A QUERY IS NOT A COMMITMENT: LEARNING TO CORRECT EXPERT ANSWERS IN ONLINE DEFERRAL

Anonymous authors
Paper under double-blind review

ABSTRACT

An inaccurate expert can still provide useful information after correction. We study online learning to defer in which the learner chooses an expert and fixes a correction function before purchasing its answer, then applies that function to the answer received. The difficulty is that observed losses reflect both expert quality and an unfinished correction: early errors can discourage queries that would be valuable after learning. We propose RESPONSE-AWARE UCB, which pools shared and expert-specific polynomial responses. A bound on cumulative response-learning error calibrates confidence-weighted risk regression and exploration, allowing the router to account for this error when deciding which answers to buy. Under bounded residuals and disagreements, a fixed feasible model of optimal responses, and linear models of free and optimal queried risk, the calibrated algorithm achieves high-probability pseudo-regret $O(\sqrt{T}\log(T+1))$ over T rounds for fixed problem parameters. The guarantee permits singular answer distributions and misspecified shared responses; optimality is relative to the bounded response class. On four test streams, the selected cubic policy has lower fee-inclusive cost than seven baselines that deploy answers unchanged. Comparisons with a common correction learner examine routing, while six-price comparisons measure cost and query rates.

1 INTRODUCTION

Learning to defer (L2D) asks a router to choose between an internal prediction and an external expert's answer (Madras et al., 2018; Mozannar & Sontag, 2020; Montreuil et al., 2026a). With several experts and consultation fees, the router must decide whom to ask and whether the expected reduction in prediction error is worth the price (Mao et al., 2024c). Standard direct deferral deploys a purchased answer unchanged. This ties the value of consultation to the accuracy of the raw answer, even when its errors are predictable. Suppose a weather model tracks day-to-day temperature changes but consistently forecasts too high. Its uncorrected error and fee may make a query unattractive, while removing the stable bias would make the same answer useful. Even a router that knows the data law cannot realize this gain if it must deploy the answer unchanged. The question is therefore not only *whom to ask*, but also *how to use the answer* once it arrives.

We study this distinction through online L2D with *post-answer responses*. On round t, the learner sees its internal prediction h_t, the public context, the available experts, and their fees. It either uses h_t or chooses an expert j and fixes a correction function $\Delta_{t,j}$. If it queries j, the final prediction is

$$\widehat{Y}_t = M_{t,j} + \Delta_{t,j}(M_{t,j} - h_t). \tag{1}$$

The function is fixed before the query and evaluated after the answer arrives. Only the selected expert's answer is observed. The outcome is revealed after the final prediction on every round, providing feedback for future corrections and query decisions. Thus the final prediction may use the purchased answer, but the decision to buy that answer is made before it is seen. We represent this sequential decision in Appendix A.

The router must decide whether to buy an expert's answer while it is still learning how to correct that answer. In direct deferral, a purchase reveals the loss of the queried action. With a learned response, it reveals the loss under the learner's current correction, while the router must assess the risk attainable with the best feasible response. The conditional gap is nonnegative and can change with the response; a confidence bound for centered observation noise alone does not cover it. Early

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.

correction errors can therefore make a useful expert appear too costly. If queries stop, its retained response loses direct training observations. We propose RESPONSE-AWARE UCB to handle this moving target. It pools bounded polynomial responses that can use the purchased answer nonlinearly: a shared fit offers an initial proposal for a new expert, while an identity-specific fit preserves what was learned when that expert returns. The shared fit need not describe every expert correctly. We bound the cumulative response-learning error relative to feasible identity responses, then use that bound to calibrate confidence-weighted risk regression. This limits the influence of observations acquired at high queried-risk uncertainty; a confidence bonus favors queries when an expert's corrected risk could still be low.

Our contributions are as follows. (i) We define a conditional-law oracle that, like the learner, selects an expert and fixes its response before seeing the answer. This comparator measures the cost of learning both which answers to buy and how to use them. (ii) Under fixed feasible models of optimal response coefficients and linear models of free and optimal queried risk, calibrated RESPONSE-AWARE UCB has high-probability pseudo-regret $O(\sqrt{T}\log(T+1))$ for fixed problem parameters. The guarantee allows singular answer distributions and misspecified shared responses. A robustness extension quantifies the cost of approximate response and risk models when bounds on their approximation errors are supplied for calibration. (iii) On four test streams, the selected cubic policy has lower cost than seven external baselines that deploy answers unchanged. Comparisons with a common correction learner examine routing, while six-price comparisons measure cost and query rates.

