FOVEAL SEARCH: LEARNING PROXY COMBINATIONS FOR CERTIFIED CORPUS DECISIONS

Anonymous authors
Research manuscript

ABSTRACT

LLMs make it possible to draw nuanced conclusions from large collections of unstructured records. But using an LLM to judge every record is costly, especially when the goal is only an overall conclusion. Cheaper signals can be obtained for the entire collection, yet relying on them alone risks drawing the wrong conclusion. Foveal Search addresses this gap by acquiring a limited number of reference judgments and making each one serve two purposes: it contributes immediately to the decision and helps determine how to combine the cheaper signals for future judgments. The method guarantees a prescribed level of decision error even when those signals are poor. We also show when the information gained from learning their combination can outweigh the judgments spent learning it, with bounds that count every acquired judgment. Across evaluated settings, Foveal Search often uses fewer judgments than methods that ignore the cheaper signals while maintaining high observed correct completion. On held-out data, learning separate weights for the signals saves 3.0% and 4.6% of judgments compared with learning one weight for their average. Source Code: https://anonymous.4open.science/r/30228.

1 INTRODUCTION

LLMs can make nuanced judgments about unstructured records, enabling analyses that depend on the meaning of each item rather than a predefined field (Gilardi et al., 2023; Patel et al., 2025). In applications such as evaluating generated responses or reviewing documents, the desired result is often an aggregate conclusion about a corpus (Saad-Falcon et al., 2024). Obtaining an LLM judgment for every record can nevertheless be expensive, especially as corpora grow or evaluation criteria change. The challenge is to reach a dependable overall conclusion without paying excessive cost.

One possibility is to use inexpensive signals available across the corpus, such as retrieval scores or predictions from smaller models (Lewis et al., 2020; Saad-Falcon et al., 2024). Such proxy signals can help prioritize or estimate judgments, but they may miss distinctions that matter to the reference evaluator. Prediction-powered inference addresses this tension by correcting abundant predictions with a smaller set of reference labels (Angelopoulos et al., 2023b); ARES applies this idea to evaluating retrieval-augmented generation (Saad-Falcon et al., 2024). For sequential corpus decisions, however, a further question arises: can the same labels both support the current decision and improve future use of inexpensive signals, with explicit accounting for the total label cost?

We introduce *Foveal Search* to answer this question for a fixed corpus. The method starts with proxy predictions for every item and acquires reference labels uniformly without replacement. Each label first updates the evidence for a decision and then helps fit the proxy combination used for subsequent labels. Because the complete proxy table is available in advance, Foveal Search can calculate how the proxies vary together without additional reference judgments; the acquired labels need only reveal how they relate to the reference outcomes. A useful combination reduces the variation left to be resolved by future labels. If the proxies prove unhelpful, the decision-error guarantee still holds, and full verification resolves any remaining uncertainty.

The potential savings must also pay for learning the combination. Existing work studies predictor combinations and resource allocation under a prescribed budget (Cowen-Breen et al., 2026), sequential inference with evolving predictors (Csillag et al., 2025), and valid inference while sampling a finite population without replacement (Waudby-Smith and Ramdas, 2020b). Our analysis focuses on the

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.

total label cost when the same sequentially acquired labels are used for both learning and a fixed-corpus decision. For Foveal Search's exact checkpoint learner, we bound the additional evidence variance due to learning even when stopping depends on the observed labels. This bound grows logarithmically with corpus size at fixed proxy rank. Under explicit evidence-growth conditions, it yields guarantees for correct completion within a label budget and bounds on expected label cost. Every acquired label is counted, with numerical recovery accounted for separately.

We evaluate whether these guarantees accompany useful savings in practice. Across synthetic settings and two real-world datasets, study-specific fixed configurations often require fewer labels than label-only controls while achieving high observed correct completion. Prediction-powered baselines can be cheaper in favorable settings, so these comparisons do not imply a universal cost advantage. Matched held-out ablations isolate the benefit of learning separate proxy weights: relative to a single weight on their average, they reduce mean label cost by 3.0% and 4.6% on the two datasets.

2 Problem Formulation

Our target is the aggregate judgment of a given, finite collection of records. Fixing this collection makes the desired answer precise: its reference labels determine a single verdict, and verifying every item can always reveal that verdict. Let $E = \{1, \dots, N\}$ denote the corpus, with $N \geq 2$. Each item $i \in E$ has an unknown reference label $y_i \in \{0, 1\}$ and an observed proxy vector $z_i \in [0, 1]^m$, where $m \geq 1$. Verifying an item reveals y_i at unit cost. The full proxy table is available before verification, as in prediction-assisted settings with abundant predictions and scarce reference labels (Angelopoulos et al., 2023b). We place no accuracy or correlation requirement on these proxies.

Given a threshold $\tau \in (0, 1)$, the corpus mean μ determines the target verdict:

$$\mu = \frac{1}{N} \sum_{i \in E} y_i, \qquad a = \mathbf{1}\{\mu > \tau\}. \tag{1}$$

The strict inequality makes equality a negative verdict. A procedure verifies distinct items sequentially, stops after acquiring $T \leq N$ labels, and returns $\widehat{a} \in \{0, 1\}$. Its actions may depend on the complete proxy table and the previously observed item–label pairs. Foveal Search draws uniformly from the remaining items, so its labels form a sample without replacement from the fixed corpus (Waudby-Smith and Ramdas, 2020b). At $T = N$, the verdict is known exactly.

