FUNCTIONAL AMBIGUITY REDUCTION: MINIMAX ACQUISITION UNDER OBSERVATIONAL EQUIVALENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Finite observations can leave distinct latent populations observationally indistinguishable while those populations imply different values of a target functional. We study how this target-relevant ambiguity limits recovery and how it should guide subsequent acquisition. We first characterize recoverability under a finite response design: the largest functional ambiguity is exactly twice the best uniform approximation error of the target by the observed response coordinates, and this approximation error equals the minimax recovery error of any estimator based only on the observed response vector. This characterization yields Functional Ambiguity Reduction (FAR), which selects queries according to the target disagreement that can remain among observationally compatible populations. FAR is one-step minimax-optimal for worst-case absolute target recovery under exact responses and extends to finite response batches and learned response models. Across five public datasets, FAR shows its clearest gains on CIFAR-10H and Duolingo-FR, reducing MAE on Duolingo-FR from 0.1330 for the strongest baseline to 0.1263. We further apply FAR to budgeted LLM evaluation, where five additional queries reduce score-estimation MAE by 7.0% on Qwen3-4B and 25.7% on Phi-3.5-mini relative to the strongest baseline for each model. These results connect finite-design identifiability directly to adaptive acquisition for population-level functional recovery. Code is available at `https://anonymous.4open.science/r/FAR-D5B5/`.

1 INTRODUCTION

In many learning, auditing, and scientific measurement problems, the quantities we can observe are not the quantities we ultimately care about (Bollen, 2002; Verbeke & Molenberghs, 2017). A crowdsourcing study may care about annotator-population reliability rather than individual labels; a model audit may target subgroup failure rates rather than single test outcomes; and a scientific study may seek a population-level property of hidden heterogeneity from indirect measurements (Dawid & Skene, 1979; Buolamwini & Gebru, 2018; Allman et al., 2009). Across these settings, the target depends on an underlying distribution of latent types that is accessible only through responses to selected probes. Because exhaustively characterizing this latent population is typically impractical (Settles, 2009), a natural question is: under a limited budget, which probe should be queried next to most effectively determine the target quantity of interest?

Recent work has accordingly shifted from generic predictive uncertainty toward target-aware acquisition (Filstroff et al., 2024). DGSM acquisition targets sensitivity functionals (Belakaria et al., 2024), transductive active learning focuses on designated target predictions (Hübotter et al., 2024), and Adaptive Elicitation selects questions to recover latent information and downstream responses (Wang et al., 2025). Many of these methods value queries through posterior variance, entropy, or information gain. Under partial identification, however, a finite design can leave distinct latent populations observationally compatible while implying different values of the same target functional (Figure 1a). This structural ambiguity need not coincide with either response uncertainty or posterior target uncertainty. A query can have high response uncertainty yet leave substantial target disagreement, while posterior uncertainty over compatible explanations can remain sensitive to prior or modeling choices (Moon & Schorfheide, 2012; Giacomini & Kitagawa, 2021; Zheng et al., 2022; Pace et al., 2024). Thus, reducing response or posterior uncertainty need not reduce the target ambiguity that remains supported by the observations (Figure 1b).

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

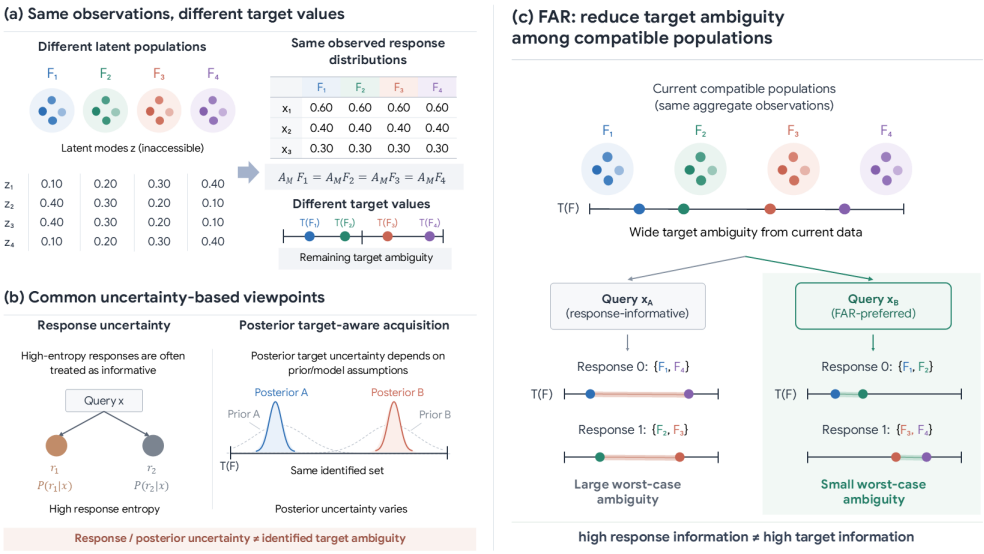


Figure 1: **Observational equivalence can hide target ambiguity.** (a) Distinct latent populations can induce the same finite observed responses while implying different values of the same target functional. (b) High response uncertainty or posterior target uncertainty need not match the identified target ambiguity that remains under the observations. (c) FAR selects queries that minimize the target disagreement remaining among response-compatible populations.

One recent approach addresses this issue more directly. TSIED (Ailer et al., 2024) adapts an indirect-experiment policy to tighten upper and lower bounds on a partially identified causal query. This shows that what is identified from the observations can itself guide acquisition. In the population-level problems above, however, the key issue is what a finite collection of responses actually determines about the target. Distinct latent populations may remain observationally indistinguishable under the acquired probes, and the relevant question is which of these unresolved differences still matter for the target. The relationship between the observation design, observational equivalence, and the target is not yet characterized in a general form.

We address this gap with Functional Ambiguity Reduction (FAR), which bases acquisition directly on target ambiguity under observational equivalence. For a given response vector on the acquired design, the compatible latent populations form an observational-equivalence fibre, whose image under the target functional is the range of target values still supported by the observations. We refer to the width of this range as functional ambiguity. FAR does not seek to identify the latent population itself. Instead, FAR evaluates each candidate probe by the target disagreement remaining among currently compatible populations with the same response distribution at that probe. It selects the probe with the smallest residual ambiguity (Figure 1c). We show that functional ambiguity has an exact recovery interpretation: its half-width equals the minimax absolute target-recovery error achievable from the observed response vector. Our contributions are:

- We characterize finite-design target recoverability under observational equivalence. Worst-case functional ambiguity is exactly twice the best uniform approximation error and, equivalently, twice the optimal minimax target-recovery error attainable from the observed response vector.
- We derive FAR from this recovery characterization: a target-specific query rule that minimizes residual functional ambiguity and is one-step minimax-optimal for absolute target recovery under exact responses. We extend the framework to finite response batches and learned response models.
- We evaluate FAR across adaptive acquisition, controlled observational-equivalence, and budgeted LLM-evaluation settings. On Duolingo-FR, FAR reduces MAE from 0.1330 to 0.1263 relative to the strongest baseline; with five additional queries, it reduces LLM score-estimation MAE by 7.0% on Qwen3-4B and 25.7% on Phi-3.5-mini.

2 Related work

Recovery under partial identification. Recent work has increasingly studied learning and evaluation when available observations do not point-identify the quantity of interest, replacing point estimates with bounds over observationally compatible models or distributions (Melnychuk et al., 2024; Penn et al., 2025). This perspective has been used to bound model-performance functionals without ground-truth labels (Maia Polo et al., 2024a), characterize causal quantities under flexible structural models (Tan et al., 2024), and study achievable robustness when robust risk is itself only partially identified (Kostin et al., 2024). Other methods tighten or compute identified sets using optimal transport, learned instrument representations, and data fusion (Gao et al., 2025; Lin et al., 2025; Schweisthal et al., 2025; Lanners et al., 2025), while related work derives prediction or decision guarantees uniformly over data-generating processes compatible with the observations (Chen et al., 2025). Recent optimal-recovery work further connects limited-information estimation to statistical minimax risk (DeVore et al., 2025). FAR instead uses the finite response design to define which latent populations remain observationally equivalent, and evaluates recoverability through their disagreement on the target functional; additional queries are valuable insofar as they eliminate target-relevant ambiguity among these compatible populations.

Version spaces and active discrimination. Version-space methods maintain hypotheses that remain compatible with observed data and select subsequent queries to eliminate competing explanations (Yan & Zhang, 2024). Recent work has developed scalable version-space algorithms for interactive model development (Palma et al., 2024), extended version-space reasoning beyond standard single-label settings (Wang et al., 2023), and adapted disagreement-based selection to multi-group objectives (Rittler & Chaudhuri, 2023). Related query-by-committee methods query where plausible hypotheses disagree, with recent work using hypothesis perturbations to make such strategies practical for deep models (Cho et al., 2024). Active testing and model discrimination similarly select tests or experimental inputs to distinguish candidate states or models, including information-maximizing binary testing and adaptive experimental-design approaches (Chattopadhyay et al., 2024; Shen et al., 2025; Niu et al., 2026). These methods generally value queries by how effectively they distinguish among compatible hypotheses. FAR instead asks which remaining distinctions matter for recovering the declared functional: compatible latent populations need not be distinguished if they imply the same target value, whereas populations that remain observationally similar but imply different target values are precisely the ambiguities that acquisition should resolve.

3 Problem formulation

Latent population and responses. Let $\mathcal{X}$ denote the space of available probes and $\mathcal{Y}$ a finite response space. An unknown distribution $F \in \mathcal{P}(\mathcal{Z})$ describes a population of latent types on a compact metric space $\mathcal{Z}$. For a probe $x \in \mathcal{X}$, the response distribution of type $z \in \mathcal{Z}$ is specified by a kernel $k_\theta(\cdot \mid x, z)$. Since the latent type is unobserved, the population response distribution is

$$q_F(y \mid x) = \int_{\mathcal{Z}} k_\theta(y \mid x, z)\, dF(z). \tag{1}$$

We condition on the response kernel in the structural analysis and assume that $z \mapsto k_\theta(y \mid x, z)$ is continuous for every x and y. The distribution $q_F(\cdot \mid x)$ describes the population's single-response probabilities and is not observed directly. A query yields a finite batch whose observation law also depends on the sampling protocol.

Target of interest. Let $h : \mathcal{Z} \to \mathbb{R}$ be a prespecified continuous target function. The corresponding population-level functional is

$$T_h(F) = \int_{\mathcal{Z}} h(z)\, dF(z). \tag{2}$$

$h(z)$ specifies the target value associated with latent type z, while $T_h(F)$ aggregates this quantity over the population. The objective is to recover $T_h(F)$ from responses to selected probes.

Observational equivalence and target ambiguity. For a finite probe design $D_M = \{x_1, \ldots, x_M\}$, define the response operator

$$\mathcal{A}_M F = [q_F(\cdot \mid x_1), \ldots, q_F(\cdot \mid x_M)]. \tag{3}$$

Two populations F and G are observationally equivalent under D_M if $\mathcal{A}_M F = \mathcal{A}_M G$. Such populations induce the same response distributions at every probe in the design but may satisfy $T_h(F) \neq T_h(G)$. Hence, even exact knowledge of the response distributions on a finite design need not determine the target. We study this structural ambiguity under exact response probabilities and later incorporate the sampling uncertainty arising from finite response batches.

Sequential acquisition task. Let $\mathcal{H}_t$ denote the acquisition history after t queried probes. An acquisition policy π maps the current history to the next probe, $x_{t+1} = \pi(\mathcal{H}_t)$. In the exact-response setting, querying x_{t+1} reveals the population response distribution $q_F(\cdot \mid x_{t+1})$. With finite batches, the query instead returns B responses sampled from this distribution. Given a finite acquisition budget, the objective is to choose probes adaptively and estimate the prespecified target $T_h(F)$ from the resulting history. Performance is measured by absolute target error.

4 METHODOLOGY

4.1 FUNCTIONAL AMBIGUITY AND RECOVERY LIMITS

Under a finite observation design, latent populations can remain observationally compatible while implying different target values. For a finite design $D_M = x_1, \ldots, x_M$, let $p_M = \mathcal{A}_M F$ denote the population response vector. The populations compatible with p_M form the observational fibre

$$\mathcal{F}_M(p_M) = \{F \in \mathcal{P}(\mathcal{Z}) : \mathcal{A}_M F = p_M\}. \tag{4}$$

This fibre induces a sharp target interval $[L_M, U_M]$, where $L_M = \inf_{F \in \mathcal{F}_M(p_M)} T_h(F)$ and $U_M = \sup_{F \in \mathcal{F}_M(p_M)} T_h(F)$. We denote its width by $W_M(p_M) = U_M - L_M$.

Theorem 1 (Fibrewise minimax recovery)**.** *For every nonempty fibre, the smallest worst-case absolute error of any scalar estimate $t \in \mathbb{R}$ of the target is*

$$\inf_{t \in \mathbb{R}} \sup_{F \in \mathcal{F}_M(p_M)} |t - T_h(F)| = \frac{W_M(p_M)}{2}. \tag{5}$$

The minimum is attained at $t^\star = (L_M + U_M)/2$.

A proof is given in Appendix A.1. Thus functional width is not merely an uncertainty diagnostic: its half-width is exactly the smallest worst-case target error achievable given the current response vector. This result is conditional on the realized response vector p_M. We next characterize the worst target ambiguity induced by the finite design D_M. For a finite response space, remove one redundant outcome probability at each probe and let $\kappa_M(z) \in \mathbb{R}^{M(|\mathcal{Y}|-1)}$ collect the remaining kernel coordinates $k_\theta(y \mid x_i, z)$. Define $\mathcal{V}_M = \operatorname{span}\{1, \kappa_{M,j} : j = 1, \ldots, M(|\mathcal{Y}| - 1)\}$. By construction, if $\mathcal{A}_M F = \mathcal{A}_M G$, then F and G agree on the population average of every $v \in \mathcal{V}_M$. We define $d_M(h) = \inf_{v \in \mathcal{V}_M} \|h - v\|_\infty$ as the distance of the target function h from $\mathcal{V}_M$, and let $\Delta_M(h) = \sup_{\mathcal{A}_M F = \mathcal{A}_M G} |T_h(F) - T_h(G)|$ denote the largest target disagreement between observationally equivalent populations.