2 RELATED WORK

Learning to defer. L2D learns when to predict and when to rely on an expert (Madras et al., 2018; Mozannar & Sontag, 2020). In regression, routing compares the prediction losses and consultation costs of several experts (Mao et al., 2024c). In classification, Montreuil et al. (2026a) separate classifier and expert scores to address coupling in joint prediction–deferral training. Online L2D extends routing to varying experts and selected-action feedback, primarily in classification settings (Duy et al., 2026), while time-series deferral also models changes in expert performance (Montreuil et al., 2026f). Our setting differs in that querying an expert does not end the decision process. After purchasing an answer, the learner must also learn how to transform that answer into the final prediction. The response function is fixed before acquisition, but its output depends on the answer received. Errors in learning this response therefore affect both the quality of the current prediction and which experts are queried in the future.

Online bandits. Contextual bandits learn action values from selected-action feedback using linear confidence bounds (Li et al., 2010; Abbasi-Yadkori et al., 2011), parameter sampling (Agrawal & Goyal, 2013), or neural confidence models (Zhou et al., 2020). Our comparisons cover these alternatives through SleepingLinUCB, SharedLinUCB, LinTS, and NeuralUCB. SleepingLinUCB maintains a separate model for each expert and restricts selection to available experts, following the sleeping-bandit setting (Kleinberg et al., 2010), while SharedLinUCB uses a shared context–identity model. To account for changing performance, we also compare with discounted regression (D-LinUCB; Russac et al. 2019) and detector-triggered restarts (CUSUM-LinUCB and GLR-LinUCB), adapted from the non-contextual methods of Liu et al. (2018) and Besson et al. (2022). All seven baselines use our availability and feedback protocol. Our confidence weighting follows He et al. (2022). The additional analytical step is to bound the bias generated by learning the responses and use this bound to calibrate acquisition. Appendix G develops these connections.

3 BACKGROUND

Learning to defer. In learning to defer (L2D), a policy decides whether to use an internal prediction or delegate to an expert (Madras et al., 2018; Mozannar & Sontag, 2020). The expert may be a human or another model. For regression, let $Y \in \mathbb{R}$ be the outcome and $M = (M_1, \ldots, M_K) \in \mathbb{R}^K$ the potential answers from $K \geq 1$ expert identities. The public input $x \in \mathcal{X}$ determines the internal prediction $h(x)$, available identities $\mathcal{E}(x) \subseteq \{1, \ldots, K\}$ and consultation fees $\beta_j(x) \geq 0$. An admissible policy, or *router*, is a measurable map $\pi : \mathcal{X} \to \{0, \ldots, K\}$ with $\pi(x) \in \{0\} \cup \mathcal{E}(x)$. Action 0 uses the internal prediction $h(x)$; action j deploys expert answer M_j unchanged. We call

squared prediction error the *loss* and loss plus the consultation fee the *cost*:

$$C^{\mathrm{def}}(x, Y, M; i) = \begin{cases} (Y - h(x))^2, & i = 0, \\ (Y - M_j)^2 + \beta_j(x), & i = j \in \mathcal{E}(x). \end{cases} \tag{2}$$

L2D minimizes $\mathbb{E}[C^{\mathrm{def}}(X, Y, M; \pi(X))]$ over a fixed nonempty class Π of admissible routers, where X is the random public input (Mao et al., 2024c;b). Querying an expert improves expected cost over the internal prediction when its conditional loss reduction exceeds its fee.

Online decisions and feedback. Online L2D learns from sequential feedback (Duy et al., 2026). On round t of $T \geq 1$ rounds, the pre-query information $\mathcal{F}_t$ contains the completed history and current public input x_t; mathematically, it is a sigma-field. A history-dependent router chooses $I_t = \pi_t(x_t) \in \{0\} \cup \mathcal{E}(x_t)$: $I_t = 0$ uses $h(x_t)$, while $I_t = j$ selects expert j. These choices are called *arms* in bandit terminology. Availability may change while expert identities persist. The internal predictor may also learn, with its current prediction included in x_t. Write $M_t = (M_{t,1}, \ldots, M_{t,K})$ for the potential-answer vector. Public quantities such as x_t may be random through the history; they are known when conditioning on $\mathcal{F}_t$.

Feedback determines which losses can be used to update the router. Full information reveals every available action's loss; bandit feedback reveals only the chosen action's loss (Auer et al., 2002). If outcomes are revealed on every round but expert answers only upon purchase, the internal loss is observed even when an expert is selected. The losses of unqueried experts are generally unavailable.