For a prescribed error tolerance $\delta \in (0, 1)$, we aim to reduce expected verification cost $\mathbb{E}[T]$ s.t.

$$\Pr\{\widehat{a} \neq a\} \leq \delta. \tag{2}$$

The corpus and proxy table remain fixed; probability and expectation are over the procedure's randomization. Every acquired label counts toward T, whether it is used for the decision, for learning a proxy combination, or for both. Thus the objective measures the net benefit of learning from proxies after paying for the labels needed to use them. Appendix N summarizes the notation.

3 Foveal Search

Section 2 provides inexpensive proxy outputs for every item, while reference labels must be purchased. Foveal Search uses the proxies to reduce the variation left for those labels to resolve. Each purchased label first updates evidence for the corpus decision, then helps fit the proxy combination used for later queries. We describe the fit and the decision rule in turn. Section 4 analyzes their error control and total label cost.

3.1 Learning Proxy Combinations

A proxy can help even when some of its individual predictions are wrong: its known corpus mean can be combined with a labeled estimate of its mean error. Foveal Search learns a linear combination $\beta^\top z_i$, with $\beta \in \mathbb{R}^m$ shared across the corpus. For any fixed β,

$$\mu = \beta^\top \bar{z} + \frac{1}{N} \sum_{i \in E} (y_i - \beta^\top z_i).$$

The first term is known from the proxies; purchased labels estimate the residual term. This prediction-and-correction structure is related to prediction-powered inference (Angelopoulos et al., 2023b). We fit β to reduce residual variance, rather than to calibrate predicted probabilities.

Combining proxies requires knowing how they vary together and how they relate to the reference labels. The complete proxy table gives the first relationship without purchasing labels. Define

$$\bar{z} = \frac{1}{N}\sum_{i\in E} z_i, \qquad x_i = z_i - \bar{z}, \qquad G = \frac{1}{N}\sum_{i\in E} x_i x_i^\top. \tag{3}$$

The unknown relationship is the cross moment $b = N^{-1}\sum_{i\in E} x_i y_i$. Because $\sum_i x_i = 0$, this also equals $N^{-1}\sum_i x_i(y_i - \mu)$. Together, G and b determine residual variance:

$$\sigma^2(\beta) = \mathrm{Var}_E(y_i - \beta^\top z_i) = \mu(1-\mu) - 2b^\top\beta + \beta^\top G\beta. \tag{4}$$

Here Var_U is uniform variance on a nonempty set U, with denominator $|U|$. Since $\mu(1-\mu)$ does not depend on β, fitting needs an estimate of b but can retain the known G. This choice underlies the calibration bound in (11). Multiple-predictor inference also studies how predictor dependence affects efficient combinations, but under a different budgeted estimation setting (Cowen-Breen et al., 2026).

We restrict coefficients to $\mathcal{K} = \{\beta \in \mathbb{R}^m \mid \|\beta\|_1 \le R\}$, with $R \ge 1$ fixed before verification. Since $z_i \in [0,1]^m$, this constraint bounds the proxy adjustment used in the evidence process and permits safe betting fractions. Coefficients may be signed and need not sum to one.

We acquire labels by sampling unverified items uniformly without replacement. Let I_s be the item queried at step s and $S_k = \{I_1, \dots, I_k\}$ the first k verified items, with $S_0 = \varnothing$. Each prefix of a prespecified size is then a uniform corpus subset, as required by the finite-population calibration analysis (Waudby-Smith and Ramdas, 2020b). At a fitting checkpoint, we use

$$\widehat{b}_k = \frac{1}{k}\sum_{i\in S_k} x_i(y_i - \tfrac{1}{2}), \qquad \widehat{\beta}_k \in \operatorname*{arg\,min}_{\beta\in\mathcal{K}}\{\beta^\top G\beta - 2\widehat{b}_k^\top\beta\}. \tag{5}$$

Subtracting $1/2$ preserves b because $\sum_i x_i = 0$. For binary labels it also gives $(y_i - \frac{1}{2})^2 = 1/4$, the identity used in (11).

We refit when the number of purchased labels doubles, at $\mathcal{T}_N = \{1, 2, 4, \dots\} \cap \{1, \dots, N-1\}$. A fit is reused until the next checkpoint or termination. Write β_t for the coefficients used at query t, with $\beta_1 = \mathbf{0}$. This schedule limits fitting frequency and the number of queries affected by each fit; both enter the allowance B_N in (12). Every label used to fit these coefficients also contributes to decision evidence.

