = \begin{cases} \log(1 + u + u^2/2), & u \geq 0, \\ -\log(1 - u + u^2/2), & u < 0. \end{cases}$$

Both roots use eighty bisection steps on $[\min z, \max z]$. At each step, a positive score moves the lower endpoint to the midpoint; a nonpositive score moves the upper endpoint to the midpoint. The returned estimate is the midpoint of the final bracket, which also specifies the choice when the Huber score has an interval of roots. A constant vector therefore returns its constant. These equations, scale conventions and fixed iteration count uniquely specify the empirical comparators.

We selected one tuning multiplier per method using both budgets and three query repetitions on a development set of 190 target model–benchmark pairs: the first ten target filenames in lexicographic order from each of the 19 benchmarks. These targets also occur in the main evaluation, but the query runs used for tuning are separate. Thus the development and evaluation sets are not disjoint at the target level. No FAQ evaluation runs are used to select the tuning multipliers. The initial grids were $\{.5, 1, 1.5, 2, 3, 4\}$ for Winsor and Huber and $\{.25, .5, 1, 2\}$ for Catoni; exact ties favored less correction. The selected multipliers were 4, 4 and .25. Because all three selected multipliers lay on their respective grid boundaries, after examining that output we expanded Winsor/Huber to $\{.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 32, 64\}$ and Catoni to $\{.03125, .0625, .125, .25, .5, 1, 2\}$. The expanded choices are 64, 64 and .03125. The grids were expanded after inspecting the initial results, so the expanded-grid comparison is a tuning-sensitivity analysis. All three selected multipliers produce estimates close to the arithmetic mean.

Table 3 reports the complete aggregate result. Only the aggregation rule changes in this empirical comparison; standard guarantees for independent observations do not automatically apply to these adaptively collected, fitted-surrogate contributions.

Uncertainty calculation. The intervals in Table 15 use a paired delta-method normal approximation, rather than the percentile bootstrap used in Table 10. For each benchmark and method,

RMSE is the square root of the mean squared error over fixed targets and twenty query repetitions. To estimate the variance of the RMSE difference, squared errors are linearized using the derivative $1/(2\,\mathrm{RMSE})$. The baseline and ASA-LV terms are subtracted within the same target and repetition before estimating variance across repetitions, retaining their empirical covariance. With B benchmarks, N_h fixed targets in benchmark h, and $R = 20$ repetitions, each target's linearized difference receives weight $1/(BN_h)$. Its sample variance across repetitions is therefore multiplied by $1/(RB^2N_h^2)$, and these terms are summed over targets and benchmarks. This calculation assumes independent query randomization across targets. Each interval is the estimated difference plus or minus 1.96 standard errors. Targets and histories remain fixed; these are approximate marginal intervals for query variability, not simultaneous guarantees or intervals for new target populations.

Table 14: Expanded-grid tuning on the 190-target development set. Development RMSE gives every local benchmark and budget equal weight.

Family	Initial choice	Expanded choice	Development RMSE
Winsor	4	64	.03239487
Huber	4	64	.03239469
Catoni	.25	.03125	.03239431

Table 15: ASA-LV against the strongest expanded-grid alternative in each primary setting and budget. Gains and paired delta-normal 95% intervals are absolute RMSE reductions, conditional on the fixed targets and histories.

Setting	Budget	Comparator	Gain	Paired 95% interval
19-benchmark IRT	5%	Catoni	.00027446	[.00000222, .00054670]
19-benchmark IRT	10%	Catoni	.00022219	[.00003867, .00040571]
FAQ factors	5%	Winsor	.00016546	[.00015747, .00017345]
FAQ factors	10%	Winsor	.00005395	[.00005074, .00005715]

The aggregate comparison does not mean every benchmark benefits. At both budgets, ASA-LV has lower RMSE in 11 of 19 local benchmarks than Winsor, Huber and Catoni, and in 19 of 19 than MOM. With released factors it wins in both benchmarks against every comparator. Thus the equal-benchmark aggregate advantage is not a claim of per-benchmark dominance.

Calibration sensitivity. The primary configuration uses three chronological pilot blocks and mass $c = \log(10)/3$. Replaying the same contributions at the smaller $c = \log(2)/3$ gives RMSE .040853/.028992 on 19-benchmark IRT and .011550/.007732 with released factors at 5%/10%; the primary values are .040810/.028906 and .011437/.007694. The primary calibration mass gives lower RMSE than the smaller mass in both evaluation settings at both budgets. This is a two-point sensitivity check, not a data-driven calibration rule.