Theorem 2 (Exact functional ambiguity)**.** *Suppose $\mathcal{Z}$ is compact, $h \in C(\mathcal{Z})$, and the response coordinates are continuous. Then*

$$\boxed{\Delta_M(h) = \sup_{p_M : \mathcal{F}_M(p_M) \neq \varnothing} W_M(p_M) = 2 d_M(h).} \tag{6}$$

The proof is given in Appendix A.2. The theorem shows that the largest target disagreement hidden by observational equivalence is exactly twice the distance of h from $\mathcal{V}_M$. The same approximation error also determines the best possible recovery from the observed response vector: over all estimators τ based only on $\mathcal{A}_M F$, the minimax error is $\inf_\tau \sup_{F \in \mathcal{P}(\mathcal{Z})} |\tau(\mathcal{A}_M F) - T_h(F)| = d_M(h)$. This error is attained by a response-affine estimator constructed from a best uniform approximation of h in $\mathcal{V}_M$. Consequently, T_h is identified under D_M for every feasible response vector if and only if $d_M(h) = 0$, equivalently, $h \in \mathcal{V}_M$. The uniform-recovery result and its proof are given in Appendix A.3.

Together with Theorem 1, these results distinguish the realized and design-level recovery limits: $W_M(p_M)/2$ is the minimax error for the current response vector, whereas $d_M(h)$ is the uniform minimax error induced by the design D_M. This suggests selecting the next probe by the target disagreement that remains among observationally compatible populations.

4.2 FUNCTIONAL AMBIGUITY REDUCTION

The recovery characterization directly suggests an acquisition rule: rather than distinguishing all currently compatible populations, a query should resolve only those distinctions that affect the target. Suppose the current exact response vector is p_M. For a candidate probe $x \in \mathcal{X} \setminus D_M$, define its residual functional ambiguity as

$$R_h(x \mid p_M) = \sup_{\substack{F,G \in \mathcal{F}_M(p_M) \\ q_F(\cdot|x) = q_G(\cdot|x)}} |T_h(F) - T_h(G)|. \tag{7}$$

The constraint $q_F(\cdot \mid x) = q_G(\cdot \mid x)$ restricts attention to populations that remain observationally indistinguishable after the response distribution at x is revealed. In effect, the query partitions the current fibre according to the response distribution at x, and $R_h(x \mid p_M)$ is the largest target disagreement remaining within any resulting successor fibre. As shown in Appendix A.4, $R_h(x \mid p_M)/2$ is exactly the worst post-query minimax absolute target-recovery error. FAR therefore selects a candidate minimizing this residual ambiguity, i.e., $x_{\mathrm{FAR}} \in \arg\min_{x \in \mathcal{X} \setminus D_M} R_h(x \mid p_M)$.

4.3 FAR FROM FINITE RESPONSE BATCHES

The exact-response rule assumes that querying x reveals $q_F(\cdot \mid x)$, whereas in practice we observe only a finite batch of B responses, summarized by an outcome $c \in \mathcal{O}_x$. The fibrewise minimax interpretation still applies after each nonempty batch outcome, but acquisition must now also account for the distribution of these outcomes. Let s contain the acquired history and the fitted response model. Write $P_{s,x}(c \mid F)$ for the conditional probability of observing outcome c when querying x at state s under population F. The observation model specifies this distribution before the next batch is observed, with $P_{s,x}(c \mid F) \geq 0$ and $\sum_{c \in \mathcal{O}_x} P_{s,x}(c \mid F) = 1$.

Let $\mathcal{F}_0 \subseteq \mathcal{P}(\mathcal{Z})$ encode any known population restrictions. For an outcome c, let $\mathcal{D}_{s,x}(c) \subseteq \mathcal{F}_0$ contain the populations accepted by the confidence construction, accounting for sampling uncertainty and any response-model discrepancy. This set may impose joint constraints using the acquired history. Response-coordinate constraints are the special case $\mathcal{D}_{s,x}(c) = \{F \in \mathcal{F}_0 : q_F(\cdot \mid x) \in \mathcal{C}_{s,x}(c)\}$. After history $\mathcal{H}_M = \{(x_t, c_t)\}_{t=1}^M$, the compatible populations form

$$\mathcal{F}_s = \mathcal{F}_0 \cap \bigcap_{t=1}^{M} \mathcal{D}_{s_{t-1},x_t}(c_t). \tag{8}$$

For an admissible candidate x and an outcome $c \in \mathcal{O}_x$, define the successor fibre $\mathcal{F}_{s,x,c} = \mathcal{F}_s \cap \mathcal{D}_{s,x}(c)$. Let $\mathcal{X}_s$ be the nonempty finite set of candidates permitted by the sampling protocol and remaining budget that have at least one nonempty successor fibre. For every nonempty branch, let $L_{x,c} = \inf_{F \in \mathcal{F}_{s,x,c}} T_h(F)$ and $U_{x,c} = \sup_{F \in \mathcal{F}_{s,x,c}} T_h(F)$ define its sharp enclosing target interval. The midpoint $m_{x,c} = (L_{x,c} + U_{x,c})/2$ is minimax under absolute loss by the argument of Theorem 1. The largest midpoint error over compatible branches is

$$C_{\mathrm{mid}}(s,x) = \max_{\substack{c \in \mathcal{O}_x: \\ \mathcal{F}_{s,x,c} \neq \varnothing}} \frac{U_{x,c} - L_{x,c}}{2}. \tag{9}$$

Let $C^\star(s) = \min_{x \in \mathcal{X}_s} C_{\mathrm{mid}}(s,x)$. Given prespecified slack parameters $\varepsilon_{\mathrm{safe}}, \eta_0 \geq 0$, retain

$$\mathcal{A}_{\mathrm{safe}}(s) = \{x \in \mathcal{X}_s : C_{\mathrm{mid}}(s,x) \leq (1 + \varepsilon_{\mathrm{safe}}) C^\star(s) + \eta_0\}. \tag{10}$$

Because the midpoint is minimax within each successor fibre, $C_{\mathrm{mid}}(s,x)$ is the worst branchwise minimax recovery error. The safety set retains candidates that are optimal or near-optimal under this criterion, after which FAR uses the outcome probabilities to minimize worst-population expected midpoint error. For each compatible population F, evaluate the expected midpoint error using its observation law $P_{s,x}(\cdot \mid F)$ and target $T_h(F)$. This keeps the outcome probabilities and target value coupled through the same population. For an empty successor fibre, fix $m_{x,c}$ at the prespecified global-range midpoint. FAR then defines

$$\boxed{\begin{aligned} J_{\mathrm{mid}}(s,x) &= \sup_{F \in \mathcal{F}_s} \sum_{c \in \mathcal{O}_x} P_{s,x}(c \mid F)\, |m_{x,c} - T_h(F)|, \\ x_{\mathrm{FAR}}(s) &\in \arg\min_{x \in \mathcal{A}_{\mathrm{safe}}(s)} J_{\mathrm{mid}}(s,x). \end{aligned}} \tag{11}$$

An exact solution therefore preserves branchwise minimax safety and minimizes worst-population expected midpoint risk among candidates satisfying that constraint. This gives one-step constrained optimality for the fixed midpoint readout, distinct from the unrestricted one-step minimax result in the exact-response setting. When a learned response model admits several observation laws for a population, the risk additionally takes the supremum over those admissible laws. Appendix B develops this extension and two sampling-model instantiations, including their compatible sets, computational representations, and conventions for inconsistent histories.

4.4 Computation and statistical guarantees

Computation. On a finite latent sieve, populations are represented by simplex weights w, and $T_h(F)$ is linear in w. Linear compatibility constraints yield polytopal fibres whose sharp target endpoints are obtained by linear programs; discrete population restrictions instead lead to integer optimization. Evaluating Eq. (11) also requires the observation law and preserves its coupling with the target through the same population. The resulting representation and risk solver therefore depend on the chosen observation model. Where certified risk bounds are available, they support candidate pruning and bound the returned action's selection error. Appendix B develops the model-specific endpoint and risk computations and distinguishes certified optimization from numerical search.

Statistical validity. Separately, let $K = \lfloor N/B \rfloor$ be the maximum number of acquired batches, and let s_j be the state after batch j. Suppose $F^\star \in \mathcal{F}_0$, the response-model allowances are simultaneously valid with probability at least $1 - \beta$, and the sampling confidence construction is simultaneously valid over the acquired history with probability at least $1 - \alpha$. The latter condition can follow from a joint confidence sequence or from conditional per-batch noncoverage bounds α_j with $\sum_{j=1}^{K} \alpha_j \le \alpha$. If these validity events imply $F^\star \in \mathcal{D}_{s_{j-1},x_j}(c_j)$ at every round, a union bound yields

$$\Pr\big(F^\star \in \mathcal{F}_{s_j} \text{ for all } j \le K\big) \ge 1 - \alpha - \beta. \tag{12}$$

Consequently, on the simultaneous validity event, the target interval contains $T_h(F^\star)$ at every round and therefore at any stopping round up to K. In a finite-sieve implementation, this guarantee further requires either that $F^\star$ be represented on the sieve or that the compatibility constraints account for discretization error. Appendix B.2.1 gives the corresponding approximation bounds. Adaptive selection therefore requires simultaneous validity over the acquired history, as formalized in Appendix A.5.

5 Experiments

We evaluate FAR in three settings. First, we study functional recovery across adaptive acquisition tasks under a shared observation–target protocol. Second, we directly evaluate acquisition under observational equivalence and measure the remaining functional ambiguity and recovery error. Finally, we consider budgeted LLM evaluation, where a small number of benchmark queries is used to recover the model's full-benchmark accuracy.

Setting. We evaluate FAR on CIFAR-10H (Peterson et al., 2019), Duolingo-FR (Settles et al., 2018), and ASSISTments (Feng et al., 2009), and use OpinionQA (Santurkar et al., 2023) and the 2024 Cooperative Election Study (CES) (Schaffner et al., 2025) for controlled observational-equivalence experiments with four prespecified targets per dataset. We compare against ActiveCQ (Gao & Sejdinovic, 2026), Adaptive Elicitation (Wang et al., 2025), BECAT (Zhuang et al., 2023), DGSM-IG (Belakaria et al., 2024), ITL (Hübotter et al., 2024), TSIED (Ailer et al., 2024), Active Measurement (Hamilton et al., 2025), and μPOCA (Astorga et al., 2024), with Random added for observational equivalence. All methods share the same frozen split, initial observations, candidate pool, target, budget, response bank, and terminal midpoint estimator, with held-out target responses used only for evaluation. We report MAE, P95, Err-AUC, N.Width, CCR, and N.EndErr, and additionally $W_{\max}$, $W_{\exp}$, and RTE for observational equivalence. Full experimental details are provided in Appendix C.

Our main findings are: (1) FAR improves functional recovery under shared acquisition protocols, with consistent gains across diverse tasks. (2) FAR reduces residual ambiguity under observational equivalence, with the clearest improvements on CES. (3) FAR extends ambiguity-aware acquisition to budgeted LLM evaluation, recovering benchmark accuracy from limited queries.

Table 1: We compare FAR with Active Measurement (Hamilton et al., 2025), ActiveCQ-TVR (Gao & Sejdinovic, 2026), Adaptive Elicitation EIG/MCTS (Wang et al., 2025), BECAT (Zhuang et al., 2023), DGSM-IG (Belakaria et al., 2024), ITL (Hübotter et al., 2024), TSIED (Ailer et al., 2024), and μPOCA (Astorga et al., 2024). We report normalized interval width (N.Width), mean absolute error (MAE), 95th-percentile error (P95), error–budget area under the curve (Err-AUC), covered contraction rate (CCR), and normalized endpoint error (N.EndErr). ↑/↓ indicate the preferred direction. Bold marks the best value in each column.

Method	N.Width↓	MAE↓	P95↓	Err-AUC↓	CCR↑	N.EndErr↓
CIFAR-10H						
Active Measurement	0.9380	0.0523	0.0895	0.0557	0.5667	0.7223
ActiveCQ-TVR	0.8551	0.0442	0.0836	0.0477	0.9000	0.6395
Adaptive Elicitation EIG/MCTS	0.9065	0.0498	0.0908	0.0549	0.7000	0.6915
BECAT	0.8338	0.0422	0.0819	0.0472	**0.9667**	0.6194
DGSM-IG	0.8923	0.0483	0.0889	0.0550	0.7667	0.6768
FAR (ours)	**0.8084**	**0.0395**	**0.0772**	**0.0453**	**0.9667**	**0.5938**
ITL	0.8923	0.0483	0.0889	0.0550	0.7667	0.6768
TSIED	0.8112	0.0399	0.0792	0.0462	**0.9667**	0.5972
μPOCA	0.8582	0.0446	0.0836	0.0484	0.9000	0.6426
Duolingo-FR						
Active Measurement	0.9766	0.1441	0.2000	0.1474	0.2667	0.7338
ActiveCQ-TVR	0.9401	0.1333	0.1974	0.1380	0.5889	0.6973
Adaptive Elicitation EIG/MCTS	0.9451	0.1344	0.1926	0.1429	0.4444	0.7024
BECAT	0.9578	0.1384	0.1978	0.1451	0.4778	0.7151
DGSM-IG	0.9595	0.1387	0.1978	0.1431	0.3889	0.7167
FAR (ours)	**0.9151**	**0.1263**	**0.1710**	**0.1364**	**0.7333**	**0.6735**
ITL	0.9444	0.1344	0.1978	0.1396	0.5333	0.7016
TSIED	0.9482	0.1358	0.1884	0.1411	0.5556	0.7058
μPOCA	0.9393	0.1330	0.1974	0.1383	0.5889	0.6965
ASSISTments						
Active Measurement	1.1107	0.0075	**0.0346**	**0.0067**	0.1667	0.7615
ActiveCQ-TVR	1.0554	0.0079	0.0350	0.0075	0.2095	**0.7210**
Adaptive Elicitation EIG/MCTS	1.2728	0.0083	0.0350	0.0075	0.0714	0.8442
BECAT	1.2053	0.0098	0.0362	0.0077	0.1524	0.8225
DGSM-IG	1.3236	0.0079	0.0349	0.0076	0.0524	0.8680
FAR (ours)	**1.0159**	0.0089	0.0351	0.0083	**0.2714**	0.7463
ITL	1.3236	0.0079	0.0349	0.0076	0.0524	0.8680
TSIED	1.1459	**0.0073**	**0.0346**	0.0067	0.1667	0.7740
μPOCA	1.1850	0.0080	0.0353	0.0076	0.1381	0.7940

5.1 Functional recovery across adaptive acquisition tasks

We evaluate adaptive functional recovery on CIFAR-10H (Peterson et al., 2019), Duolingo-FR (Settles et al., 2018), and ASSISTments (Feng et al., 2009) under a shared observation–target protocol against eight recent acquisition methods. As shown in Table 1, FAR achieves the strongest gains on CIFAR-10H and Duolingo-FR. On CIFAR-10H, it is best on N.Width, MAE, P95, Err-AUC, and N.EndErr, with a significant MAE improvement over TSIED after multiplicity correction (-0.000439, 95% CI $[-0.000987, -0.000072]$, $p = .0128$). On Duolingo-FR, FAR is best on all six metrics, reducing MAE from 0.1330 to 0.1263. On ASSISTments, FAR attains the smallest N.Width and the highest CCR, while other methods achieve lower MAE and Err-AUC. This contrast illustrates that ambiguity reduction provides tighter control of the compatible target set, but need not coincide with the lowest realized midpoint error in every dataset. Budget-wise trajectories are reported in Appendix D.1.