3.2 Constructing Decision Evidence

Lower residual variance is useful only if changing the weights does not change what the evidence tests. Before query t, let $V_t = S_{t-1}$ be the verified set, $U_t = E \setminus V_t$ the unread set, $n_t = |U_t|$, and $A_t = \sum_{i\in V_t} y_i$. At the boundary $\mu = \tau$, the unread labels must sum to $N\tau - A_t$. We give the proxy control that same total:

$$b_t^{\mathrm{null}} = \frac{N\tau - A_t}{n_t}, \qquad \bar{z}_{U_t} = \frac{1}{n_t}\sum_{i\in U_t} z_i, \qquad f_{\beta,t}(i) = b_t^{\mathrm{null}} + \beta^\top(z_i - \bar{z}_{U_t}). \tag{6}$$

The proxy adjustment sums to zero over U_t, so $\sum_{i\in U_t} f_{\beta,t}(i) = N\tau - A_t$ for every β. Controls remain unclipped, even if they fall outside $[0,1]$, to preserve this identity.

The uniformly sampled item I_t supplies the residual correction

$$D_t = \frac{n_t}{N}\{y_{I_t} - f_{\beta_t,t}(I_t)\}. \tag{7}$$

The factor n_t/N puts the unread-set correction on the full-corpus scale. Let $\mathcal{F}_{t-1}$ contain the complete proxy table and the previous item–label pairs. Since β_t is fixed before observing y_{I_t},

$$\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = \mu - \tau, \qquad v_t = \mathrm{Var}(D_t \mid \mathcal{F}_{t-1}) = \left(\frac{n_t}{N}\right)^2 \mathrm{Var}_{U_t}(y_i - \beta_t^\top z_i). \tag{8}$$

The first identity supports Proposition 1; the second connects the fit to Theorem 1. Each label updates evidence using its pre-query coefficients before it contributes to a future fit. This ordering keeps

Algorithm 1 Foveal Search

Require: $\{z_i\}_{i\in E}$, VERIFY, τ, δ, R, and mixture $(\mathcal{C}, \{\pi_c\})$
Ensure: Verdict and acquired-label count $(\widehat{a}, T)$

Compute $\bar{z}, \{x_i\}, G$ ▷ known proxy geometry, (3)
$U \leftarrow E$; $A \leftarrow k \leftarrow 0$; $s \leftarrow \beta \leftarrow \mathbf{0}$; $E_c^{\pm} \leftarrow 1$ for all $c \in \mathcal{C}$
while $U \neq \varnothing$ **do**
  if $A > N\tau$ **then return** $(1, k)$; **if** $A + |U| \leq N\tau$ **then return** $(0, k)$ ▷ exact bounds
  $M^{\pm} \leftarrow \sum_{c\in\mathcal{C}} \pi_c E_c^{\pm}$ ▷ combine evidence
  if $M^{+} \geq 2/\delta$ **then return** $(1, k)$; **if** $M^{-} \geq 2/\delta$ **then return** $(0, k)$ ▷ positive side first
  $n \leftarrow |U|$; fix $f_c, \lambda_c^{\pm}$ from $\mathcal{F}_k$ ▷ pre-query controls; (6)
  $I \sim \mathrm{Unif}(U)$; $y_I \leftarrow \textsc{Verify}(I)$; $k \leftarrow k + 1$ ▷ one paid label
  $D_c \leftarrow (n/N)\{y_I - f_c(I)\}$ for all $c \in \mathcal{C}$ ▷ residual correction, (7)
  $E_c^{\pm} \leftarrow E_c^{\pm}(1 \pm \lambda_c^{\pm} D_c)$ for all $c \in \mathcal{C}$ ▷ certified evidence first
  $s \leftarrow s + x_I(y_I - \frac{1}{2})$ ▷ reuse the label, (5)
  if $k \in \mathcal{T}_N$ **then** $\beta \leftarrow \widehat{\beta}_k$ with $\widehat{b}_k = s/k$ ▷ future weights, (5)
  $A \leftarrow A + y_I$; $U \leftarrow U \setminus \{I\}$ ▷ remove the verified item
end while
return $(\mathbf{1}(A/N > \tau), N)$ ▷ all labels verified
On recoverable fit/evidence failure: verify all unrecorded items; then **return** $(\mathbf{1}(N^{-1}\sum_{i\in E} y_i > \tau), N)$

the updates predictable, as required in sequential betting and e-value methods (Waudby-Smith and Ramdas, 2024; Csillag et al., 2025).

We maintain evidence from the learned combination alongside fixed reference controls: zero adjustment, individual proxies, and their average. Each component $c \in \mathcal{C}$ forms a correction $D_{c,t}$ by (7). Its nonnegative betting fraction $\lambda_{c,t}^{\pm}$ is fixed in advance or chosen from pre-query bounds so that $1 \pm \lambda_{c,t}^{\pm} D_{c,t} \geq 0$ for every possible next observation. With $E_{c,0}^{\pm} = 1$ and fixed weights $\pi_c \geq 0$ summing to one, update

$$E_{c,t}^{\pm} = E_{c,t-1}^{\pm}(1 \pm \lambda_{c,t}^{\pm} D_{c,t}), \qquad M_t^{\pm} = \sum_{c\in\mathcal{C}} \pi_c E_{c,t}^{\pm}. \tag{9}$$

The positive and negative mixtures accumulate evidence against $\mu \leq \tau$ and $\mu \geq \tau$, respectively. All components share the purchased labels. The mixture provides the anytime guarantee in Proposition 1; its fixed-fraction learned-control components also support Theorems 2 and 3. Appendix E.1 specifies the components and fractions.