For Π fixed before interaction, realized direct-deferral regret compares cumulative cost with the best router in hindsight (Cesa-Bianchi & Lugosi, 2006):

$$\mathrm{Reg}_T^{\mathrm{def}}(\Pi) = \sum_{t=1}^{T} C^{\mathrm{def}}(x_t, Y_t, M_t; I_t) - \inf_{\pi \in \Pi} \sum_{t=1}^{T} C^{\mathrm{def}}(x_t, Y_t, M_t; \pi(x_t)). \tag{3}$$

The infimum is outside the sum: one router is used throughout, but it may select different experts at different inputs. Both costs use the same realized sequence, including unobserved potential answers. A *conditional risk* is expected prediction loss given $\mathcal{F}_t$. Our *conditional pseudo-regret* in section 4 compares expected costs with an oracle's best feasible action and response on each round, chosen before any current answer is revealed.

4 Protocol and Comparator

We now specify the allowed corrections and the oracle against which we measure the cost of learning to query and use an answer.

Protocol. Write $h_t = h(x_t)$, $\mathcal{E}_t = \mathcal{E}(x_t)$ and $\beta_{t,j} = \beta_j(x_t)$. Before querying, Y_t and the potential answers $(M_{t,j})_{j \in \mathcal{E}_t}$ are hidden. We assume non-intervention: purchasing an answer does not change their conditional joint law, as when a forecast does not affect the weather. Using only $\mathcal{F}_t$, the learner fixes I_t and a response function. Action 0 deploys h_t; a query to j pays $\beta_{t,j}$, reveals only $M_{t,j}$ and applies the correction in (1). The outcome is then revealed on every round. Our policy is deterministic, with public orders for ties and registration.

Disagreement. Assume $|Y_t - h_t| \leq R$ and $|M_{t,j} - h_t| \leq D$ almost surely for every round and active expert, with known $R, D > 0$. Set $G = R + D$, $d_{t,j} = M_{t,j} - h_t$ and $e_{t,j} = Y_t - M_{t,j}$; then $|e_{t,j}| \leq G$. We learn the response by predicting the expert residual $e_{t,j}$ from disagreement $d_{t,j}$. Since $M_{t,j} = h_t + d_{t,j}$ and h_t is public, centering loses no information about the answer given $\mathcal{F}_t$. It gives a bounded response input without requiring bounded absolute prediction levels.

Response class. We model the correction as a low-degree polynomial in disagreement. Polynomial features allow nonlinear adjustments while keeping the model linear in its coefficients (Hu et al., 2020; Liu et al., 2021), so fitting the response by squared loss remains a convex problem. Keeping the degree low also limits how many coefficients must be learned from each expert's purchased answers. Because purchases depend on earlier fits, we use an online least-squares comparison that does not require independent samples (Hazan, 2023). It bounds cumulative fitting loss relative to a fixed feasible model, allowing us to account for response-learning error in the routing analysis (lemma D.1).

To keep the corrections bounded, we use scaled Legendre features and restrict the coefficient norm. For an integer order $q \geq 1$, let $\mathbb{B}_q = \{a \in \mathbb{R}^{q+1} : \|a\|_2 \leq 1\}$, with $\|\cdot\|_2$ the Euclidean norm. Using

the degree-k Legendre polynomial P_k, define

$$\phi_q(d) = \frac{(1, DP_1(d/D), \ldots, DP_q(d/D))^\top}{\sqrt{1+qD^2}}, \qquad \Delta_a(d) = Ga^\top \phi_q(d). \tag{4}$$

For a query to expert j, the learner fixes $a_{t,j} \in \mathbb{B}_q$ before purchase. Its deployed prediction is

$$\widehat{Y}_t = \begin{cases} h_t, & \text{if } I_t = 0, \\ M_{t,j} + \Delta_{a_{t,j}}(d_{t,j}), & \text{if } I_t = j \in \mathcal{E}_t. \end{cases} \tag{5}$$

For $d \in [-D, D]$, the bound $|P_k(d/D)| \le 1$ (Szegő, 1975) gives $\|\phi_q(d)\|_2 \le 1$ and $|\Delta_a(d)| \le G$ for $a \in \mathbb{B}_q$. The factor D preserves the linear coordinate $DP_1(d/D) = d$. Since $|e_{t,j}| \le G$, the queried prediction loss $(e_{t,j} - \Delta_a(d_{t,j}))^2$ is at most $4G^2$. This bound supplies bounded labels for risk estimation. The scaling and coefficient ball define the response budget. The normalization depends on q, so the bounded classes need not be nested. The analysis requires bounded features and linear coefficients, not orthogonality under the answer distribution, and extends to other fixed bounded dictionaries under their corresponding response and risk models.

For the experiments, we choose $q = 3$: four coefficients at a fixed context capture bias correction, linear rescaling, and quadratic and cubic adjustments. The cubic term allows the rescaling to vary with disagreement magnitude.