5.2 Recovery under observational equivalence

We evaluate acquisition under observational equivalence on OpinionQA (Santurkar et al., 2023) and the 2024 Cooperative Election Study (CES) (Schaffner et al., 2025), where each shared initial state admits distinct latent populations with identical observed response probabilities but different target

Table 2: We compare FAR with Active Measurement (Hamilton et al., 2025), ActiveCQ-TVR (Gao & Sejdinovic, 2026), Adaptive Elicitation EIG/MCTS (Wang et al., 2025), BECAT (Zhuang et al., 2023), DGSM-IG (Belakaria et al., 2024), ITL (Hübotter et al., 2024), TSIED (Ailer et al., 2024), μPOCA (Astorga et al., 2024), and Random on 200 shared states per dataset. We report worst-branch interval width ($W_{\max}$), worst-population expected interval width ($W_{\exp}$), worst-population expected midpoint error (RTE), farthest-endpoint error (End.Err), efficiency regret (Eff.Reg.), and oracle-action agreement (Oracle). ↑/↓ indicate the preferred direction. Bold marks the best value in each column.

Method	$W_{\max}$ (pp)↓	$W_{\exp}$ (pp)↓	RTE↓	End.Err (pp)↓	Eff.Reg.↓	Oracle↑
OpinionQA						
FAR (ours)	**74.1592**	**71.7172**	**0.2852**	**58.7183**	**0.0000**	**1.0000**
TSIED	74.1631	71.7283	0.2854	58.7657	0.0003	0.9950
Adaptive Elicitation	74.5171	72.1127	0.2920	60.0302	0.0008	0.9200
ActiveCQ-TVR	74.2746	71.8700	0.2873	59.1239	0.0008	0.9750
DGSM-IG	74.4214	72.0422	0.2890	59.4555	0.0011	0.9600
ITL	84.3465	83.3195	0.4053	81.1703	0.0435	0.0500
μPOCA	82.3242	80.7735	0.3761	75.6850	0.0104	0.3750
Active Measurement	86.8680	86.3132	0.4239	84.9412	0.0138	0.0550
BECAT	88.0225	87.7479	0.4366	87.3510	0.0394	0.0000
Random	87.4589	87.0554	0.4297	86.0905	0.0167	0.0300
CES						
FAR (ours)	**71.6669**	**71.0862**	**0.3148**	**63.9309**	**0.0000**	**1.0000**
TSIED	71.8039	71.2022	0.3155	64.1089	0.0006	0.7150
Adaptive Elicitation	72.4779	71.7715	0.3184	64.9104	0.0011	0.4000
ActiveCQ-TVR	73.8130	72.8123	0.3241	66.4428	0.0012	0.3200
DGSM-IG	73.8674	72.8713	0.3246	66.5452	0.0013	0.2450
ITL	73.8629	72.8524	0.3243	66.4954	0.0012	0.2450
μPOCA	73.8769	72.8774	0.3246	66.5461	0.0012	0.2500
Active Measurement	83.9684	83.5059	0.4072	81.8016	0.0021	0.0550
BECAT	85.5706	85.2937	0.4240	84.8402	0.0022	0.0350
Random	84.0440	83.4486	0.4053	81.5709	0.0020	0.0150

values. From the same two-question history, each method selects one query and receives 40 responses. FAR achieves the best value on all six metrics in both datasets, reducing residual functional ambiguity and recovery error. On CES, FAR reduces $W_{\max}$ from 71.8039 to 71.6669, $W_{\exp}$ from 71.2022 to 71.0862, and End.Err from 64.1089 to 63.9309, while attaining the lowest RTE and Eff.Reg. and perfect oracle-action agreement. Figure 2 further examines separation of target-conflicting populations. On CES, FAR reaches a discrimination error of 0.50% at 40 responses compared with 1.07% for TSIED, while showing comparable discrimination performance on OpinionQA. Together, these results show that FAR reduces target-relevant ambiguity over both the full compatible population set and population pairs that remain observationally equivalent under the initial design. Multi-round results are provided in Appendix D.1.

5.3 Full-score recovery in budgeted LLM evaluation

We next study budgeted LLM evaluation, where the goal is to estimate benchmark accuracy from a small subset of model responses. We evaluate Qwen3-4B (Yang et al., 2025) and Phi-3.5-mini-instruct (Abdin et al., 2024) on an MMLU-Pro subset (Wang et al., 2024) containing 40 prompts from each of 14 subjects. Starting from 14 shared responses, each method selects five additional prompts and estimates accuracy over the benchmark subset. FAR maintains compatible subject-level correct totals and allocates the remaining queries to reduce ambiguity in the overall score, using the midpoint of the resulting score interval as its estimate. Figure 3 compares FAR with Random, RSL (Zhang et al., 2025), AIPW (Zhang et al., 2025), tinyBenchmarks (Maia Polo et al., 2024b), and SparseEval (Zhang et al., 2026). FAR achieves the lowest mean MAE on both evaluated models, reducing error by 7.0% on Qwen3-4B relative to Random and by 25.7% on Phi-3.5-mini relative to SparseEval. These results show that functional-ambiguity-aware acquisition can extend effectively to finite-budget LLM evaluation. Full protocol and implementation details are provided in Appendix C.5.

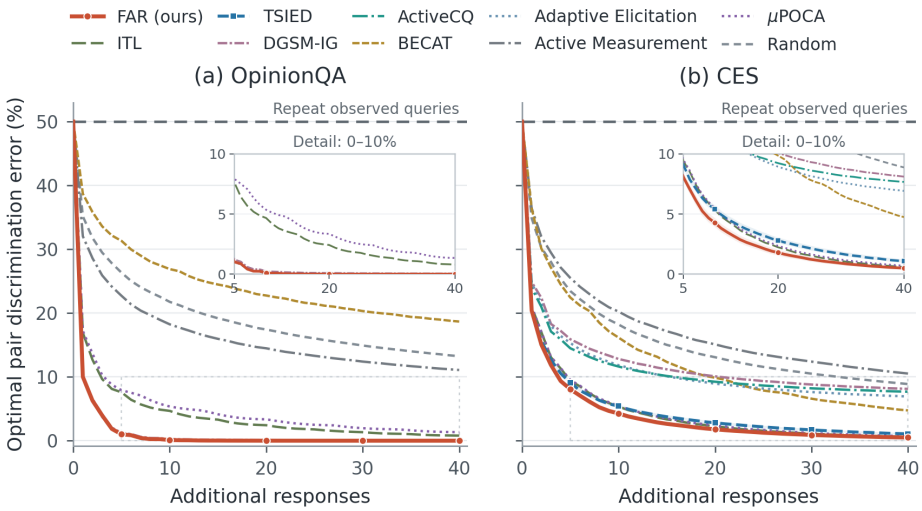


Figure 2: **Distinguishing target-conflicting populations.** For each method, we fix its selected query and measure pair-discrimination error as the response budget increases. Lower error indicates better separation of observationally equivalent populations with different target values.

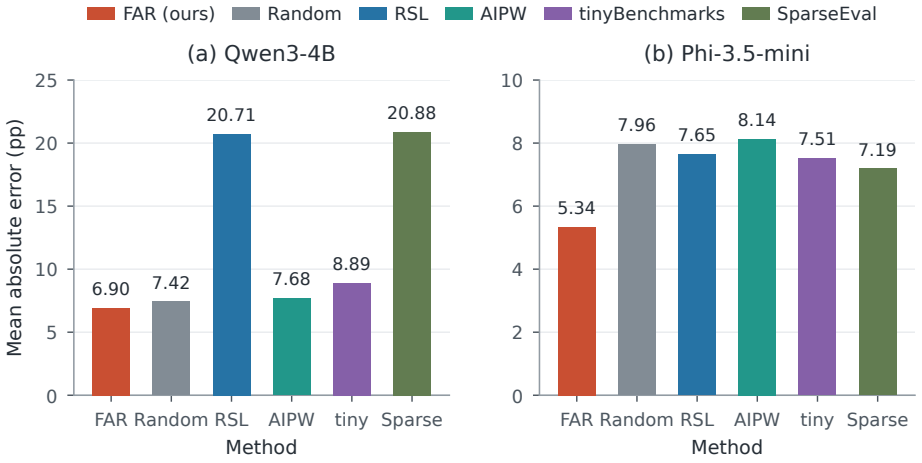


Figure 3: **Budgeted LLM score recovery.** Mean absolute error in estimating full MMLU-Pro subset accuracy after five additional prompt evaluations, starting from 14 shared observations. Results average over 60 histories and are reported in percentage points (lower is better).

6 Conclusion

FAR turns target ambiguity under observational equivalence into a principled acquisition objective, with functional ambiguity directly characterizing minimax target-recovery error. Across adaptive acquisition and controlled observational-equivalence benchmarks, FAR reduces target-relevant ambiguity and yields strong recovery performance. The same principle also extends to finite-budget LLM evaluation, where FAR improves benchmark-score recovery on the two evaluated models. At the same time, the experiments highlight an important distinction between the objective optimized by FAR and realized point accuracy: reducing worst-case functional ambiguity does not necessarily lead to lower point error under every data distribution. Better understanding this gap between minimax recovery guarantees and average-case performance is an important direction for further study.

AI USE STATEMENT

Generative AI tools were used to assist with the refinement of theoretical arguments, mathematical proofs, code debugging, and manuscript editing. All theoretical claims and proofs were independently verified by the authors. The authors also verified the experimental protocols and implementations, numerical results, and cited references, and manually reviewed all AI-assisted content. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide complete proofs of the theoretical results in Appendix A and detailed algorithmic and implementation specifications in Appendix B, including the finite-batch observation models, compatible-set construction, optimization procedures, and numerical certification. Appendix C describes the experimental protocols, baseline implementations, evaluation metrics, statistical procedures, and the budgeted LLM score-recovery setup. Additional acquisition trajectories, ablations, sensitivity analyses, and LLM evaluation results are reported in Appendix D. Together, these materials document the assumptions, algorithms, and evaluation procedures used in this work. Anonymous source code is available at `https://anonymous.4open.science/r/FAR-D5B5/`.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL `https://arxiv.org/abs/2404.14219`.

Elisabeth Ailer, Niclas Dern, Jason Hartford, and Niki Kilbertus. Targeted sequential indirect experiment design. *Advances in Neural Information Processing Systems*, 37:122029–122053, 2024.

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. 2009.

Nicolás Astorga, Tennison Liu, Nabeel Seedat, and Mihaela van der Schaar. Active learning with llms for partially observed and cost-aware scenarios. *Advances in Neural Information Processing Systems*, 37:20819–20857, 2024.

Syrine Belakaria, Benjamin Letham, Janardhan R Doppa, Barbara Engelhardt, Stefano Ermon, and Eytan Bakshy. Active learning for derivative-based global sensitivity analysis with gaussian processes. *Advances in neural information processing systems*, 37:53887–53917, 2024.

Kenneth A Bollen. Latent variables in psychology and the social sciences. *Annual review of psychology*, 53(1):605–634, 2002.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Aditya Chattopadhyay, Benjamin David Haeffele, Rene Vidal, and Donald Geman. Performance bounds for active binary testing with information maximization. In *Forty-first International Conference on Machine Learning*, 2024.

Jian Chen, Zhehao Li, and Xiaojie Mao. Learning with selectively labeled data from multiple decision-makers. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8480–8519. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25al.html.

Seong Jin Cho, Gwangsu Kim, and Chang D Yoo. Hypothesis perturbation for active learning. *IEEE Journal of Selected Topics in Signal Processing*, 19(1):115–128, 2024.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979.

Ronald DeVore, Robert D Nowak, Rahul Parhi, Guergana Petrova, and Jonathan W Siegel. Optimal recovery meets minimax estimation. *arXiv preprint arXiv:2502.17671*, 2025.

Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 19(3): 243–266, 2009.

Louis Filstroff, Iiris Sundin, Petrus Mikkola, Aleksei Tiulpin, Juuso Kylmäoja, and Samuel Kaski. Targeted active learning for bayesian decision-making. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=KxPjuiMgmm.

Erdun Gao and Dino Sejdinovic. Activecq: Active estimation of causal quantities. In *International Conference on Learning Representations*, volume 2026, pp. 82957–83004, 2026.

Zijun Gao, Shu Ge, and Jian Qian. Bridging multiple worlds: Multi-marginal optimal transport for causal partial-identification problem. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/gao25b.html.

Raffaella Giacomini and Toru Kitagawa. Robust bayesian inference for set-identified models. *Econometrica*, 89(4):1519–1556, 2021.

Max Hamilton, Jinlin Lai, Wenlong Zhao, Subhransu Maji, and Daniel R. Sheldon. Active measurement: Efficient estimation at scale. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/d6adf82fa531dcb8bfa53c224ce5e466-Abstract-Conference.html.

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning: Theory and applications. *Advances in Neural Information Processing Systems*, 37: 124686–124755, 2024.

Julia Kostin, Nicola Gnecco, and Fanny Yang. Achievable distributional robustness when the robust risk is only partially identified. *Advances in Neural Information Processing Systems*, 37: 83915–83950, 2024.

Quinn Lanners, Cynthia Rudin, Alexander Volfovsky, and Harsh Parikh. Data fusion for partial identification of causal effects. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/dd4e79c2228c43954cc074abf2f85a07-Abstract-Conference.html.