Before each query, we return an exact positive verdict if $A_t > N\tau$, or an exact negative verdict if $A_t + n_t \leq N\tau$. Otherwise, we return positive if $M_{t-1}^{+} \geq 2/\delta$ or negative if $M_{t-1}^{-} \geq 2/\delta$, checking the positive side first. The threshold allocates error $\delta/2$ to each side. Complete verification returns the exact verdict at $T = N$ and is the fallback for recoverable fit or evidence-certification failures. These rules ensure completion and count every acquired label in the cost results of Section 4. In Algorithm 1, k counts acquired labels and s accumulates the cross moment.

4 Theoretical Analysis

Foveal Search uses each purchased label first for decision evidence and then to fit weights for future queries. We establish that this reuse preserves error control, bound the variance cost of learning through stopping, and translate that bound into completion and label-cost guarantees. Throughout, the binary label–proxy table is fixed, queries are uniform without replacement, and coefficients lie in $\mathcal{K}$. These sampling assumptions follow the finite-population setting studied in Serfling (1974); Bardenet and Maillard (2015); Waudby-Smith and Ramdas (2020b). Proofs and bounded-response extensions appear in Appendices A and B.

4.1 Decision Correctness

Before each query, the control in (6), its coefficients, and its betting fraction are fixed from the available history. Hence, learning from previous labels does not change the conditional mean $\mu - \tau$

in (8). This predictable-update requirement is standard in time-uniform inference and betting-based evidence processes (Howard et al., 2021; Waudby-Smith and Ramdas, 2024; Csillag et al., 2025). Under either null hypothesis, the corresponding nonnegative mixture in (9) is a supermartingale. The logical bounds and census fallback in Algorithm 1 give exact verdicts when invoked.

Proposition 1 (Anytime validity)**.** *For every fixed binary label–proxy table and prescribed $\delta \in (0, 1)$, Foveal Search returns a verdict $\widehat{a}$ satisfying*

$$\mathbb{P}\{\widehat{a} \neq \mathbf{1}(\mu > \tau)\} \leq \delta, \qquad T \leq N. \tag{10}$$

The guarantee permits any predictable coefficients in $\mathcal{K}$ used with the specified evidence construction, including coefficients learned from previously acquired labels.

Thus proxy quality affects label cost, not the error guarantee. Bounding that cost requires analyzing the particular learner in (5).

4.2 Calibration at a Data-Dependent Stop

The learner minimizes the residual variance in (4), which enters the conditional evidence variance in (8). Let $\beta_* \in \arg\min_{\beta \in \mathcal{K}} \sigma^2(\beta)$ be an analysis-only optimum, $\sigma_*^2 = \sigma^2(\beta_*)$, and $r = \operatorname{rank}(G)$. Since (3) supplies G from the complete proxy table, fitting estimates only the proxy–label cross moment. For an exact fit on a uniform k-subset,

$$\mathbb{E}[\sigma^2(\widehat{\beta}_k) - \sigma_*^2] \leq \frac{N-k}{k(N-1)}\chi, \qquad \chi = \frac{r}{4} - b^\top G^\dagger b \geq 0. \tag{11}$$

Here $G^\dagger$ is the Moore–Penrose inverse. The finite-population factor vanishes at census, and singular G is allowed. MultiPPI also studies predictor combinations and their estimation costs, but under prescribed observation budgets (Cowen-Breen et al., 2026).

Stopping depends on the observed labels, so prefixes conditional on continued execution need not be uniform subsets. We instead charge nonnegative fitting excess along a complete uniform permutation. For the doubling checkpoints $\mathcal{T}_N$ in Section 3.1, define

$$\Delta_0 = \mu(1-\mu) - \sigma_*^2, \quad \Xi_N = \sum_{k \in \mathcal{T}_N} \frac{\min(k, N-k)(N-k)}{k(N-1)}, \quad B_N = \Delta_0 + \chi \Xi_N. \tag{12}$$

The first query contributes Δ_0. A fit after k labels serves at most $\min(k, N-k)$ later queries.

Theorem 1 (Calibration at a data-dependent stop)**.** *For the exact-fit doubling learner and every observation-based stopping time $1 \leq T \leq N$,*

$$\boxed{\mathbb{E}\left[\sum_{t=1}^{T} v_t\right] \leq \sigma_*^2 \mathbb{E}[T] + B_N \leq \sigma_*^2 \mathbb{E}[T] + \frac{1 + r\lceil \log_2 N \rceil}{4}.} \tag{13}$$

Every label used for calibration is included in T.

The allowance B_N measures cumulative-variance cost, not extra labels. It grows at most logarithmically with N at fixed rank. Appendix C compares this checkpoint-specific result with online-learning regret bounds (Hazan, 2023a); Appendix D examines other fitting schedules.

4.3 Correct Completion Within a Budget

Validity alone permits completion at census. To bound earlier completion, we analyze a fixed-fraction component using the learned control in (9). Its log evidence grows with the decision margin, while Theorem 1 bounds the variance that can slow that growth. Prediction-powered e-values provide related analyses of sequential evidence with evolving predictors (Csillag et al., 2025).