Oracle and regret. The comparator class consists of all $\mathcal{F}_t$-measurable rules choosing free prediction or a pair (j, a) with $j \in \mathcal{E}_t$ and $a \in \mathbb{B}_q$. Routing may use the full public history, but the correction must belong to the specified response class. Write $L_t^0 = \mathbb{E}[(Y_t - h_t)^2 \mid \mathcal{F}_t]$ for the free risk. The best queried prediction risk for expert j is

$$W_{t,j}^* = \min_{a \in \mathbb{B}_q} \mathbb{E}[(e_{t,j} - Ga^\top \phi_q(d_{t,j}))^2 \mid \mathcal{F}_t]. \tag{6}$$

The minimization fixes one coefficient vector using $\mathcal{F}_t$; its loss is averaged over the still-hidden outcome and answer. Compactness and continuity give a minimizer; Appendix A specifies measurable versions. Since $L_t^0 \le R^2$, a query fee of at least R^2 cannot improve on free prediction. Restricting queries to $\mathcal{A}_t = \{j \in \mathcal{E}_t : \beta_{t,j} < R^2\}$ therefore preserves the oracle cost $C_t^* = \min\{L_t^0, \min_{j \in \mathcal{A}_t}(W_{t,j}^* + \beta_{t,j})\}$, with an empty inner minimum equal to $+\infty$. The oracle knows the conditional law and may change expert and response with context, but must choose both before any current answer is revealed. Set $\beta_{t,0} = 0$ and $C_t = (Y_t - \widehat{Y}_t)^2 + \beta_{t,I_t}$. Our pseudo-regret is

$$\mathrm{Reg}_T = \sum_{t=1}^{T} \{\mathbb{E}[C_t \mid \mathcal{F}_t] - C_t^*\}. \tag{7}$$

Unlike hindsight regret over the fixed direct-deferral class Π in section 3, this compares conditional expected costs with the best feasible action and response on each round. The oracle chooses before feedback, following the conditional-mean convention of linear contextual bandits (Abbasi-Yadkori et al., 2011; He et al., 2022).

5 Response-Aware UCB

Response-Aware UCB uses past purchases to learn both how to correct an expert's answer and whether that answer is worth buying. The difficulty is that an expert's early losses reflect an unfinished correction, not just the risk attainable with its best response. We first fit the corrections, then bound the error incurred while learning them, and use that bound to construct the query rule.

5.1 Learning the response

Each expert has an identity-specific fit that retains what was learned from its purchased answers. A second, shared fit uses purchases from all identities and can help with a new expert when their responses have common structure. We combine these two proposals according to their past losses on the selected identity, without requiring the shared model to be correct.

Response fits. Let $\ell_{t,j} \in \mathbb{R}^{d_\ell}$ and $s_{t,j} \in \mathbb{R}^{d_s}$ be public features of norm at most one for the identity and shared fits, respectively. They determine the correction coefficients before purchase; $\phi_q(d_{t,j})$

determines how those coefficients act on the answer after purchase. For example, an intercept and a bounded temperature covariate let a coefficient vary with the public temperature. With regularizers $\lambda_I, \lambda_P > 0$, we fit the normalized residuals by

$$\Theta_{t,j} = \operatorname*{arg\,min}_{\|\Theta\|_F \le 1} \left\{ \lambda_I \|\Theta\|_F^2 + \sum_{r<t:I_r=j} \left(\frac{e_{r,j}}{G} - \ell_{r,j}^\top \Theta \phi_q(d_{r,j}) \right)^2 \right\}, \tag{8}$$

$$\Psi_t = \operatorname*{arg\,min}_{\Psi \in \mathbb{R}^{d_s \times (q+1)}} \left\{ \lambda_P \|\Psi\|_F^2 + \sum_{r<t:I_r \neq 0} \left(\frac{e_{r,I_r}}{G} - s_{r,I_r}^\top \Psi \phi_q(d_{r,I_r}) \right)^2 \right\}. \tag{9}$$

Here $\Theta \in \mathbb{R}^{d_\ell \times (q+1)}$ and $\|\Theta\|_F$ is the Euclidean norm of its entries. The regularizers make both minimizers unique. The identity proposal $a_{t,j}^H = \Theta_{t,j}^\top \ell_{t,j}$ is feasible because $\|a_{t,j}^H\|_2 \le \|\Theta_{t,j}\|_F \|\ell_{t,j}\|_2 \le 1$. We project the shared proposal onto the same coefficient ball:

$$a_{t,j}^P = \frac{\Psi_t^\top s_{t,j}}{\max\{1, \|\Psi_t^\top s_{t,j}\|_2\}}. \tag{10}$$

Superscripts H and P denote identity history and population. The identity fit is constrained during optimization so that its cumulative loss can be compared with any fixed feasible identity model (lemma D.1 in Appendix D). The shared fit only needs to supply a feasible alternative, so projection suffices.