Sirui Lin, Zijun Gao, Jose Blanchet, and Peter Glynn. Tightening causal bounds via covariate-aware optimal transport. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37815–37845. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lin25f.html.

Felipe Maia Polo, Subha Maity, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Weak supervision performance evaluation via partial identification. *Advances in Neural Information Processing Systems*, 37:135704–135740, 2024a.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. Quantifying aleatoric uncertainty of the treatment effect: a novel orthogonal learner. *Advances in Neural Information Processing Systems*, 37:105039–105089, 2024.

Hyungsik Roger Moon and Frank Schorfheide. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012.

Ruochen Niu, Qiang Shen, and Sze Zheng Yong. Active model discrimination: A multi-stage approach with adaptive partitions. *Automatica*, 190:113067, 2026.

Alizée Pace, Hugo Yèche, David Ha, Gunnar Ratsch, and Guy Tennenholtz. Delphic offline reinforcement learning under nonidentifiable hidden confounding. In *International Conference on Learning Representations*, volume 2024, pp. 2768–2796, 2024.

Luciano Di Palma, Yanlei Diao, and Anna Liu. Efficient version space algorithms for human-in-the-loop model development. *ACM Transactions on Knowledge Discovery from Data*, 18(3):1–49, 2024.

Jordan Penn, Lee M. Gunderson, Gecia Bravo-Hermsdorff, Ricardo Silva, and David Watson. Budgetiv: Optimal partial identification of causal effects with mostly invalid instruments. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2485–2493. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/penn25a.html.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9617–9626, 2019.

Nicholas Rittler and Kamalika Chaudhuri. Agnostic multi-group active learning. *Advances in Neural Information Processing Systems*, 36:1100–1118, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Brian Schaffner, Marissa Shih, Stephen Ansolabehere, and Jeremy Pope. Cooperative Election Study Common Content, 2024, 2025. URL https://doi.org/10.7910/DVN/X11EP6.

Jonas Schweisthal, Dennis Frauen, Maresa Schröder, Konstantin Hess, Niki Kilbertus, and Stefan Feuerriegel. Learning representations of instruments for partial identification of treatment effects. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53703–53727. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/schweisthal25a.html.

Burr Settles. Active learning literature survey. 2009.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pp. 56–65, 2018.

Wanggang Shen, Jiayuan Dong, and Xun Huan. Variational sequential optimal experimental design using reinforcement learning. *Computer Methods in Applied Mechanics and Engineering*, 444: 118068, 2025.

Jiyuan Tan, Jose Blanchet, and Vasilis Syrgkanis. Consistency of neural causal partial identification. *Advances in Neural Information Processing Systems*, 37:68956–68999, 2024.

Geert Verbeke and Geert Molenberghs. Modeling through latent variables. *Annual Review of Statistics and Its Application*, 4:267–282, 2017.

Jimmy Wang, Thomas P. Zollo, Richard S. Zemel, and Hongseok Namkoong. Adaptive elicitation of latent information using natural language. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/wang25ff.html.

Ran Wang, Shuyue Chen, and Yu Yu. Extending version-space theory to multi-label active learning with imbalanced data. *Pattern Recognition*, 142:109690, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Tom Yan and Chicheng Zhang. The human-ai substitution game: active learning from a strategic labeler. In *International Conference on Learning Representations*, volume 2024, pp. 25319–25377, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Guanhua Zhang, Florian E. Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/c57bacb5ea621582235c54ce2a83136d-Abstract-Conference.html.

Taolin Zhang, Hang Guo, Wang Lu, Tao Dai, Shu-Tao Xia, and Jindong Wang. Sparseeval: Efficient evaluation of large language models by sparse optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CZAzAedGSV.

Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Bayesian inference and partial identification in multi-treatment causal inference with unobserved confounding. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3608–3626. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/zheng22a.html.

Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen, Jinze Wu, and Xin Li. A bounded ability estimation for computerized adaptive testing. *Advances in Neural Information Processing Systems*, 36:2381–2402, 2023.

A PROOFS

A.1 PROOF OF THEOREM 1

Proof. For any $t \in \mathbb{R}$, all target values attainable within $\mathcal{F}_M(p_M)$ lie between L_M and U_M. By the definitions of the infimum and supremum, target values in the fibre can approach both endpoints, and therefore

$$\sup_{F \in \mathcal{F}_M(p_M)} |t - T_h(F)| = \max\{|t - L_M|,\ |t - U_M|\} \geq \frac{U_M - L_M}{2}.$$

The lower bound is attained by choosing $t^\star = (L_M + U_M)/2$, for which the two endpoint errors are equal. Hence the minimax absolute recovery error is $W_M(p_M)/2$, with the midpoint as an optimal estimate. □

A.2 PROOF OF THEOREM 2

Proof. For any nonempty fibre $\mathcal{F}_M(p_M)$, its target diameter is

$$W_M(p_M) = \sup_{F,G \in \mathcal{F}_M(p_M)} |T_h(F) - T_h(G)|.$$

Since two populations are observationally equivalent exactly when they belong to the same fibre, it follows that

$$\Delta_M(h) = \sup_{p_M : \mathcal{F}_M(p_M) \neq \varnothing} W_M(p_M).$$

It therefore remains to show that $\Delta_M(h) = 2d_M(h)$. For the upper bound, let F and G be observationally equivalent. They agree on the expectation of every $v \in \mathcal{V}_M$, and hence

$$\begin{aligned} |T_h(F) - T_h(G)| &= \left| \int_{\mathcal{Z}} (h - v)\, dF - \int_{\mathcal{Z}} (h - v)\, dG \right| \\ &\leq \int_{\mathcal{Z}} |h - v|\, dF + \int_{\mathcal{Z}} |h - v|\, dG \leq 2\|h - v\|_\infty. \end{aligned}$$

Taking the infimum over $v \in \mathcal{V}_M$ and then the supremum over observationally equivalent pairs gives $\Delta_M(h) \leq 2d_M(h)$. For the reverse inequality, suppose $d_M(h) > 0$. Since $\mathcal{V}_M$ is a finite-dimensional closed subspace of $C(\mathcal{Z})$, the Hahn–Banach theorem gives a continuous linear functional $\Lambda \in C(\mathcal{Z})^*$ satisfying

$$\|\Lambda\| = 1, \qquad \Lambda|_{\mathcal{V}_M} = 0, \qquad \Lambda(h) = d_M(h).$$

By the Riesz–Markov representation theorem, write $\Lambda(f) = \int_{\mathcal{Z}} f\, d\eta$ for a finite signed regular Borel measure η with $\|\eta\|_{\mathrm{TV}} = 1$. Since $1 \in \mathcal{V}_M$, we have $\eta(\mathcal{Z}) = 0$. Thus, for the Jordan decomposition $\eta = \eta^+ - \eta^-$,

$$\eta^+(\mathcal{Z}) = \eta^-(\mathcal{Z}) = \frac{1}{2}, \qquad F^\star := 2\eta^+, \qquad G^\star := 2\eta^-$$

define two probability measures. Because Λ annihilates every response coordinate in $\mathcal{V}_M$, $F^\star$ and $G^\star$ have identical expectations of all nonredundant coordinates $\kappa_{M,j}$ and hence $\mathcal{A}_M F^\star = \mathcal{A}_M G^\star$. At the same time,

$$T_h(F^\star) - T_h(G^\star) = 2 \int_{\mathcal{Z}} h\, d\eta = 2\Lambda(h) = 2d_M(h).$$

Therefore $\Delta_M(h) \geq 2d_M(h)$. If $d_M(h) = 0$, the previously established upper bound already gives $\Delta_M(h) = 0$. Combining the two directions proves the result. □

A.3 Uniform minimax recovery and identification

Let $v^\star(z) = a_0^\star + (a^\star)^\top \kappa_M(z) \in \mathcal{V}_M$ be a best uniform approximation to h, so that $\|h - v^\star\|_\infty = d_M(h)$. Since $\mathcal{V}_M$ is finite-dimensional, the infimum defining $d_M(h)$ is attained. Consider the response-affine estimator

$$\tau^\star(\mathcal{A}_M F) = a_0^\star + (a^\star)^\top \int_{\mathcal{Z}} \kappa_M(z)\, dF(z).$$

For every $F \in \mathcal{P}(\mathcal{Z})$,

$$|\tau^\star(\mathcal{A}_M F) - T_h(F)| = \left| \int_{\mathcal{Z}} (v^\star - h)\, dF \right| \le \|v^\star - h\|_\infty = d_M(h).$$

Hence the uniform minimax recovery error is at most $d_M(h)$. For the reverse inequality, the case $d_M(h) = 0$ is immediate. Suppose $d_M(h) > 0$. The proof of Theorem 2 constructs observationally equivalent populations $F^\star$ and $G^\star$ such that

$$\mathcal{A}_M F^\star = \mathcal{A}_M G^\star, \qquad |T_h(F^\star) - T_h(G^\star)| = 2 d_M(h).$$

Any estimator based only on the observed response vector must take the same value on these two populations. Therefore, by the triangle inequality,

$$\max\{|\tau(\mathcal{A}_M F^\star) - T_h(F^\star)|\,, |\tau(\mathcal{A}_M G^\star) - T_h(G^\star)|\} \ge d_M(h).$$

Combining this lower bound with the affine construction above gives

$$\inf_{\tau} \sup_{F \in \mathcal{P}(\mathcal{Z})} |\tau(\mathcal{A}_M F) - T_h(F)| = d_M(h).$$

Thus allowing nonlinear functions of the observed response vector cannot improve the uniform minimax guarantee. Finally, T_h is identified under the finite design for every feasible response vector if and only if $h \in \mathcal{V}_M$. If $h \in \mathcal{V}_M$, the affine representation above determines $T_h(F)$ from $\mathcal{A}_M F$. Conversely, identification implies that every observationally equivalent pair has the same target value, so $\Delta_M(h) = 0$. By Theorem 2, this gives $d_M(h) = 0$; since $\mathcal{V}_M$ is closed, $h \in \mathcal{V}_M$.

A.4 One-step minimax optimality of FAR

For a candidate probe x, let $\mathcal{P}_x(p_M) = \{q_F(\cdot \mid x) : F \in \mathcal{F}_M(p_M)\}$ be the set of response distributions attainable from the current fibre. For $r \in \mathcal{P}_x(p_M)$, define the successor fibre $\mathcal{F}_{M+1}(p_M, r; x) = \{F \in \mathcal{F}_M(p_M) : q_F(\cdot \mid x) = r\}$, and let $W_{M+1}(p_M, r; x)$ denote its sharp target width.

Proposition 3 (One-step minimax optimality of FAR)**.** *For every candidate x,*

$$R_h(x \mid p_M) = \sup_{r \in \mathcal{P}_x(p_M)} W_{M+1}(p_M, r; x),$$

and its worst post-query minimax absolute recovery error is

$$\sup_{r \in \mathcal{P}_x(p_M)} \inf_{a_r \in \mathbb{R}} \sup_{F \in \mathcal{F}_{M+1}(p_M, r; x)} |a_r - T_h(F)| = \frac{R_h(x \mid p_M)}{2}.$$

Consequently, minimizing $R_h(x \mid p_M)$ minimizes the worst post-query minimax absolute target-recovery error over the candidate set.

Proof. Two populations in $\mathcal{F}_M(p_M)$ have the same response distribution at x if and only if they belong to the same successor fibre. Therefore,

$$\begin{aligned} R_h(x \mid p_M) &= \sup_{\substack{F, G \in \mathcal{F}_M(p_M) \\ q_F(\cdot|x) = q_G(\cdot|x)}} |T_h(F) - T_h(G)| \\ &= \sup_{r \in \mathcal{P}_x(p_M)} \sup_{F, G \in \mathcal{F}_{M+1}(p_M, r; x)} |T_h(F) - T_h(G)| \\ &= \sup_{r \in \mathcal{P}_x(p_M)} W_{M+1}(p_M, r; x). \end{aligned}$$

By Theorem 1, the minimax absolute recovery error within each nonempty successor fibre is $W_{M+1}(p_M, r; x)/2$. Taking the supremum over attainable r therefore gives the second claim. Since the factor $1/2$ is independent of x, the same candidates minimize both criteria. □

A.5 FINITE-BATCH STATISTICAL VALIDITY

We prove the simultaneous containment statement used in Section 4.4. Fix a true population $F^\star \in \mathcal{F}_0$. Let $\mathscr{G}_{j-1}$ contain all fitted-model and calibration information, the acquired history, and policy randomization available before round j, so that the action selected at that round is measurable with respect to $\mathscr{G}_{j-1}$. Let $\mathcal{E}_{\mathrm{samp}}$ denote the simultaneous validity event supplied by the sampling construction, with $\Pr(\mathcal{E}_{\mathrm{samp}}) \geq 1-\alpha$, and let $\mathcal{E}_{\mathrm{model}}$ denote the simultaneous validity event for any response-model allowances, with $\Pr(\mathcal{E}_{\mathrm{model}}) \geq 1-\beta$. When no fitted response model is used, we take $\mathcal{E}_{\mathrm{model}} = \Omega$ and $\beta = 0$.

It suffices that, on $\mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{model}}$, every realized confidence-compatible update retains the true population. That is, if the state before round j is s_{j-1}, the selected action is a_j, and the observed outcome is c_j, then

$$F^\star \in \mathcal{D}_{s_{j-1},a_j}(c_j) \qquad \text{for every acquired round } j \leq K.$$

The concrete constructions establishing this property depend on the observation model and are given in Appendix B.

Proof of Eq. (12). On the event $\mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{model}}$, the true population satisfies every compatibility constraint added along the realized acquisition trajectory. Since $F^\star \in \mathcal{F}_0$ and each successor fibre is obtained by intersecting the current fibre with the new confidence-compatible constraint, induction over the acquired rounds gives $F^\star \in \mathcal{F}_{s_j}$ for every $j \leq K$ on this event.