Let $T_0, \widehat{a}_0$ denote label count and verdict with exact fits, certified evidence arithmetic, and no numerical recovery. Set $d = |\mu - \tau| > 0$ and $a = \mathbf{1}(\mu > \tau)$. Index the included fixed-fraction learned-control components by j. Their fractions satisfy $0 < \lambda_j(1+R) < 1/2$; p_j is the corresponding one-sided mixture weight (Appendix E.1). For a fixed deadline t, define

$$\begin{aligned} Q_t &= \sigma_*^2 \left(t - \frac{t(t-1)}{2N}\right) + B_N, \\ h_j &= \log \frac{2}{\delta p_j} + \zeta, \qquad \kappa_j = \frac{\lambda_j}{1 - \lambda_j(1+R)}. \end{aligned} \tag{14}$$

The first term of Q_t reflects the shrinking unread set, while B_N accounts for learning. Certified arithmetic uses $\zeta = 1$; exact arithmetic permits $\zeta = 0$.

Theorem 2 (Correct completion by a verification budget)**.** *For the exact doubling learner,* $1 \le t < N$ *and* $0 < \eta < 1$*, if any included component satisfies*

$$t \log(1 + \lambda_j d) \ge h_j + \frac{\kappa_j^2 Q_t}{2\eta} + \kappa_j \sqrt{\frac{1-\eta}{\eta} Q_t}, \tag{15}$$

then $\mathbb{P}(T_0 > t) \le \eta$ *and* $\mathbb{P}(\widehat{a}_0 = a,\, T_0 \le t) \ge [1 - \delta - \eta]_+$.

Here $[x]_+ = \max\{x, 0\}$. The parameter δ controls incorrect verdicts, whereas η controls missed deadlines. At a fixed margin and component, smaller σ_*^2 or B_N makes the sufficient condition easier to satisfy. The logical bounds in Section 3.2 give a second stopping route: their stopping time L satisfies $T_0 \le L \le N$ on the same query order. Appendix A.6 combines both routes, refines Q_t, and accounts for numerical recovery.

4.4 Expected Verification Cost

We finally bound expected cost across all query orders. As in prediction-powered inference, proxy predictions are corrected with labeled residuals (Angelopoulos et al., 2023b). Here those labels also learn the weights, so their contribution to evidence variance must be included. Theorem 1 supplies that bound for the same paid trajectory.

Theorem 3 (Expected completion cost)**.** *Suppose* $d > 0$*. For a fixed-fraction learned-control component* j *in the declared evidence mixture* (95)*, let* λ_j *and* p_j *be its betting fraction and prior weight, and define*

$$c_j = \lambda_j d - \lambda_j^2 (d^2 + \sigma_*^2), \qquad A_j(\delta) = \log \frac{2}{\delta p_j} + \log \frac{3}{2} + 1. \tag{16}$$

For every such component with $c_j > 0$*,*

$$\boxed{\mathbb{E}[T_0] \le \min\left\{ N,\, \frac{A_j(\delta) + \lambda_j^2 B_N}{c_j} \right\}.} \tag{17}$$

The condition $c_j > 0$ requires margin-driven growth to exceed the benchmark's quadratic penalty. In (17), $A_j(\delta)$ covers the evidence threshold, prior weight, terminal update, and rounding; $\lambda_j^2 B_N$ accounts for learning. The population quantities used to evaluate this certificate are not inputs to the deployed learner.

Recoverable numerical failure F triggers census in Algorithm 1. If a bound $\mathbb{P}(F) \le \nu$, $\nu \in [0, 1]$, is justified, then

$$\mathbb{E}[T] \le C_\delta, \qquad C_\delta = \min\left\{ N,\, \min_{j : c_j > 0} \frac{A_j(\delta) + \lambda_j^2 B_N}{c_j} + N\nu \right\}. \tag{18}$$

The inner minimum covers included components; an empty minimum is $+\infty$, giving $C_\delta = N$. The no-recovery policy uses $\nu = 0$, but zero observed recoveries do not justify that value for an implementation.

Corollary 1 (Error control and verification savings)**.** *Under the assumptions of Theorem 3, suppose numerical failure* F *triggers census recovery with a justified bound* $\mathbb{P}(F) \le \nu$*,* $\nu \in [0, 1]$*. Proposition 1 and the implemented-cost bound in* (18)*, which defines* C_δ*, imply*

$$\boxed{\mathbb{P}\{\widehat{a} \ne \mathbf{1}(\mu > \tau)\} \le \delta, \qquad \mathbb{E}[N - T] \ge N - C_\delta.} \tag{19}$$

The smallest bound over included fixed components selects a certificate, not a new deployed component. For example, $C_\delta \le 0.8N$ certifies at least 20% expected savings. Holding the calibration and evidence rules and the recovery bound fixed, a larger δ weakly improves this certificate by lowering its $\log(1/\delta)$ term.