Pooling. To adapt to whether an expert benefits from shared information, we deploy a convex mixture of the two coefficient proposals:

$$a_{t,j} = w_{t,j}^H a_{t,j}^H + w_{t,j}^P a_{t,j}^P, \qquad w_{t,j}^H, w_{t,j}^P \ge 0, \quad w_{t,j}^H + w_{t,j}^P = 1. \tag{11}$$

We initialize both weights at $1/2$ for each new identity. After querying j and observing the outcome, we evaluate the pre-query proposals using $\mathcal{L}_t^u = (e_{t,j}/G - (a_{t,j}^u)^\top \phi_q(d_{t,j}))^2/4$, $u \in \{H, P\}$. We update the weights by exponential weighting (Cesa-Bianchi & Lugosi, 2006), multiplying each by $\exp(-\eta \mathcal{L}_t^u)$ and renormalizing, with $0 < \eta \le 1/2$. Other identities keep their weights. Both losses use the same purchased answer, so comparing proposals needs no additional query.

The mixture stays in $\mathbb{B}_q$ by convexity. For the stated learning-rate range, its cumulative squared loss over queries to each identity is at most that of either proposal sequence plus $4G^2 \log 2/\eta$. Lemma D.2 in Appendix D proves this comparison using exp-concavity (Hazan, 2023). Thus the shared proposal can help without invalidating the guarantee provided by the identity fit. All fits use only past observations, and identity histories and pool weights persist during absence.

5.2 Accounting for response error

An expert may be worth buying after correction even when its early corrected predictions are poor. To account for this distinction, we separate the loss we observe from the optimal risk we want to estimate. On a queried round, the normalized deployed loss is

$$Z_t = (Y_t - \widehat{Y}_t)^2/(4G^2) \in [0, 1]. \tag{12}$$

We define the response-learning gap by $\rho_t = (\mathbb{E}[(Y_t - \widehat{Y}_t)^2 \mid \mathcal{F}_t] - W_{t,I_t}^*)/G^2$ when $I_t \neq 0$, and set $\rho_t = 0$ otherwise. Since the deployed response belongs to the oracle's coefficient ball, queried rounds satisfy

$$\mathbb{E}[Z_t \mid \mathcal{F}_t] = W_{t,I_t}^*/(4G^2) + \rho_t/4, \qquad 0 \le \rho_t \le 4. \tag{13}$$

The first term is the router's target; the second is the error from learning the response. We cannot observe ρ_t because the optimal risk is unknown. Instead of estimating each gap, we bound their cumulative size. This requires a fixed feasible identity model to represent optimal responses across contexts, as formalized in assumption 6.1.

Lemma 5.1 (Cumulative response error)**.** *Under the bounded protocol of section 4 and assumption 6.1, use the exact response fits and pooling updates of* Response-Aware UCB *with* $0 < \eta \le 1/2$*. For every* $\delta \in (0, 1)$*, with probability at least* $1 - \delta/3$ *over the interaction, simultaneously for all* t*,*

$$\sum_{s \le t} \rho_s \le E_t, \tag{14}$$

where E_t is the query-count bound in equation (24). For $t \le T$, it satisfies $E_t \le \overline{E}_T$, with the explicit horizon bound $\overline{E}_T$ given in equation (19).

Appendix D.3 proves the lemma by combining the constrained-fitting and pooling comparisons, then converting realized loss differences to conditional risk differences. On this event, $G^2 E_t$ bounds the cumulative conditional excess prediction risk caused by learning the responses. It grows only logarithmically with the number of purchases when the dimensions, registry capacity, regularizers, pool rate and δ are fixed. The calibrated router uses the deterministic horizon bound $\overline{E}_T$ defined in section 6, not the unknown individual gaps.

5.3 Choosing an action

We now estimate whether an expert's optimal queried risk plus its fee can beat free prediction. The response-error bound lets us account for the bias in purchased losses when making this comparison.