D.3 Historical-mean and regression surrogates

The historical-mean regime is retained as a fixed-surrogate reference rather than pooled with learned terminal predictors. Within that regime, ASA-LV improves on the matched FAQ estimate and on RS in RMSE and ranking. The retained AIPW-B/U rows compare uncertainty and uniform Bernoulli allocation. The regression-terminal comparison is therefore reported as a separate estimator family. It reuses the same historical-mean query trajectories and the fixed AISCV feature span. Logistic Regression–ASA-LV fits an inverse-propensity-weighted logistic ridge predictor with an unpenalized intercept and bounded sigmoid predictions, forms surrogate-augmented contributions, and applies ASA-LV. AISCV regression fits its control-variate estimator on the same logged trajectories. No query is regenerated for this comparison. Here both regression methods use the historical mean, item intercept and four locally fitted IRT coordinates as their feature inputs, including on the FAQ data. This differs from the released-factor features used by AISCV regression in the main FAQ comparison. Appendix A.4 describes feature preprocessing and regression fitting.

Tables 16 and 17 keep the terminal-surrogate families separate. The Logistic Regression–ASA-LV row has lower RMSE and MAE and higher Kendall τ_b and Spearman ρ than AISCV regression in both data regimes and at both budgets.

On the 19-benchmark collection with a fixed historical-mean surrogate, ASA-LV has lower RMSE but slightly higher MAE than FAQ at both budgets. This pattern is consistent with reducing some large estimation errors while increasing some smaller errors, because RMSE gives greater weight to large deviations; the aggregate metrics alone do not establish this mechanism. Logistic Regression–ASA-LV further reduces both RMSE and MAE relative to fixed-mean ASA-LV in both evaluation settings at both budgets. With the same historical features and queried observations available to both regression methods, it also outperforms AISCV regression on all four reported metrics throughout these comparisons.

The RS mean values on the FAQ data differ from those in Table 1 because these evaluations use different random-sample realizations. The historical-mean evaluation generates uniform permutations with the local-surrogate random-stream keys, whereas the released-factor evaluation uses separate released-factor keys. Both report the unadjusted mean of a uniform sample without replacement with the same nominal budget and target cohort; RS itself does not use a surrogate. The archived equal-benchmark RMSE values are $0.019768/0.013586$ here and $0.019646/0.013536$ in the released-factor evaluation at the 5%/10% budgets. These differences do not represent an effect of surrogate choice on RS.

Table 16: Historical-mean and regression-terminal error. RMSE and mean absolute error (MAE) are equal-benchmark averages; lower is better. "Fixed mean" and "regression" are separate terminal-surrogate families.

Regime	Terminal family	Method	RMSE		MAE	
			5%	10%	5%	10%
19 benchmarks	Fixed mean	ASA-LV	.052190	.036106	.040540	.028129
		FAQ	.052748	.036436	.039581	.027812
		AIPW-B	.055768	.038368	.042185	.028739
		AIPW-U	.055131	.038315	.043005	.029895
		RS mean	.057601	.040115	.045464	.031699
	Regression	**Logistic Regression–ASA-LV**	**.046448**	**.030393**	**.035293**	**.022896**
		AISCV regression	.048773	.031130	.036811	.023681
FAQ data	Fixed mean	ASA-LV	.016405	.011593	.013067	.009212
		FAQ	.016414	.011596	.013070	.009213
		AIPW-B	.016934	.011651	.013380	.009194
		AIPW-U	.017382	.011987	.013819	.009528
		RS mean	.019768	.013586	.015672	.010773
	Regression	**Logistic Regression–ASA-LV**	**.014617**	**.010254**	**.011604**	**.008121**
		AISCV regression	.014911	.010451	.011854	.008293

D.4 Covariance-aware query policy

This experiment compares uncertainty-only and covariance-aware querying on the 19-benchmark collection, using all 974 target model–benchmark pairs, twenty query repetitions, and the 5% and 10% budgets. Both policies use the same local IRT fitting procedure and ASA-LV aggregation. Within each target and repetition, they share the initial predictor and random-number stream, but generate their own adaptive query trajectories.