The two validity events need not be independent. By the union bound,

$$\Pr(\mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{model}}) \geq 1 - \Pr(\mathcal{E}_{\mathrm{samp}}^c) - \Pr(\mathcal{E}_{\mathrm{model}}^c) \geq 1 - \alpha - \beta.$$

Therefore

$$\Pr\big(F^\star \in \mathcal{F}_{s_j} \text{ for all } j \leq K\big) \geq 1 - \alpha - \beta,$$

which is Eq. (12). Since $[L_{s_j}, U_{s_j}]$ is obtained by minimizing and maximizing T_h over $\mathcal{F}_{s_j}$, population containment implies $T_h(F^\star) \in [L_{s_j}, U_{s_j}]$ simultaneously over the same rounds. The result therefore also holds at any stopping round bounded by K. □

For the Binomial population-response model, Appendix B.2 constructs the required sampling and response-model validity events through the finite-batch confidence constraints. For the fixed finite-population model, Appendix B.3.1 supplies the corresponding simultaneous event directly through its joint likelihood-ratio confidence sequence. Thus both observation models satisfy the same containment argument above, while their model-specific validity constructions remain separate.

B ALGORITHMIC AND IMPLEMENTATION DETAILS

B.1 OBSERVATION LAWS AND COMPATIBLE POPULATIONS

Finite-batch acquisition distinguishes the action selected by the policy from the random outcome subsequently observed. We write a for a generic acquisition action and $P_{s,a}(\cdot \mid F)$ for its conditional observation law under a compatible population $F \in \mathcal{F}_s$. After observing $c \in \mathcal{O}_a$, the state is updated through a confidence-compatible constraint $\mathcal{D}_{s,a}(c)$. The sampling design determines both the observation law and the corresponding state update. Binary item-level outcomes alone do not determine the distribution of a batch count.

When the response mechanism is learned rather than known, a compatible population may admit a family of observation laws $\mathcal{P}_{s,a}(F)$. The corresponding robust finite-batch objective is

$$J_{\mathrm{mid}}(s,a) = \sup_{F \in \mathcal{F}_s} \sup_{P \in \mathcal{P}_{s,a}(F)} \sum_{c \in \mathcal{O}_a} P(c)\, |m_{a,c} - T_h(F)|\,. \tag{13}$$

A singleton $\mathcal{P}_{s,a}(F)$ recovers Eq. (11). Importantly, the target value and the admissible observation laws are coupled through the same population F. Sampling uncertainty and response-model uncertainty enter through the compatible update and the admissible law family, rather than through an independent uncertainty set over target values. We use two instantiations. In the population-response

setting, $a = x$ is a probe and independent respondent sampling yields a Binomial count law with linear compatibility constraints (Appendix B.2). In the fixed finite-population setting, $a = b$ is a groupwise allocation and uniform sampling without replacement yields a Hypergeometric law with joint likelihood constraints (Appendix B.3). Both instantiate the same FAR objective and midpoint readout, but their state representations and numerical solvers are different.

B.2 Binomial population-response model

B.2.1 Observation model and compatible fibre

For a probe x_t, let $C_t = \sum_{j=1}^{B} Y_{t,j}$ denote the number of successes among B binary responses. Conditional on the population F, history $\mathcal{H}_{t-1}$, and selected probe x_t, latent respondent types are drawn independently from F, and responses are conditionally independent given these types according to the response kernel $k_\theta(\cdot \mid x_t, z)$. This allows heterogeneous respondents, since $k_\theta(1 \mid x, z)$ may vary with z. A common batch success probability arises after integrating over F.

Proposition 4 (Conditional batch count law)**.** *Under the population-response sampling model above, the responses in batch t are conditionally iid with success probability $p_t = q_F(1 \mid x_t)$. Hence,*

$$C_t \mid F, \mathcal{H}_{t-1}, x_t \sim \mathrm{Binomial}(B, p_t). \tag{14}$$

Proof. For any $(y_1, \ldots, y_B) \in \{0,1\}^B$, integrating over the independently sampled latent respondent types gives

$$\Pr(Y_{t,1} = y_1, \ldots, Y_{t,B} = y_B \mid F, \mathcal{H}_{t-1}, x_t) = \prod_{j=1}^{B} \int_{\mathcal{Z}} k_\theta(y_j \mid x_t, z)\, dF(z) = \prod_{j=1}^{B} q_F(y_j \mid x_t).$$

Thus the responses are conditionally iid Bernoulli with success probability $p_t = q_F(1 \mid x_t)$. Summing over the $\binom{B}{c}$ configurations with c successes yields Eq. (14). □

Adaptive query selection preserves the conditional count law because x_t is fixed given the acquired history. With a known response kernel, $P_{s,x}(c \mid F) = \mathrm{Binom}(c; B, q_F(1 \mid x))$. For a fitted kernel, we allow the actual success probability p to differ from the model-implied probability $q_F(1 \mid x)$ by a candidate-specific radius ρ_x. Let $I_x(c) = [\ell_{x,c}, u_{x,c}]$ denote the sampling confidence interval associated with count c. The resulting admissible success probabilities, observation laws, and confidence-compatible response region are

$$\begin{aligned} \mathcal{Q}_x(F) &= \{p \in [0,1] : |p - q_F(1 \mid x)| \le \rho_x\}, \\ \mathcal{P}_{s,x}(F) &= \{P_p : P_p(c) = \mathrm{Binom}(c; B, p), p \in \mathcal{Q}_x(F)\}, \\ \mathcal{C}_x(c) &= [\ell_{x,c} - \rho_x,\ u_{x,c} + \rho_x] \cap [0,1]. \end{aligned} \tag{15}$$

For the Binomial model, $I_x(c)$ is a Clopper–Pearson interval using the prespecified error allocation. For fitted kernels, the candidate-specific radius ρ_x accounts for response-model discrepancy. For computation, let $\mathcal{Z}_L = \{z_1, \ldots, z_L\}$ be a finite latent sieve and represent a population by weights $w \in \Delta_L$, with $k_x = (k_\theta(1 \mid x, z_1), \ldots, k_\theta(1 \mid x, z_L))^\top$. After a history $\mathcal{H}_M = \{(x_t, c_t)\}_{t=1}^{M}$, the current and successor fibres are

$$\begin{aligned} \mathcal{W}_s &= \left\{ w \in \Delta_L : k_{x_t}^\top w \in [\ell_t - \rho_{x_t}, u_t + \rho_{x_t}] \cap [0,1],\ t = 1, \ldots, M \right\}, \\ \mathcal{W}_{s,x,c} &= \left\{ w \in \mathcal{W}_s : k_x^\top w \in \mathcal{C}_x(c) \right\}. \end{aligned} \tag{16}$$

A response branch c is compatible whenever $\mathcal{W}_{s,x,c}$ is nonempty, and only compatible branches enter the branchwise safety criterion. Larger ρ_x enlarges the set of populations compatible with the response model while leaving the target functional unchanged.

The structural theory is defined over the full population space $\mathcal{P}(\mathcal{Z})$, while Eq. (16) uses a finite sieve. Let r_L denote the sieve mesh radius and let F_L be the push-forward of F to its nearest sieve points. If h is L_h-Lipschitz and the response coordinates defining the fibre are $L_{\kappa,j}$-Lipschitz, then

$$|T_h(F) - T_h(F_L)| \le L_h r_L, \qquad \left| \int \kappa_j\, dF - \int \kappa_j\, dF_L \right| \le L_{\kappa,j} r_L. \tag{17}$$

Accordingly, a continuous constraint $a_j \le \int \kappa_j \, dF \le b_j$ can be represented on the sieve by expanding its endpoints by $\delta_{L,j} = L_{\kappa,j} r_L$. If $L_{s,L}^{\delta_L}$ and $U_{s,L}^{\delta_L}$ denote the resulting sieve target endpoints, then

$$L_s \ge L_{s,L}^{\delta_L} - L_h r_L, \qquad U_s \le U_{s,L}^{\delta_L} + L_h r_L. \tag{18}$$

Hence the expanded interval $[L_{s,L}^{\delta_L} - L_h r_L, \, U_{s,L}^{\delta_L} + L_h r_L]$ contains the continuous target interval. When the relevant population is represented exactly on the sieve, no discretization allowance is required.

B.2.2 Response–target reduction and optimization

Let $h_L = (h(z_1), \dots, h(z_L))^\top$. Since the current and successor fibres are polytopes, feasibility and sharp target endpoints are obtained by linear programs. For each compatible branch,

$$L_{x,c} = \min_{w \in \mathcal{W}_{s,x,c}} h_L^\top w, \qquad U_{x,c} = \max_{w \in \mathcal{W}_{s,x,c}} h_L^\top w, \qquad m_{x,c} = \frac{L_{x,c} + U_{x,c}}{2}. \tag{19}$$

Removing the branch-specific constraint gives the current target interval, while the successor programs provide the branchwise midpoints used in the finite-batch rule. For a candidate x, the observation law depends on a compatible population through its admissible success probability p, while the target is $t = h_L^\top w$. We therefore project the compatible fibre onto the coupled response–target set

$$\mathcal{R}_x^p(s) = \left\{ (p,t) : \exists \, w \in \mathcal{W}_s, \; |p - k_x^\top w| \le \rho_x, \; 0 \le p \le 1, \; t = h_L^\top w \right\}. \tag{20}$$

This compact convex polygon preserves the coupling between response probability and target value through the same compatible population. Its boundary can be recovered by linear-program support queries. Equivalently, fixing p and minimizing or maximizing $h_L^\top w$ gives the lower and upper target boundaries, which are piecewise linear. The robust finite-batch objective then becomes

$$g_x(p,t) = \sum_{c=0}^{B} \binom{B}{c} p^c (1-p)^{B-c} |m_{x,c} - t|, \qquad J_{\mathrm{mid}}(s,x) = \max_{(p,t) \in \mathcal{R}_x^p(s)} g_x(p,t). \tag{21}$$

For fixed p, $g_x(p,t)$ is convex in t, so its maximum over each vertical section is attained on the upper or lower boundary. Along a boundary edge $t = ap + b$, splitting at the points where $t = m_{x,c}$ fixes the sign of each absolute-value term, reducing the objective on each segment to a univariate polynomial of degree at most $B+1$. The solver therefore optimizes over polygon boundaries, using feasible evaluations as lower bounds and Bernstein-form subdivision to obtain upper bounds until the desired numerical tolerance is reached.

B.2.3 Numerical certificates and feasibility handling

For each safe candidate, the solver maintains certified lower and upper bounds on the projected risk and prunes candidates whose lower bound exceeds the best available upper bound. If $\widehat{x}$ is the returned candidate, define

$$\underline{J}_x \le J_{\mathrm{mid}}(s,x) \le \overline{J}_x, \qquad \gamma_{\mathrm{sel}} = \overline{J}_{\widehat{x}} - \min_{x \in \mathcal{A}_{\mathrm{safe}}(s)} \underline{J}_x. \tag{22}$$

Then $\gamma_{\mathrm{sel}} \le \varepsilon_{\mathrm{alg}}$ guarantees

$$J_{\mathrm{mid}}(s,\widehat{x}) \le \min_{x \in \mathcal{A}_{\mathrm{safe}}(s)} J_{\mathrm{mid}}(s,x) + \varepsilon_{\mathrm{alg}}.$$

This certificate quantifies numerical optimization error for the finite-sieve one-step objective; statistical coverage is treated separately in Appendix A.5. A realized history may yield an empty finite-sieve fibre because of response-model discrepancy, discretization, or sampling variation. We restore feasibility by applying the smallest common enlargement to the accumulated probability constraints,

$$\lambda_s = \min_{w \in \Delta_L, \; \lambda \ge 0} \left\{ \lambda : \ell_t - \rho_{x_t} - \lambda \le k_{x_t}^\top w \le u_t + \rho_{x_t} + \lambda, \; t = 1, \dots, M \right\}. \tag{23}$$

All candidates at that state are then evaluated using the corresponding relaxed fibre. Here $\lambda_s = 0$ when the original fibre is feasible, while $\lambda_s > 0$ gives the minimum relaxation required to restore compatibility and serves as an incompatibility diagnostic.

B.3 HYPERGEOMETRIC FINITE-POPULATION MODEL

B.3.1 OBSERVATION MODEL AND JOINT COMPATIBILITY

Consider a fixed collection of $N_{\mathrm{tot}} = \sum_{g=1}^{G} N_g$ items partitioned into G groups, with one fixed binary outcome per item. Let $K_g \in \{0, \dots, N_g\}$ denote the unknown number of successes in group g. Associate each integer vector $K = (K_1, \dots, K_G)$ with the population

$$F_K = \frac{1}{N_{\mathrm{tot}}} \sum_{g=1}^{G} \left[(N_g - K_g)\delta_{(g,0)} + K_g \delta_{(g,1)} \right], \qquad T_h(F_K) = \frac{\sum_g K_g}{N_{\mathrm{tot}}}, \tag{24}$$

where $h(g, y) = y$. The admissible population class is therefore indexed by the integer group totals $K_1, \dots, K_G$.

At state s, group g has supplied n_g distinct observations, of which c_g are successes. An action $b = (b_1, \dots, b_G) \in \mathbb{Z}_{\geq 0}^{G}$ allocates the next batch subject to $0 \leq b_g \leq N_g - n_g$ and $\sum_g b_g = B$. Conditional on the acquired history and allocation, items are sampled uniformly without replacement within each group, independently across groups. If z_g denotes the number of newly observed successes in group g, then

$$P_{s,b}(z \mid F_K) = \prod_{g=1}^{G} \frac{\binom{K_g - c_g}{z_g} \binom{N_g - n_g - K_g + c_g}{b_g - z_g}}{\binom{N_g - n_g}{b_g}}. \tag{25}$$

The probability is zero outside the feasible Hypergeometric support. After observing z, the sufficient counts update to $(n + b, c + z)$. Allocations may depend on previous observations and groups may be queried repeatedly, while each item is observed at most once. The resulting randomness comes from sampling unobserved items from the fixed finite population.

Compatibility is defined directly over the unknown integer totals. Up to factors independent of K, the likelihood of an acquired history is

$$L_s(K) = \prod_{g=1}^{G} \frac{(K_g)_{c_g} (N_g - K_g)_{n_g - c_g}}{(N_g)_{n_g}}, \qquad (a)_r = a(a-1)\cdots(a-r+1). \tag{26}$$

Impossible totals have zero likelihood. Batch-count and policy-selection factors that do not depend on K cancel in likelihood ratios, so the same expression applies under adaptive allocation.