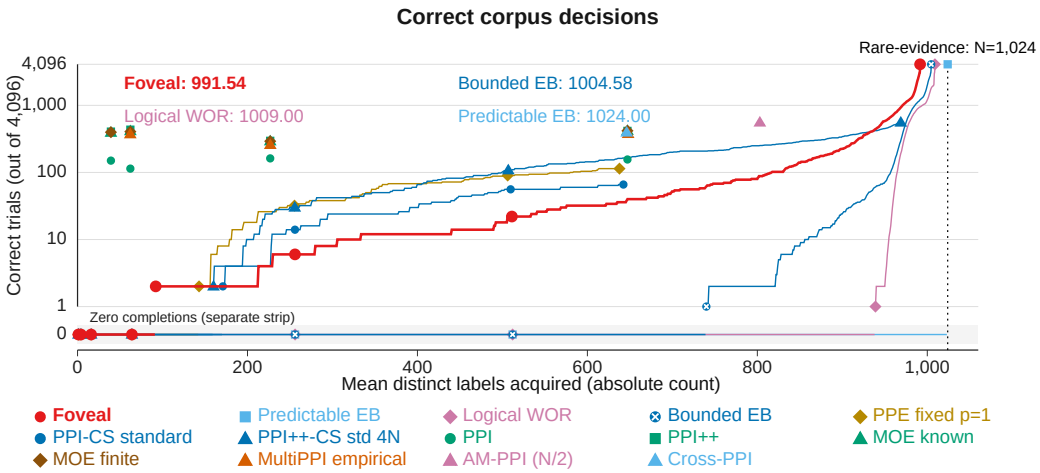


Figure 1: **Correct completion and label cost when decisive evidence is rare.** Each configuration is evaluated over 4096 development trials with $N = 1024$. The exploratory Foveal variant that permits statistical stopping from label 64 completes all trials correctly at 991.54 mean distinct labels, the lowest cost among the evaluated configurations with full observed correct completion on this panel. The minimum prefix and displayed configuration selection are post hoc; Appendix H.3 reports the original policy and all sensitivity variants.

Table 1: Real-data comparisons at declared endpoints. Costs count distinct labels; correct completion is the percentage of all planned trials yielding a correct verdict. The upper and lower blocks contain sequential and fixed-time methods, respectively. Appendix I.5 details budgets and eligibility.

Method	Housing labels	Correct (%)	Hypothyroid labels	Correct (%)
Foveal Search	1076.56	100.0	103.06	100.0
EB-CS (constant)	2553.19	100.0	131.19	100.0
PPI++-CS (pooled)	941.58	94.8	60.19	62.5
PPI++-CS (selected)	119.00	100.0	34.81	100.0
PPE (admissible random forest)	1006.31	100.0	Not applicable	
MOE-PPI (known mean)	3258.63	100.0	92.38	100.0
MultiPPI (Ledoit–Wolf shrinkage)	3268.81	87.5	92.38	37.5
AM-PPI	Time limit exceeded		117.38	18.8

5 Experiments

Setup. We first test decisions driven by rare evidence in synthetic corpora of $N = 1024$ items, with $16 \pm h$ rare items for $h \in \{1, 4\}$. Responses combine rare-item membership and an independently shuffled, balanced binary background with weights $7/8$ and $1/8$; the threshold corresponds to 16 rare items. Paired constructions and repeated query orders produce 4096 trials from 64 independent base populations. We also evaluate Housing ($N = 5160$), with six regression and tree-based proxies (Gu et al., 2026), and Hypothyroid ($N = 145$), with two gradient-boosted proxies (Brawand et al., 2026).[1] Baselines include label-only EB-CS and exact logical bounds (Waudby-Smith and Ramdas, 2020a); fixed-time PPI, PPI++, MOE-PPI, MultiPPI, AM-PPI, and Cross-PPI (Angelopoulos et al., 2023a;c; Gu et al., 2026; Cowen-Breen et al., 2026; Brawand et al., 2026; Zrnic and Candès, 2024b); and sequential PPE and PPI-CS/PPI++-CS (Csillag et al., 2025; Kilian et al., 2025). Correct completion uses all planned trials as the denominator, and mean distinct-label cost includes calibration, errors, and undecided runs; predictor training and computation are excluded. Unless specified otherwise, Foveal uses $\delta = .05$, $R = 1$, and doubling fits from the first label. Appendices I.4, I.3, and M give baseline protocols and accounting details.

[1] Appendices H.1 and I.1 specify dataset construction and real-data protocols.

Table 2: Held-out proxy-weight ablations. Entries are mean label counts over 64 paired query orders per population. All 640 trials return correct verdicts without numerical recovery.

Dataset	Foveal ($R = 1$)	Foveal ($R = 2$)	SCALAR ($R = 1$)	ONS ($R = 1$)	ZERO ($R = 1$)
Housing	626.00	698.19	645.22	620.92	700.34
Hypothyroid	96.23	94.69	100.88	100.47	102.64

Table 3: Mean labels over 16 query orders and upper bounds on $\mathbb{E}[T_0]$ for the Housing dataset ($N = 5160$, $R = 1$). The bounds account for conservative evidence rounding used for certification.