Risk estimates. We use public features $x_t^0 \in \mathbb{R}^{d_0}$ and $c_{t,j} \in \mathbb{R}^{d_c}$ of norm at most one. Unlike $\ell_{t,j}, s_{t,j}$, which determine correction coefficients, these features predict free and optimal queried risk, respectively. All four feature vectors are known under $\mathcal{F}_t$ and cannot use the current answer; they may use the same public input. The guarantee assumes that both normalized risks are linear in these features (assumption 6.2); this is separate from the response-model assumption above. The free loss is observed every round, whereas Z_t is available only after a query. With positive-definite priors Λ_0, Λ_* and $c_s = c_{s,I_s}$ on queried rounds, we use the full-history ridge estimates below, where ω_s is a sample weight defined next:

$$V_t^0 = \Lambda_0 + \sum_{s<t} x_s^0 (x_s^0)^\top, \qquad \widehat{\theta}_t^0 = (V_t^0)^{-1} \sum_{s<t} x_s^0 \frac{(Y_s - h_s)^2}{R^2},$$
$$V_t^* = \Lambda_* + \sum_{s<t:I_s\neq 0} \omega_s c_s c_s^\top, \qquad \widehat{\theta}_t^* = (V_t^*)^{-1} \sum_{s<t:I_s\neq 0} \omega_s c_s Z_s. \tag{15}$$

The design matrices V_t^0, V_t^* record the observed feature directions; the priors regularize directions with little data.

Confidence weighting. To turn the cumulative error bound into a useful confidence bound, we also control each observation's influence. A biased observation in a poorly measured direction can strongly affect later risk estimates. As in linear UCB (Abbasi-Yadkori et al., 2011), we measure uncertainty in the queried-risk direction by $u_{t,j} = \sqrt{c_{t,j}^\top (V_t^*)^{-1} c_{t,j}}$. For a threshold $\tau > 0$ fixed before interaction, we assign each queried round a weight before observing the selected answer:

$$\omega_t = \min\{1, \tau / u_{t,I_t}\}, \tag{16}$$

using one at zero width. Whereas $w_{t,j}^H, w_{t,j}^P$ mix response coefficients, ω_t weights one queried-risk observation. This confidence-weighting rule comes from corrupted linear bandits (He et al., 2022). It caps the product $\omega_t u_{t,I_t}$ at τ, so the total response error in lemma 5.1 controls its influence on the risk estimate. Equation (46) gives the bound; Appendix D.4 proves the resulting confidence guarantee. Both the design and target sum use the same sample weight. Response fits and proposal-loss updates remain unweighted: they learn the correction directly, rather than treating its current loss as optimal risk.

Acquisition. We express the forecasts in squared-loss units so they can be compared with fees: $\widehat{L}_t^0 = R^2 \operatorname{clip}_{[0,1]}((x_t^0)^\top \widehat{\theta}_t^0)$ and $\widetilde{W}_{t,j} = 4G^2 c_{t,j}^\top \widehat{\theta}_t^*$. Clipping truncates to the indicated interval and cannot increase the free forecast's error because $L_t^0 \in [0, R^2]$. The queried forecast stays unclipped, as used in its confidence bound.

For a confidence multiplier $\alpha \ge 0$, we use the acquisition scores

$$S_{t,j} = \widehat{L}_t^0 - \widetilde{W}_{t,j} - \beta_{t,j} + 4\alpha G^2 u_{t,j}, \qquad S_{t,0} = 0, \qquad j \in \mathcal{A}_t. \tag{17}$$

The first three terms estimate the saving from consultation, including its fee. The bonus favors queries whose optimal risk could be lower than the current forecast. Under the assumptions and calibration in theorem 6.3, $\widetilde{W}_{t,j} - 4\alpha G^2 u_{t,j}$ is a lower confidence bound for $W_{t,j}^*$, accounting for both label noise and response-learning bias. No exploration bonus is needed for free prediction because its loss is observed under every action; its estimation error is still included in the regret bound.

We choose the largest score, preferring free prediction on an exact tie and then the earliest registered expert. Thus we query only when the selected expert's score is positive. We deploy its frozen pooled response as in equation (5), or use h_t if no query is selected. After observing Y_t, we update the free-risk fit on every round and the queried-risk fit, response fits and selected identity's pool weights only after a purchase.

The theorem couples α and τ through equation (21); the benchmarks retain the same updates but select these constants independently by cost on the validation block. Appendix B gives the benchmark feature construction, Algorithm 1, and the recursive implementation.

6 Theory

Can we learn which answers are worth purchasing while still learning how to correct them? We answer this question by bounding regret against the conditional response oracle in equation (7). The guarantee includes both query mistakes and the prediction error incurred while learning the responses, rather than assuming that the corrections are already known.

6.1 Model assumptions

We use one model to learn optimal corrections across contexts and another to estimate their risk before purchasing an answer. Both assumptions hold almost surely for every round and eligible expert; public contexts and availability may depend on history.