Curvature-based covariance approximation. Let $\phi_i \in \mathbb{R}^5$ contain question i's four fixed historical IRT features and an intercept coordinate. Immediately before query t, define the random predictive-variance term $W_{t,i} = F_{t-1}(i)\{1 - F_{t-1}(i)\}$. Let $\mathcal{J}_{t-1}$ be the set of distinct queried questions used in the most recent scheduled predictor fit, and let $M_{t-1} = |\mathcal{J}_{t-1}|$. We construct

$$\Sigma_{t-1} = \left[\sum_{j \in \mathcal{J}_{t-1}} W_{t,j} \phi_j \phi_j^\top + \lambda \max\{M_{t-1}, 1\} I_5 \right]^{-1}, \qquad \lambda = 0.05. \tag{11}$$

For $M_{t-1} > 0$, the precision matrix inside the brackets is the logistic curvature for the sum of the fitted observations' losses, plus $\lambda M_{t-1} I_5$. The latter scaling corresponds to multiplying the

Table 17: Historical-mean and regression-terminal model-ranking correlations. Kendall τ_b and Spearman ρ are computed within each benchmark and query repetition, then averaged over repetitions and equally across benchmarks; higher is better.

Regime	Terminal family	Method	Kendall τ_b		Spearman ρ	
			5%	10%	5%	10%
19 benchmarks	Fixed mean	ASA-LV	.740286	.794145	.868019	.905404
		FAQ	.740195	.793602	.867758	.905183
		AIPW-B	.732951	.794263	.863827	.908386
		AIPW-U	.723364	.783711	.857284	.901741
		RS mean	.715675	.777396	.840217	.889446
	Regression	**Logistic Regression–ASA-LV**	**.758928**	**.822301**	**.881875**	**.927603**
		AISCV regression	.749508	.816901	.874154	.924260
FAQ data	Fixed mean	ASA-LV	.893993	.920388	.984781	.991185
		FAQ	.893871	.920338	.984755	.991176
		AIPW-B	.893226	.921508	.984633	.991466
		AIPW-U	.889988	.918813	.983733	.990899
		RS mean	.876158	.908253	.978284	.987843
	Regression	**Logistic Regression–ASA-LV**	**.903140**	**.927997**	**.987131**	**.992703**
		AISCV regression	.901481	.926718	.986744	.992475

mean-loss ridge objective by M_{t-1} before forming its curvature. We invert this matrix to obtain a curvature-based approximation to parameter covariance. Before any target answers are fitted, $\Sigma_0 = (\lambda I_5)^{-1}$. The predictor fitting procedure remains unchanged. The predictor and covariance approximation are refreshed after every four new distinct queried answers, and remain fixed between refreshes. Repeated queries do not add duplicate observations to the fit or the curvature sum.

Sampling scores. The uncertainty-only score is $S_t^{\mathrm{unc}}(i) = \sqrt{W_{t,i}}$. The covariance-aware score is $S_t^{\mathrm{cov}}(i) = \sqrt{W_{t,i} + W_{t,i}^2 \phi_i^\top \Sigma_{t-1} \phi_i}$. The added term approximates prediction variance due to uncertainty in the fitted parameters, using the logistic slope factor $W_{t,i}$. Its coefficient is fixed at one. For either score, queries are sampled with replacement using $Q_t(i) = \tau/n + (1 - \tau)S_t(i)/\sum_{j=1}^n S_t(j)$, with $\tau = 0.01$; if all scores vanish, sampling is uniform. Query probabilities are refreshed with the predictor and covariance approximation. Both policies use the last scheduled predictor to reconstruct the final contributions with their own logged propensities, followed by ASA-LV aggregation with $b = 3$ and $c = \log(10)/3$.

Results. Table 18 reports RMSE and MAE computed over targets and repetitions within each benchmark and then averaged equally across benchmarks. Covariance-aware querying reduces RMSE by 1.11% and 2.86%, and MAE by 1.79% and 2.65%, at the 5% and 10% budgets, respectively. Per-benchmark RMSE is lower on 11 of 19 benchmarks at the 5% budget and 15 of 19 at the 10% budget. These results show an aggregate benefit from incorporating the curvature-based covariance term, with variation across benchmarks. The primary tables retain uncertainty-only querying; the improvement here comes from changing acquisition while retaining the same surrogate-fitting and aggregation procedures.

Table 18: Query-policy ablation on the 19-benchmark collection with all 974 target model–benchmark pairs and twenty repetitions. Entries are equal-benchmark mean RMSE/MAE; lower values are better. Both policies use ASA-LV.

Query policy	5% budget	10% budget
Uncertainty-only	.040810/.030875	.028906/.021864
Covariance-aware	**.040357/.030322**	**.028079/.021285**