Metric	Default	Capped	Fixed
Mean labels	1076.56	1135.63	1205.38
Cost upper bound	5160.00	2973.89	2901.83

5.1 Completion and Label-Cost Tradeoffs

On the rare-evidence panel, the original Foveal policy correctly completes 4088 of 4096 trials at 985.84 mean labels, with no undecided runs. An exploratory variant that delays statistical stopping until label 64 correctly completes all 4096 trials at 991.54 labels (Figure 1), 1.30% below bounded-response EB-CS and 3.17% below predictable EB-CS; the minimum prefix was selected after inspecting development outcomes (Appendix H.4). On the real-data development populations, the original $R = 1$ policy correctly completes all 16 query orders at each central threshold, using 1076.56 labels on Housing and 103.06 on Hypothyroid, respectively 57.8% and 21.4% fewer than constant-bet EB-CS (Table 1). Some prediction-powered configurations cost less: admissible fixed-proxy PPE on Housing, known-mean MOE-PPI on Hypothyroid, and retrospectively selected PPI++-CS predictors on both datasets. The pooled PPI++-CS configurations have lower correct-completion rates, while standard PPI++-CS leaves 3512 rare-evidence trials undecided at its prespecified $4N$ observation horizon. These native-protocol comparisons show Foveal's observed completion and savings over label-only controls, while its cost ranking depends on the proxy regime.

5.2 Effect of Learning Proxy Weights

On held-out Housing ($N = 5160$) and Hypothyroid ($N = 146$) populations, we compare five configurations using 64 paired query orders per population. Foveal uses $R = 1$ or $R = 2$ in (5); the matched CALAR control learns one signed coefficient on the equal-weight proxy combination, ZERO disables the learned adjustment, and approximate Online Newton Step (ONS) updates weights after every label. Foveal and SCALAR use doubling checkpoints, so the ONS comparison changes both the learner and update schedule. All configurations share proxy inputs and fixed-reference evidence components, with settings fixed before held-out outcomes were accessed. All 640 deployments finish correctly without numerical recovery. As shown in Table 2, Foveal ($R = 1$) reduces mean label cost relative to SCALAR by 3.0% on Housing and 4.6% on Hypothyroid, and relative to ZERO by 10.6% and 6.2%. The prespecified equal-population mean of $(T_{\mathrm{Foveal}} - T_{\mathrm{Scalar}})/N$ is $-.01775$, with a 95% paired-bootstrap interval of $[-.02228, -.01361]$ across query orders within these fixed populations. This matched contrast supports learning separate proxy weights; ONS is cheaper on Housing, and Foveal ($R = 2$) is cheaper on Hypothyroid but more costly than $R = 1$ on Housing. Appendices J.3 and L report separate covariance and proxy-mechanism analyses.

5.3 Certifying Expected Label Savings

On the Housing development population ($N = 5160$, $R = 1$), we compare three policies that share uniform acquisition, checkpoint calibration, and fixed-reference evidence components but differ in their learned-component betting fractions. DEFAULT adapts fractions to pre-query correction bounds; CAPPED limits them to 1.25 times safe fixed reference values; and FIXED uses those values throughout (Appendix F).[2] As shown in Table 3, over 16 query orders, CAPPED and FIXED use 1135.63 and 1205.38 mean labels, while their analytic expected-cost upper bounds are 2973.89 and

[2]Appendix F derives the modified-policy cost bounds.

2901.83. These bounds certify at least 42.4% and 43.8% expected savings relative to census for the corresponding exact-fit policies. DEFAULT has the lowest measured cost at 1076.56 labels, but its available bound equals N. The certificates use full-population moments only for evaluation and assume exact fits and certified evidence arithmetic without numerical recovery. Appendix K.2 reports all betting scales and the recovery allowance; runtime and implementation tradeoffs appear in Appendices K, D, and E.

6 Related Work

Foveal Search connects prediction-assisted estimation, sequential inference, and costly corpus evaluation. Our focus is the label cost of reaching a reliable corpus-level decision when the same labels support both proxy calibration and sequential evidence. Appendix G compares estimators, assumptions, and guarantees in detail.

Prediction-assisted estimation. Control variates and semi-supervised estimators use auxiliary moments to reduce estimation variance (Cheng and Feast, 1980; Chakrabortty and Cai, 2018; Azriel et al., 2021). PPI and PPI++ combine predictions with corrections from reference labels (Angelopoulos et al., 2023a;c). With multiple predictors, MultiPPI optimizes their combination and observation budgets (Cowen-Breen et al., 2026), MOE-PPI learns variance-reducing mixtures (Gu et al., 2026), and AM-PPI jointly considers predictor routing and label acquisition (Brawand et al., 2026). These works motivate using proxy information, but their estimation and allocation analyses do not directly give a stopped label-cost bound for the fixed-corpus checkpoint learner studied here. Foveal Search uses known full-corpus proxy covariance to bound the remaining calibration error, then carries that bound through stopping into completion and expected-cost guarantees.

Sequential and active inference. Finite-population confidence sequences permit repeated inference under sampling without replacement (Waudby-Smith and Ramdas, 2020a; 2024). Prediction-Powered E-Values accommodates evolving predictors with anytime-valid evidence (Csillag et al., 2025), while Bayes-assisted PPI constructs asymptotically valid confidence sequences (Kilian et al., 2025). Active Statistical Inference and Prediction-Powered Active Testing instead adapt which items are labeled (Zrnic and Candès, 2024a; Ashouritaklimi et al., 2026). Foveal Search assumes uniform acquisition to obtain its finite-population calibration bound. Beyond validity at an arbitrary stop, it gives conditions for completing a correct decision within a label budget and bounds expected cost while counting the labels used to learn proxy weights.