Assumption 6.1 (A feasible model of optimal responses)**.** *For every identity j, there is a fixed $\Theta_j^* \in \mathbb{R}^{d_\ell \times (q+1)}$ with $\|\Theta_j^*\|_F \le 1$ such that $(\Theta_j^*)^\top \ell_{t,j}$ minimizes the conditional queried risk over the entire ball $\mathbb{B}_q$ in equation (6).*

We model the *optimal coefficients* across contexts with one fixed matrix per identity. This connects the fixed-comparator guarantee of online regression (Hazan, 2023) to the conditional response oracle. The conditional mean residual itself need not be polynomial: the model represents a best correction within the prescribed ball. The guarantee also permits a misspecified shared model and singular answer covariance. Appendix B.5 relates this condition to realizability in contextual bandits.

Assumption 6.2 (Linear free and optimal queried risks)**.** *Fixed parameters $\theta_0^* \in \mathbb{R}^{d_0}$ and $\theta_* \in \mathbb{R}^{d_c}$ satisfy*

$$L_t^0/R^2 = (x_t^0)^\top \theta_0^*, \qquad W_{t,j}^*/(4G^2) = c_{t,j}^\top \theta_*. \tag{18}$$

Known finite constants Q_0, Q_* bound $\|\theta_0^*\|_{\Lambda_0} \le Q_0$ and $\|\theta_*\|_{\Lambda_*} \le Q_*$, where $\|v\|_\Lambda = \sqrt{v^\top \Lambda v}$.

Fixed linear conditional means, bounded features and parameter-norm bounds are used in OFUL (Abbasi-Yadkori et al., 2011), linear Thompson sampling (Agrawal & Goyal, 2013) and confidence-weighted linear bandits (He et al., 2022). We apply this structure to free risk and risk *after optimizing the response*; the latter is modeled separately from the optimal coefficients. Both models allow context dependence through their public features. For example, take $h = 0$, Y uniform on $\{-0.3, 0.3\}$, and an expert answering $M = Y + 0.2 + 0.1x$ for public $x \in [-1, 1]$. With $q = 3$, $R = 0.3$, $D = 0.7$, and $\ell = (1, x)^\top/\sqrt{2}$, its optimal correction $-0.2 - 0.1x$ has one fixed coefficient matrix with squared Frobenius norm $0.247 < 1$. The free and optimal queried risks are 0.09 and zero. Appendix B.6 gives the matrix and risk features satisfying both assumptions.

6.2 Confidence calibration

We choose the exploration multiplier α and weight threshold τ to cover two sources of estimation error: centered observation noise and the upward bias from unfinished responses in equation (13). For a horizon $T \ge 1$ and failure probability $\delta \in (0, 1)$, lemma 5.1 gives $\sum_{t \le T} \rho_t \le \overline{E}_T$ with probability at least $1 - \delta/3$, using the following bound known before interaction:

$$\overline{E}_T = 2\left[K\lambda_I + 8Kp_I \log\left(1 + \frac{T}{Kp_I\lambda_I}\right) + \frac{4K\log 2}{\eta}\right] + 64\log\frac{3}{\delta}. \tag{19}$$

Here $p_I = d_\ell(q+1)$ counts coefficients per identity. The terms account for fitting the identity responses, pooling the two proposals, and converting observed loss comparisons to conditional risk comparisons (Appendix B.1).

For centered noise and regularization, self-normalized regression bounds (Abbasi-Yadkori et al., 2011) give confidence radii B_0 for free risk and B_* for queried risk. For $v \in \{0, *\}$, set $\lambda_v = \lambda_{\min}(\Lambda_v) > 0$ and $d_* = d_c$, and take

$$B_v = \tfrac{1}{2}\sqrt{d_v \log\left(1 + \frac{T}{d_v \lambda_v}\right) + 2\log(3/\delta)} + Q_v. \tag{20}$$

The square-root term covers noise; Q_v covers regularization bias. For queried risk, the weights cap $\omega_t u_{t,I_t}$ at τ (He et al., 2022). The response-learning contribution to the radius is therefore at most $\tau\overline{E}_T/4$ (equation (46)). We balance this bound with B_*:

$$\tau = \frac{4B_*}{\overline{E}_T}, \qquad \alpha = B_* + \frac{\tau\overline{E}_T}{4} = 2B_*. \tag{21}$$

On the joint confidence event, $\widetilde{W}_{t,j} - 4\alpha G^2 u_{t,j}$ is then a lower confidence bound for optimal queried risk, accounting for both sources of error. Appendix D.4 proves the simultaneous confidence statement.