Let π be a full-support mixing distribution over the integer total vectors and define $M_s = \sum_K \pi(K) L_s(K)$. For the uniform product mixture, $M_s = \prod_g \mathrm{B}(c_g + 1, n_g - c_g + 1)$. For any fixed true total vector $K^\star$, the process $M_s / L_s(K^\star)$ is a nonnegative test supermartingale starting at one. Its maximal inequality yields the nested compatible set

$$\mathcal{K}_s = \bigcap_{u \preceq s} \left\{ K \in \prod_{g=1}^{G} \{0, \dots, N_g\} : L_u(K) \geq \alpha M_u \right\}, \tag{27}$$

which contains $K^\star$ at all retained checkpoints with probability at least $1 - \alpha$. Here $u \preceq s$ ranges over the retained checkpoints. Identifying $\mathcal{F}_s = \{F_K : K \in \mathcal{K}_s\}$ gives the confidence-compatible population set used by FAR. Because the construction is joint across groups and acquisition rounds, it does not require separate per-group or per-round error allocation.

B.3.2 INTEGER OPTIMIZATION AND ALLOCATION SEARCH

For a candidate allocation b and attainable future count vector z, let $\mathcal{K}_{s,b,z}$ denote the corresponding successor compatible set. Since the target is the finite-population mean, its sharp successor interval is obtained by minimizing and maximizing the total number of successes:

$$L_{b,z} = \frac{1}{N_{\mathrm{tot}}} \min_{K \in \mathcal{K}_{s,b,z}} \sum_g K_g, \qquad U_{b,z} = \frac{1}{N_{\mathrm{tot}}} \max_{K \in \mathcal{K}_{s,b,z}} \sum_g K_g, \qquad m_{b,z} = \frac{L_{b,z} + U_{b,z}}{2}. \tag{28}$$

Taking logarithms of the likelihood constraints in Eq. (27) makes them additive across groups, allowing the endpoint problems to be solved by dynamic programming over groupwise totals and

accumulated likelihood costs. An equivalent integer-linear formulation selects one admissible total for each group.

Substituting the Hypergeometric observation law into the robust midpoint objective gives

$$J_{\mathrm{mid}}(s,b) = \max_{K \in \mathcal{K}_s} \sum_z P_{s,b}(z \mid F_K) \left| m_{b,z} - \frac{\sum_g K_g}{N_{\mathrm{tot}}} \right|, \tag{29}$$

where the sum ranges over feasible future count vectors under allocation b. The same compatible population determines both the target value and the distribution of future observations. Unlike the Binomial case, the grouped Hypergeometric law depends on the vector of remaining group totals, so optimization proceeds directly over the integer state rather than through a scalar response–target projection.

For each evaluated allocation, successor endpoints and future-count expectations are computed exactly over the retained integer state, while the outer search over candidate allocations is numerical. Empty hypothetical successor branches use the prespecified global score midpoint $1/2$.

C Experimental setup

C.1 Experimental protocols

We consider two population-response acquisition settings and one finite-benchmark LLM evaluation setting. In the functional-recovery and observational-equivalence experiments, each selected query is answered by B sampled respondents, producing B binary responses whose success count is Binomial as described in Appendix B.2. In the LLM experiment, each benchmark prompt has one fixed outcome, so querying unseen prompts amounts to sampling without replacement from a finite population and gives the Hypergeometric model of Appendix B.3. Within each experiment, all methods share the same frozen split, initial history, candidate pool, target, query budget. The functional-recovery and observational-equivalence experiments additionally use a common terminal estimator, whereas the LLM baselines retain their respective score estimators.

Batch model and response sampling. The functional-recovery and observational-equivalence experiments in Sections 5.1 and 5.2 use the Binomial population-response model of Appendix B.2, with $B = 10$ and $B = 40$ responses per selected query, respectively. Each batch consists of responses from B distinct respondent records, modeled as independent draws from the underlying population F. The use of distinct records prevents duplicate use of the same realized response within a batch; records are sampled separately across queries and may therefore recur at different queries. For paired evaluation, response batches are pre-drawn for every candidate and shared across methods within each repetition. The LLM experiment follows the separate finite-population sampling protocol in Appendix C.5.

Functional-recovery protocol. We evaluate functional recovery on CIFAR-10H (Peterson et al., 2019), Duolingo-FR (Settles et al., 2018), and ASSISTments (Feng et al., 2009). For each dataset, response-model fitting, acquisition, and target evaluation use separate data roles, with an additional calibration split when candidate-specific response-model radii are required. The target functionals and candidate pools are fixed before acquisition and are constructed so that target responses are not available to the policy. All methods start from the same initial history and proceed for the same number of adaptive rounds, receiving the same pre-drawn response batch whenever a query is selected. Final recovery is evaluated by the midpoint of the resulting compatible target interval. Dataset-specific role splits, target definitions, candidate construction, and acquisition budgets are detailed in Appendix C.2.

Observational-equivalence protocol. We study observational equivalence on OpinionQA (Santurkar et al., 2023) and CES (Schaffner et al., 2025). For each dataset, we prespecify four target functionals and construct 200 initial states in which distinct compatible populations agree on the observed two-query response probabilities but imply different target values. From each shared state, all methods select one query from the same candidate pool and observe the same batch of 40 responses. We then recompute the successor compatible fibre and evaluate the remaining target

ambiguity and midpoint recovery under the common response model. This protocol directly tests how effectively each acquisition rule resolves target-relevant ambiguity left by the initial observations.

Population-pair discrimination diagnostic. For Figure 2, we measure how well each selected query distinguishes a fixed pair of target-conflicting populations that are observationally equivalent on the initial design. If the selected query has response probabilities p_F and p_G under the two populations, n responses induce Binomial count distributions $\mathrm{Binomial}(n, p_F)$ and $\mathrm{Binomial}(n, p_G)$. We report their equal-prior optimal discrimination error,

$$e_n(x; F, G) = \frac{1}{2} \sum_{c=0}^{n} \min\{\mathrm{Binom}(c; n, p_F), \mathrm{Binom}(c; n, p_G)\}.$$

For an initially observed query, observational equivalence implies $p_F = p_G$ and hence $e_n = 1/2$; informative new queries reduce this error as additional responses are collected.

Latent representation, response kernel, and target functional. For the functional-recovery experiments, the finite support $\mathcal{Z}_L = \{z_1, \ldots, z_L\}$ is fixed by the response model fitted on the designated response-model split. For each query x, $k_\theta(1 \mid x, z_\ell)$ is the predicted success probability of latent type z_ℓ, and FAR represents an unknown population by weights $w \in \Delta_L$ over this support. For a target reference set $\mathcal{R}_h$, we define

$$h(z_\ell) = \frac{1}{|\mathcal{R}_h|} \sum_{r \in \mathcal{R}_h} k_\theta(1 \mid r, z_\ell),$$

so that $T_h(F_w) = h^\top w$ is the population's mean predicted success probability on the target set. CIFAR-10H and Duolingo-FR use 121 latent support points, while ASSISTments uses the 734-point empirical latent support. Posterior particles used by posterior-based baselines are separate from the simplex weights optimized by FAR.

Replication and sampling. For functional recovery (Section 5.1; Appendix D.1), we use 30 random splits per dataset, with three targets for CIFAR-10H and Duolingo-FR and seven for ASSISTments, giving 90, 90, and 210 target–split trajectories per method. The one-step observational-equivalence evaluation (Section 5.2) uses 50 repetitions for each of four targets, yielding 200 shared states per dataset; the component ablations and safety-slack analysis reuse these states (Appendix D.2). The known-kernel multi-round evaluation uses five response seeds for each of four targets and three populations, giving 60 trajectories per dataset and method (Appendix D.1). For LLM score recovery (Section 5.3; Appendices C.5 and D.3), we use 60 initial-history sampling seeds shared across both target models, all methods, and additional budgets $B \in \{1, \ldots, 5\}$. Target-model responses are generated once and held fixed. RSL and AIPW additionally average 2,000 shared subset draws within each history to estimate conditional expected error.

C.2 Dataset construction and task definitions

CIFAR-10H. CIFAR-10H (Peterson et al., 2019) provides repeated human annotations of CIFAR-10 images, which we convert to binary correctness responses relative to the reference label. Each frozen split uses disjoint annotator roles for response-model fitting, calibration, acquisition, and target evaluation, with 30 candidate images and 78 disjoint reference images. We consider three target functionals, measuring mean annotator reliability over all reference images, the animal subset, and the vehicle subset. All methods share two initial queries followed by four adaptive queries, with $B = 10$ acquisition-role responses revealed for each selected image. Reference-image responses from the target role are used only for evaluation.

Duolingo-FR. Duolingo-FR is constructed from the French-learning track of SLAM 2018 (Settles et al., 2018). We use only word-bank translation (`reverse_tap`) and retain each learner's earliest attempt on a prompt. Learners are split into disjoint kernel, calibration, acquisition, and target-evaluation roles. Each split contains 30 candidate prompts and 78 reference prompts, with the reference set divided into short and long source-prompt strata. The three target functionals measure mean correctness over all reference prompts and over each of the two strata. All methods use the same two initial queries followed by four adaptive queries, with $B = 10$ responses per query.

ASSISTments. For ASSISTments (Feng et al., 2009), we define seven held-out skill-level target functionals from the student response data. For each target, the target skill is excluded from acquisition, and a frozen 20-problem candidate pool is constructed from the remaining eligible problems before held-out target responses are revealed. Response-model and target-evaluation students are kept disjoint from the acquisition population. All methods share the same candidate pool and pre-drawn response bank. Acquisition starts from two common initial problems and proceeds for four adaptive rounds with $B = 10$ responses per query.

OpinionQA. OpinionQA (Santurkar et al., 2023) is used for the controlled observational-equivalence experiment. We prespecify four target questions and construct shared initial states from two observed questions such that the resulting compatible population fibre contains populations that agree on the observed response probabilities but differ in target value. For each target, the target question and the two initial questions are excluded from acquisition, leaving a common candidate pool for all methods. Each method selects one additional query and observes the same batch of $B = 40$ responses. We use 50 repetitions for each target, giving 200 shared states in total.

Cooperative Election Study (CES). We construct the CES experiment from the 2024 Cooperative Election Study (Schaffner et al., 2025) using the same observational-equivalence protocol as OpinionQA. Four target questions are fixed in advance, and each shared state is defined by two initial questions whose compatible fibre leaves nonzero target ambiguity. The target question and the initial questions are excluded from the common acquisition pool. Each method then selects one query and receives the same batch of $B = 40$ responses. We again use 50 repetitions for each target, yielding 200 shared states.

C.3 Baselines

All baselines are evaluated under the same frozen action–response–target protocol as FAR. We preserve the acquisition principle of each published method while replacing task-specific modelling components, when necessary, with the response model and candidate set of the corresponding benchmark. Within each dataset, all methods receive the same history, candidate pool, response bank, query budget, and target definition, and are evaluated using the same terminal midpoint estimator.

ActiveCQ-TVR. ActiveCQ selects measurements to reduce posterior uncertainty in a target causal quantity (Gao & Sejdinovic, 2026). We use its total-variance reduction criterion, scoring each candidate by the expected reduction in posterior variance of the declared target after the next response batch and selecting the candidate with the largest reduction.

Adaptive Elicitation EIG/MCTS. Adaptive Elicitation uses a learned response model to simulate future observations and select questions that reduce uncertainty about latent information and downstream responses (Wang et al., 2025). We use its target-directed expected-information-gain criterion with finite lookahead search, generating predictive branches under the shared response model.

BECAT. BECAT selects questions so that the gradient induced by the acquired subset approximates the gradient obtained from the full question bank (Zhuang et al., 2023). We use its expected gradient-difference criterion and greedily select the candidate that most improves the corresponding gradient-coverage objective under the shared response model.

DGSM-IG. DGSM active learning targets uncertainty in a global sensitivity functional using a Gaussian-process surrogate (Belakaria et al., 2024). We use its information-gain variant, scoring each candidate by the expected information gain about the declared scalar target under the shared response model.

ITL. Information-based Transductive Learning (ITL) selects accessible observations to reduce uncertainty about a designated set of prediction targets (Hübotter et al., 2024). We score each candidate by its conditional mutual information with the prespecified target under the shared response model.

TSIED. Targeted Sequential Indirect Experiment Design (TSIED) adaptively selects experiments to tighten bounds on a partially identified target query (Ailer et al., 2024). We use its bound-directed criterion, selecting the candidate that minimizes the expected width of the compatible target interval after the next response.

Active Measurement. Active Measurement combines predictive modeling with adaptive sampling to estimate population aggregates (Hamilton et al., 2025). We apply its acquisition criterion in settings where the target admits the corresponding aggregate-measurement formulation, using the common candidate pool and response model.

μPOCA. μPOCA addresses partially observed, cost-aware acquisition by selecting observations according to their reduction in predictive uncertainty relative to acquisition cost (Astorga et al., 2024). Since all queries have equal cost in our benchmarks, we rank candidates by their expected reduction in target-relevant predictive uncertainty under the shared response model.

Random. Random is included in the observational-equivalence experiments as a non-adaptive reference. At each decision, it samples uniformly from the remaining candidate pool and is evaluated using the same responses, target interval, and terminal estimator as the other methods.

C.4 Evaluation metrics and statistical procedures

For the functional-recovery experiments, let $[L_j, U_j]$ be the compatible target interval after acquisition round j, with midpoint $m_j = (L_j + U_j)/2$ and width $W_j = U_j - L_j$, and let $T^\star$ denote the held-out target value. Unless stated otherwise, terminal metrics are evaluated after the final acquisition round K.

Mean absolute error (MAE). We measure terminal recovery accuracy by the absolute error of the midpoint estimator, $|m_K - T^\star|$, averaged across evaluation units. MAE is our primary point-recovery metric; lower values indicate more accurate recovery of the target functional.

P95 error. We report the 95th percentile of the terminal midpoint absolute error $|m_K - T^\star|$ across evaluation units to characterize upper-tail recovery error. Lower values indicate better tail performance.

Error area under the curve (Err-AUC). Err-AUC summarizes midpoint recovery error over the full acquisition trajectory. It is computed as the area under the error curve $\{|m_j - T^\star|\}_{j=0}^{K}$ across the common acquisition checkpoints. Lower values indicate better recovery performance over the acquisition budget.

Normalized width (N.Width). We measure contraction of the compatible target interval by

$$\mathrm{N.Width} = \frac{W_K}{W_0}.$$

This normalization accounts for differences in initial ambiguity across evaluation units. Values below one indicate contraction, and lower values correspond to less residual target ambiguity at the terminal round.

Covered contraction rate (CCR). CCR is the fraction of evaluation units for which the final fibre is nonempty, the terminal interval contains $T^\star$, and $W_K < W_0$. Higher values indicate that contraction is achieved while retaining coverage of the target.