LLM evaluation and verification systems. ARES combines learned judges, human annotations, and PPI to estimate retrieval-augmented generation quality (Saad-Falcon et al., 2024). Semantic-query cascades use proxy scores and selective oracle calls to control precision and recall or trade off quality against cost (Liskowski and Schmaus, 2026). AMOS interleaves retrieval and verification for agenda-driven knowledge discovery (Yeh et al., 2026). These systems address different evaluation outputs and acquisition workflows. In contrast, Foveal Search focuses on a fixed corpus-level threshold decision and provides reference-label error control together with conditions for early completion and expected label savings.

7 Conclusion

We introduced Foveal Search for reliable corpus-level threshold decisions using inexpensive proxies and costly reference labels. Each label updates evidence before informing calibration with known full-corpus proxy covariance. Our analysis guarantees finite-sample error control and completion by full verification. It also gives conditions for earlier correct completion and expected label savings, accounting for calibration through data-dependent stopping. Across evaluated regimes, study-specific fixed configurations maintain high correct completion and often save labels over label-only controls; held-out ablations separately show savings from learned proxy weights. Future work will investigate proxy-guided acquisition and broader LLM judgment workloads.

AI USE STATEMENT

Generative AI assisted with derivations, implementation, literature comparison, and writing. The authors remain responsible for correctness, attribution, and the exact scope of reported results. The author-side proof audit is not represented as independent external or proof-assistant verification.

REFERENCES

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

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

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. arXiv:2311.01453, 2023c. URL https://arxiv.org/abs/2311.01453. Accepted for publication in The Annals of Applied Statistics.

Kianoosh Ashouritaklimi, Valentin Kilian, Daolang Huang, Tom Rainforth, and François Caron. Prediction-powered active testing. arXiv:2607.08347v1, 2026. URL https://arxiv.org/abs/2607.08347v1.

David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. arXiv:1612.02391v4, 2021. URL https://arxiv.org/abs/1612.02391v4.

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

Nicholas Peter Brawand, Nima Leclerc, Anhthy Ngo, Matthew Peterson, Sriram Vishwanath, Laith Alhussein, and Ben Wellner. Active multiple-prediction-powered inference. arXiv:2605.08429v1, 2026. URL https://arxiv.org/abs/2605.08429v1.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018. doi: 10.1214/17-AOS1594.

R. C. H. Cheng and G. M. Feast. Control variables with known mean and variance. *Journal of the Operational Research Society*, 31(1):51–56, 1980. doi: 10.1057/jors.1980.6.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W. Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gJZ5rf2bS4.

Daniel Csillag, Claudio José Struchiner, and Guilherme Tegoni Goedert. Prediction-powered E-values. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11493–11514, 2025. URL https://proceedings.mlr.press/v267/csillag25a.html.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

Yanwu Gu, Linglong Kong, and Dong Xia. Prediction-powered inference by mixture of experts. arXiv:2604.27892v1, 2026. URL https://arxiv.org/abs/2604.27892v1.

Elad Hazan. Introduction to online convex optimization, 2023a. URL https://arxiv.org/abs/1909.05207.

Elad Hazan. Introduction to online convex optimization. Second edition, arXiv:1909.05207v3, 2023b. URL https://arxiv.org/abs/1909.05207v3. Chapter 4, Lemma 4.2 and Theorems 4.4–4.5.

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

Valentin Kilian, Stefano Cortinovis, and Francois Caron. Anytime-valid, bayes-assisted, prediction-powered inference. In *Advances in Neural Information Processing Systems*, volume 38, pages 900–918, 2025. doi: 10.52202/085713-0033.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Paweł Liskowski and Kyle Schmaus. Streaming model cascades for semantic SQL. arXiv preprint 2604.00660, 2026. URL https://arxiv.org/abs/2604.00660v2. Version 2, July 2026.

Liana Patel, Siddharth Jha, Melissa Pan, Harshit Gupta, Parth Asawa, Carlos Guestrin, and Matei Zaharia. Semantic operators and their optimization: Enabling LLM-based data processing with accuracy guarantees in LOTUS. *Proceedings of the VLDB Endowment*, 18(11):4171–4184, 2025. doi: 10.14778/3749646.3749685.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.20.

R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974. doi: 10.1214/aos/1176342611.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, pages 20204–20214, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/e96c7de8f6390b1e6c71556e4e0a4959-Abstract.html.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, pages 20204–20214, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/e96c7de8f6390b1e6c71556e4e0a4959-Abstract.html.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

C.-Y. Yeh, W.-C. Huang, H.-W. Chen, C. Wang, D.-N. Yang, W.-C. Lee, P. S. Yu, and M.-S. Chen. AMOS: Active mining with observable steps for agenda-driven review knowledge discovery. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2026.

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

Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024b. doi: 10.1073/pnas.2322083121. URL https://www.pnas.org/doi/10.1073/pnas.2322083121.