Normalized endpoint error (N.EndErr). We evaluate the recovered interval by

$$\mathrm{N.EndErr} = \frac{\max\{|L_K - T^\star|,\ |U_K - T^\star|\}}{W_0},$$

where lower values indicate that both interval endpoints remain closer to the target. For the observational-equivalence experiments, we evaluate each selected candidate under the Binomial

model of Appendix B.2 using its response–target projection $\mathcal{R}_x^p(s)$. For state s, candidate x, and response count c, let $[L_{x,c}, U_{x,c}]$ denote the successor target interval and $m_{x,c} = (L_{x,c} + U_{x,c})/2$ its midpoint.

Worst-branch width ($W_{\max}$). We define the largest residual target ambiguity after querying x as

$$W_{\max}(s, x) = \max_{c:\mathcal{F}_{s,x,c}\neq\varnothing} \big(U_{x,c} - L_{x,c}\big).$$

It is the maximum successor-interval width over all compatible response branches; lower values indicate less worst-branch ambiguity.

Worst-population expected width ($W_{\exp}$). We define the largest expected successor width over compatible response–target pairs as

$$W_{\exp}(s, x) = \sup_{(p,t)\in\mathcal{R}_x^p(s)} \sum_{c=0}^{B} \mathrm{Binom}(c; B, p)\big(U_{x,c} - L_{x,c}\big).$$

Lower values indicate less expected residual ambiguity under the worst compatible population.

RTE. Our primary recovery metric for observational equivalence is

$$\mathrm{RTE}(s, x) = \sup_{(p,t)\in\mathcal{R}_x^p(s)} \sum_{c=0}^{B} \mathrm{Binom}(c; B, p)|m_{x,c} - t|.$$

RTE is the largest expected midpoint error over compatible populations; lower values indicate better target recovery.

Endpoint error (End.Err). We measure the expected error of the farther interval endpoint by

$$\mathrm{End.Err}(s, x) = \sup_{(p,t)\in\mathcal{R}_x^p(s)} \sum_{c=0}^{B} \mathrm{Binom}(c; B, p)\max\{|L_{x,c} - t|,\ |U_{x,c} - t|\}.$$

Lower values indicate that the recovered interval remains closer to the target.

Efficiency regret (Eff.Reg). We define

$$\mathrm{Eff.Reg}(s, x) = \mathrm{RTE}(s, x) - \min_{a\in\mathcal{A}_{\mathrm{eval}}(s)} \mathrm{RTE}(s, a),$$

where $\mathcal{A}_{\mathrm{eval}}(s)$ is the common evaluation candidate set. A value of zero indicates that the selected query minimizes RTE among the evaluated candidates.

Oracle agreement. Oracle agreement is the fraction of shared states in which the selected query attains the minimum RTE over $\mathcal{A}_{\mathrm{eval}}(s)$. Higher values indicate closer agreement with the common recovery criterion.

Statistical procedures. For the functional-recovery and observational-equivalence experiments, all comparisons are paired over the shared splits or states. We report paired 95% bootstrap confidence intervals using the corresponding replication unit. Pairwise significance is assessed with one-sided Wilcoxon signed-rank tests and Holm correction within each dataset and metric family. P95 uncertainty is estimated by replication-level bootstrap.

C.5 LLM SCORE-RECOVERY PROTOCOL

Model responses and scoring. We evaluate Qwen3-4B (Yang et al., 2025) and Phi-3.5-mini-instruct (Abdin et al., 2024) on a subject-balanced subset of MMLU-Pro containing 14 subjects and 40 prompts per subject. Each model generates one response per prompt using its official chat template. Qwen3-4B uses thinking mode with temperature 0.6, top-$p = 0.95$, top-$k = 20$, and a 32,768-token output limit, while Phi-3.5-mini-instruct uses greedy decoding with an 8,192-token limit. Both models receive zero-shot reasoning prompts requesting a final JSON answer. A response is scored as correct when generation terminates normally and the parsed answer matches the reference; truncation or parsing failure is counted as incorrect. Responses are generated once and fixed throughout evaluation, yielding full-set accuracies of $370/560$ for Qwen and $233/560$ for Phi.

Shared history and observation model. Each of the 60 evaluation histories begins with one uniformly sampled prompt from each subject, giving 14 initial observations. All methods receive the same initial prompts and responses and select five additional, previously unobserved prompts, for a total of 19 observations. Because each prompt has a fixed outcome and prompts are sampled without replacement, the observation process follows the Hypergeometric model of Appendix B.3. For FAR, $G = 14$, $N_g = 40$, $n_g = 1$, and the additional allocation satisfies $\sum_g b_g = 5$. The target is the accuracy over the full 560-prompt MMLU-Pro subset,

$$T(K) = \frac{\sum_g K_g}{560}.$$

Compatibility and optimization. FAR uses the joint likelihood-ratio compatible set of Appendix B.3.1 with $\alpha = 0.05$. Successor score intervals and future-count integration follow Appendix B.3.2. We set $\varepsilon_{\text{safe}} = \eta_0 = 0$ and search over structured allocations together with 48 random proposals followed by one local-search round.

Baselines. Random selects five of the remaining 546 prompts uniformly without replacement and estimates accuracy from all 19 observed outcomes. RSL (Zhang et al., 2025) uses random prompt subsets together with its learned benchmark-performance predictor. AIPW (Zhang et al., 2025) uses the same sampled subsets with its augmented inverse-probability-weighted estimator. For tinyBenchmarks (Maia Polo et al., 2024b), we use the GP-IRT implementation in BenchPred, fixing the 14 initial representatives and optimizing five additional representatives. SparseEval (Zhang et al., 2026) uses its MLP variant, with the 14 initial anchors fixed and five additional anchor positions selected from the unobserved prompts. RSL, AIPW, and tinyBenchmarks are fitted using responses from 100 historical models, while SparseEval additionally uses 20 historical validation models.

Evaluation metric. For method a, history s_h, and full-benchmark accuracy $T^\star$, we report mean absolute error in percentage points,

$$\mathrm{MAE}_a = \frac{100}{60} \sum_{h=1}^{60} \mathbb{E}\left[\left|\widehat{T}_a - T^\star\right| \mid s_h\right].$$

FAR and Random compute this conditional expectation exactly under the Hypergeometric model, while RSL and AIPW approximate it using 2,000 shared subset draws per history. For tinyBenchmarks and SparseEval, each initial history determines a single additional prompt set and corresponding score estimate, so no additional Monte Carlo averaging is required.

D Additional results

D.1 Recovery over the acquisition budget

Functional recovery. Figure 4 complements the terminal comparisons in Table 1 by showing recovery throughout acquisition. On CIFAR-10H, FAR achieves lower mean error than TSIED from the second adaptive round onward, with similar terminal errors. On Duolingo-FR, the gap becomes more pronounced in the final two rounds. ASSISTments shows a different trajectory: FAR's mean error increases, while its normalized interval width first decreases and then rises slightly above its initial level.

Observational equivalence. We further evaluate multi-round recovery under observational equivalence using the fitted response kernel. For each target, we construct two populations with identical initial response probabilities and maximally separated target values, together with their equal-weight mixture. All methods start from the same two initial queries and select four additional queries with ten responses each. Figure 5 shows that FAR achieves the lowest terminal mean absolute error on both datasets. At 60 responses, FAR and TSIED obtain 15.95 and 16.44 percentage points on OpinionQA, and 20.57 and 20.91 on CES, respectively. ActiveCQ performs slightly better than FAR at the first adaptive round on CES, while FAR attains the lowest error at the terminal budget.

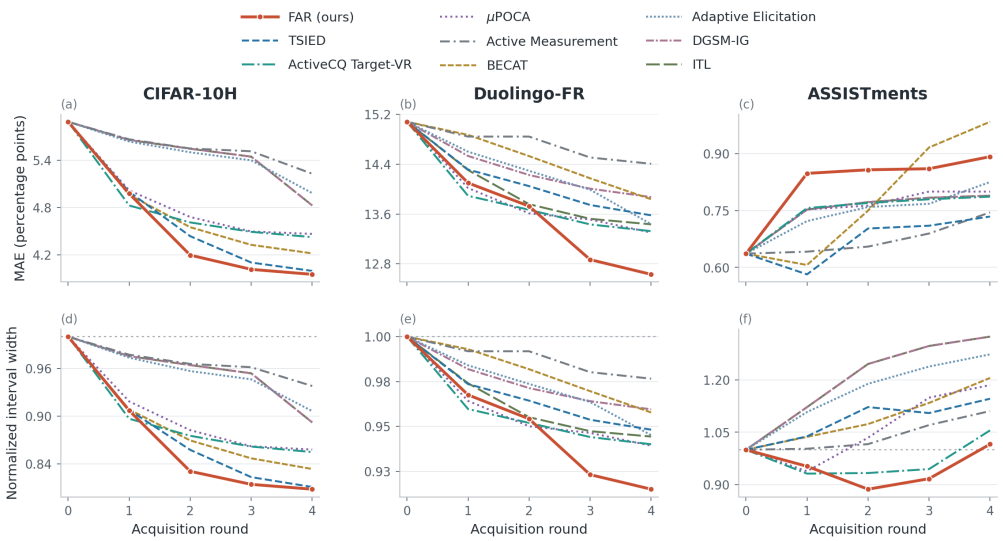


Figure 4: **Target recovery over the acquisition budget.** Columns correspond to CIFAR-10H, Duolingo-FR, and ASSISTments, with all nine methods shown. The top row reports mean absolute midpoint error in percentage points, and the bottom row reports interval width normalized by its initial value; the dotted line marks a ratio of one. Round zero denotes the shared initial observations, followed by four adaptive acquisition rounds. Curves average all targets over 30 splits per dataset.

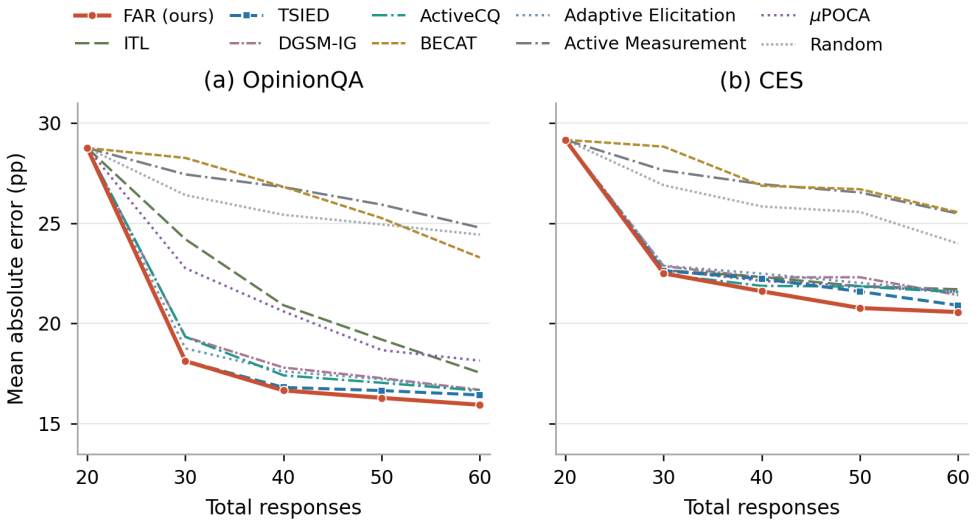


Figure 5: **Multi-round target recovery under observational equivalence.** All methods start from two shared initial queries and select four additional queries with ten responses each, increasing the total number of responses from 20 to 60. Curves show mean absolute midpoint error over four targets, three populations per target, and five response seeds.

D.2 Analysis under observational equivalence

Response–target geometry. We use a representative OpinionQA state to illustrate how different acquisition criteria interact with the same response–target geometry.

Figure 6 shows how query choice affects functional ambiguity. Each candidate query maps the compatible populations into a response–target region, where the horizontal axis gives the query response probability and the vertical axis gives the target value. A vertical section contains populations

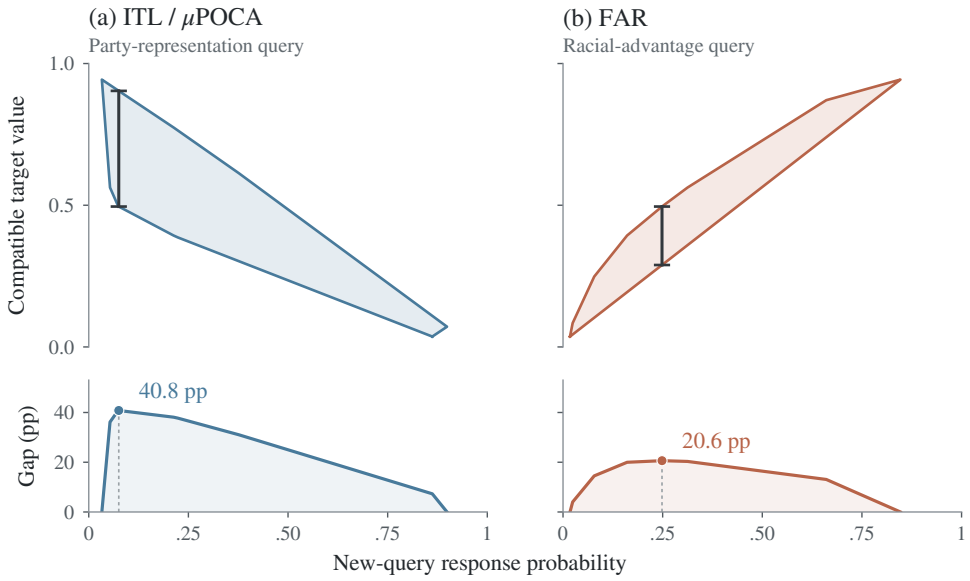


Figure 6: **Illustrative response–target geometry on OpinionQA.** Top: attainable response–target pairs for two candidate queries, with black segments marking their widest vertical sections. Bottom: the corresponding section widths as a function of response probability. ITL and μPOCA select the left query, whereas FAR selects the right query.

that induce the same response probability, and its height measures their remaining target disagreement. In this example, FAR selects the query with the smaller maximum section width.

Figure 7 extends the same example to finite-batch recovery. Both panels use the same compatible population fibre and the same target functional; only the selected query differs, so their least-favourable populations may occur at different target values. Accounting for the finite-batch response distribution and the corresponding midpoint estimator, the FAR-selected query has a smaller maximum recovery risk, 32.81 versus 37.87 percentage points for TSIED.

Component ablation. We evaluate the contributions of joint response–target risk modeling and midpoint readout on the same 200 OpinionQA and CES states used in the observational-equivalence experiments. The *w/o joint risk coupling* variant replaces FAR's joint response–target region with the product of its marginal intervals when scoring candidates, while retaining the same safety set and branch midpoints. The *w/o midpoint readout* variant keeps FAR's selected queries but replaces the midpoint with the posterior median used by TSIED, clipped to satisfy the same branchwise safety constraint. We also include TSIED with the same safety filter as an acquisition comparator.

Table 3: **Component ablation of FAR.** Mean worst-population expected absolute error (J) and worst-branch readout error (C) on OpinionQA and CES. Both metrics are reported in percentage points; lower is better. Bold denotes the lowest mean.

	OpinionQA		CES	
Variant	$J \downarrow$	$C \downarrow$	$J \downarrow$	$C \downarrow$
FAR	**28.5190**	**37.0796**	**31.4794**	**35.8334**
w/o joint risk coupling	29.6716	37.2869	31.6220	35.8973
w/o midpoint readout	30.6109	38.9336	34.1367	37.6151
TSIED + same safety	28.5442	37.0815	31.5425	35.8774

Here J denotes worst-population expected absolute error and C the worst-branch readout error; for midpoint readout, $C = W_{\max}/2$. Table 3 shows that removing joint risk coupling increases mean

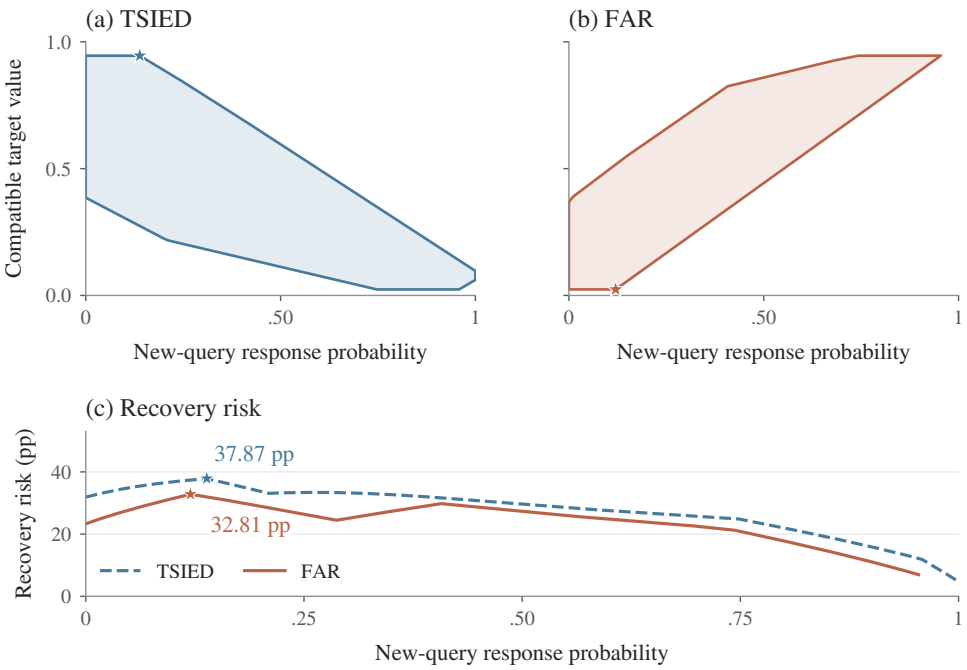


Figure 7: **Finite-batch view of the same OpinionQA example.** Top: response–target regions for the queries selected by TSIED and FAR. Bottom: the corresponding midpoint recovery risk as a function of response probability. Stars mark the respective worst-case points. TSIED selects the left query, whereas FAR selects the right query.

J by 4.04% on OpinionQA and 0.45% on CES, while replacing the midpoint increases it by 7.34% and 8.44%, respectively. FAR also attains slightly lower mean risk than the safety-matched TSIED comparator on both datasets.

Sensitivity to the safety slack. We vary $\varepsilon_{\text{safe}}$ over 200 shared states per dataset with additional budget $B = 40$. Figure 8 shows that increasing the safety slack admits more candidate queries, while FAR's mean recovery risk remains stable on OpinionQA and changes only modestly on CES.

D.3 Additional LLM evaluation results

Error distributions. Figure 9 complements the mean-error results for budgeted LLM score recovery by showing the cumulative distribution of absolute score-estimation error on MMLU-Pro for Qwen3-4B and Phi-3.5-mini after five additional queries. At any error tolerance, a higher curve indicates a larger fraction of outcomes recovered within that tolerance. The curves cross for some methods, showing that relative performance can vary with the required accuracy level and providing a more detailed view of the error distribution than mean error alone.

Performance across evaluation budgets. We vary the additional query budget from $B = 1$ to 5, yielding 15–19 observed responses. All budgets use the same 60 initial histories and the models, estimators, and settings in Appendix C.5, with each budget allocated independently from the initial state. Table 4 reports MAE, RMSE, P90, P95, and CVaR_{95} of the absolute score error. P90 and P95 characterize upper-tail errors, while CVaR_{95} reports the mean error within the worst 5% of the error distribution.

Across evaluation budgets, FAR shows the strongest overall performance on both models, particularly on RMSE and tail-sensitive metrics. On Phi-3.5-mini, FAR achieves the lowest MAE, RMSE, P90, P95, and CVaR_{95} at nearly all budgets, indicating consistently accurate score recovery across both typical and high-error cases. On Qwen3-4B, Random is slightly better in MAE at the two smallest

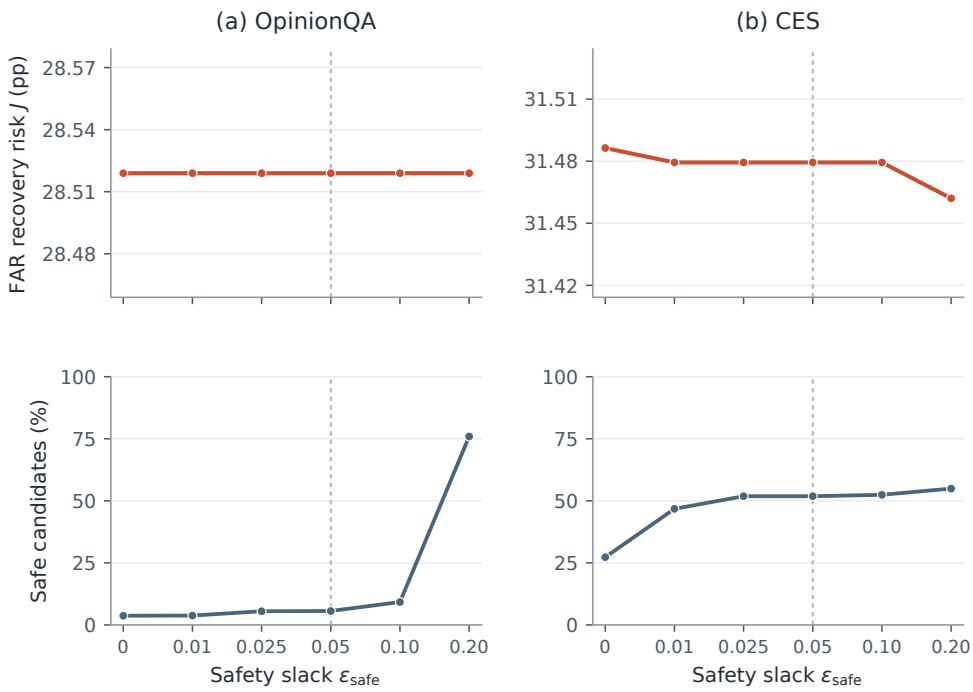


Figure 8: **FAR sensitivity to the safety slack.** Top: mean worst-population expected midpoint error J in percentage points. Bottom: mean fraction of the 27 candidates satisfying the safety constraint. Vertical dotted lines mark $\varepsilon_{\text{safe}} = 0.05$. Results are averaged over 200 shared states per dataset.

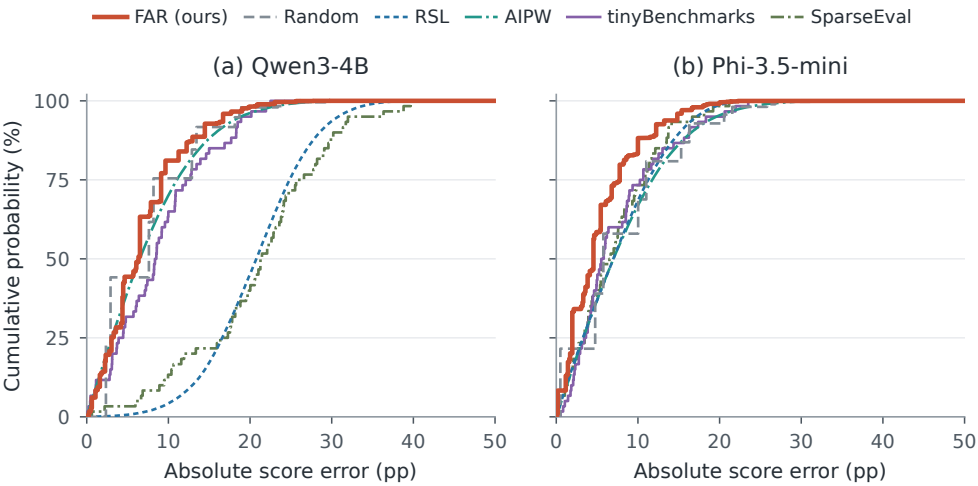


Figure 9: **Distribution of LLM score-estimation errors.** Cumulative distributions of absolute score error on MMLU-Pro for Qwen3-4B and Phi-3.5-mini, averaged over 60 shared histories and subsequent acquisition outcomes. At a given error tolerance, a higher curve indicates a greater probability of recovering the benchmark score within that tolerance.

budgets, whereas FAR becomes best for $B \geq 3$ and maintains the lowest RMSE and CVaR_{95} throughout. As the budget increases, FAR's MAE generally decreases on both models, with the advantage becoming clearer once several additional prompts can be allocated across subjects. These

Table 4: **LLM score recovery across evaluation budgets.** B denotes the number of additional queries after 14 initial responses. All metrics are reported in percentage points; lower is better. Bold denotes the best value within each model and budget.

		Qwen3-4B					Phi-3.5-mini				
B	Method	MAE	RMSE	P90	P95	CVaR_{95}	MAE	RMSE	P90	P95	CVaR_{95}
1	**FAR (ours)**	8.467	**9.500**	15.446	**15.446**	**19.411**	**5.924**	**7.166**	**12.946**	**12.946**	**15.948**
	Random	**7.540**	9.898	**13.929**	19.405	23.625	8.737	11.149	18.393	21.607	25.892
	RSL	21.224	22.074	29.418	31.569	34.081	7.877	9.729	16.576	18.152	19.710
	AIPW	8.313	10.513	18.234	20.832	24.567	9.237	11.385	18.477	22.052	26.493
	tinyBenchmarks	9.071	10.755	16.312	19.048	21.772	8.698	10.764	18.201	20.697	23.571
	SparseEval	19.699	21.120	30.394	32.256	36.406	7.725	9.700	16.018	19.066	22.902
2	**FAR (ours)**	7.798	**9.070**	**14.821**	16.875	**19.230**	**5.920**	**7.158**	**12.321**	**14.286**	**16.395**
	Random	**7.796**	9.724	16.071	**16.071**	23.625	8.640	10.878	16.607	22.857	25.305
	RSL	21.069	21.947	29.099	31.434	33.962	7.830	9.597	16.094	17.957	19.654
	AIPW	8.165	10.266	16.885	20.677	24.000	8.933	11.049	17.714	21.474	25.558
	tinyBenchmarks	9.234	11.191	17.553	18.394	26.195	8.731	11.022	18.050	20.925	23.584
	SparseEval	20.706	22.378	30.616	31.353	36.535	8.021	9.736	16.836	18.327	20.332
3	**FAR (ours)**	**7.394**	**8.642**	**12.946**	**16.161**	**18.716**	**6.431**	**7.626**	**11.518**	**13.929**	**16.635**
	Random	7.574	9.550	16.282	19.013	22.498	8.246	10.623	18.078	23.099	24.987
	RSL	20.916	21.807	28.866	31.196	33.689	7.762	9.459	15.644	17.698	19.526
	AIPW	7.984	10.009	16.407	19.934	23.449	8.642	10.712	17.412	20.768	24.664
	tinyBenchmarks	9.239	10.990	15.086	19.651	25.416	7.676	9.940	17.312	20.413	22.294
	SparseEval	19.354	21.298	30.525	33.149	34.650	7.662	9.272	14.215	16.338	18.632
4	**FAR (ours)**	**7.104**	**8.532**	**13.750**	**16.786**	**19.195**	**5.690**	**7.156**	**12.500**	**13.929**	**17.018**
	Random	7.334	9.379	16.071	17.262	21.957	8.327	10.382	19.385	19.504	23.825
	RSL	20.805	21.698	28.674	30.989	33.477	7.716	9.364	15.392	17.487	19.459
	AIPW	7.818	9.796	16.117	19.329	22.926	8.396	10.437	16.977	20.318	24.146
	tinyBenchmarks	8.926	10.654	17.748	20.015	21.356	7.045	9.204	14.769	17.425	23.214
	SparseEval	19.532	21.464	30.322	32.980	35.413	7.959	9.508	14.622	17.844	20.115
5	**FAR (ours)**	**6.901**	**8.483**	14.464	**16.696**	**19.835**	**5.344**	**6.901**	**12.232**	**14.821**	**17.123**
	Random	7.417	9.213	**13.440**	18.703	21.402	7.959	10.154	16.288	20.555	23.573
	RSL	20.710	21.600	28.514	30.809	33.290	7.645	9.249	15.116	17.181	19.231
	AIPW	7.678	9.615	15.834	18.848	22.471	8.136	10.122	16.532	19.615	23.355
	tinyBenchmarks	8.888	10.618	18.299	18.898	21.536	7.510	9.340	16.245	18.329	22.028
	SparseEval	20.883	22.527	30.190	31.986	38.272	7.189	8.757	13.717	16.638	19.495

results suggest that ambiguity-aware allocation is especially beneficial when the evaluation budget is large enough to distribute queries selectively, while also providing more stable control of large estimation errors.