6.3 Regret guarantee

Theorem 6.3 (Regret of response-aware routing)**.** *Assume the bounded non-interventional protocol of section 4 and assumptions 6.1 and 6.2. Fix $T \geq 1$, $\delta \in (0, 1)$, feature maps and positive-definite priors before interaction; public feature vectors may depend on history. Run* Response-Aware UCB *with exact constrained response fits, full-history weighted queried-risk estimates, $0 < \eta \leq 1/2$ and equation (21). With probability at least $1 - \delta$ over the joint interaction law,*

$$\mathrm{Reg}_T \leq C_{\mathrm{prob}} \left\{ \sqrt{T} \log \frac{T+1}{\delta} + \log \frac{T+1}{\delta} \right\}, \tag{22}$$

where C_{prob} is finite and depends only on the residual bounds, dimensions, registry capacity, regularizers, priors, parameter-norm bounds and pool rate, not on T or δ.

For fixed problem parameters and $\delta = (T+1)^{-2}$, the bound is $O(\sqrt{T}\log(T+1))$. Bounded one-round regret also gives vanishing expected average regret. Thus the cost of learning the corrections and query decisions is sublinear relative to the oracle, even though only purchased answers train the responses. Appendix B.7 gives a fully explicit bound with sharper dependence on query counts.

Proof idea. Lemma 5.1 controls the total error from learning responses. The calibrated weights prevent this error from invalidating the risk confidence bounds. Optimistic acquisition then compares each chosen action with the oracle; summing the estimation and response errors gives the regret bound. Appendix D gives the proof.

Comparison with direct deferral. Along the same public history, the learner's cumulative conditional cost minus that of the direct-deferral oracle equals Reg_T minus the cumulative oracle gain from correction. On the theorem's event, the learner therefore improves on direct deferral whenever this gain exceeds the regret bound. If the average oracle gain stays bounded away from zero, the linear cumulative benefit eventually exceeds the sublinear learning cost. Appendix C.2 characterizes when correction produces such a gain. For approximate response and risk models, Appendix D.6 bounds the additional regret using supplied approximation-error bounds.

7 Experiments

We examine (i) the cost of learning to query and correct answers against seven external baselines on four streams (table 1); (ii) sensitivity to six consultation prices (table 2); and (iii) routing with a common response learner, together with learning and query-rate curves, in Appendix E.

Setting. Synthetic regimes control when free prediction, direct deferral, or correction is useful. Melbourne daily minimum temperatures (Hyndman & Yang, 2018) pair real outcomes with simulated sensors to control answer bias and noise. On Jena (Max Planck Institute for Biogeochemistry, n.d.) and Delhi (sumanthvrao, 2019), experts are frozen random forests, extra trees, and gradient-boosted trees; Jena also uses radial-basis kernel ridge. Jena uses contemporaneous weather covariates;

Melbourne and Delhi use past observations and calendar features. Availability schedules include expert absence and return.

We compare SleepingLinUCB and SharedLinUCB, our available-action LinUCB adaptations (Li et al., 2010; Kleinberg et al., 2010); D-LinUCB (Russac et al., 2019); LinTS (Agrawal & Goyal, 2013); NeuralUCB (Zhou et al., 2020); and CUSUM-LinUCB and GLR-LinUCB, linear adaptations of the change detectors of Liu et al. (2018) and Besson et al. (2022). For each method and dataset, we select from 256 configurations by validation cost and report fee-inclusive cost on separate test blocks. Appendices E and F describe evaluation, warmup, features, and experts.

Table 1: Mean test cost, including consultation fees. Entries show mean ± sample standard deviation across five learner seeds. Lower is better.

Method	Synthetic	Melbourne	Jena	Delhi
RESPONSE-AWARE UCB	**0.08092**	**4.49069**	**3.72844**	**2.59240**
SleepingLinUCB	0.08488	4.61400	4.03045	2.66401
SharedLinUCB	0.08316	4.90137	4.02217	2.64185
D-LinUCB	0.08790	5.11775	4.07844	2.63729
LinTS	0.08609 ± 0.00389	5.42987 ± 0.15762	4.11937 ± 0.04932	2.74807 ± 0.04935
NeuralUCB	0.10183 ± 0.00604	5.58128 ± 0.26329	4.18453 ± 0.13384	2.92431 ± 0.20347
CUSUM-LinUCB	0.10532 ± 0.00107	5.61922 ± 0.12936	4.33342 ± 0.03940	2.79313 ± 0.05511
GLR-LinUCB	0.10321 ± 0.00141	5.24715 ± 0.16709	4.34858 ± 0.05603	2.91288 ± 0.05895

Results. RESPONSE-AWARE UCB has lower mean test cost than every external baseline on all four streams. At benchmark fees, it queries on every scored round on Jena and Delhi. The six-price Jena study then tests acquisition as fees increase. Our method has the lowest cost at all six prices (table 2), with equal search budgets per price. Across these price-specific configurations, its query fraction falls from 96% to 0.1%, spanning frequent consultation to near-free prediction.

Table 2: Mean Jena test cost at six consultation prices. Columns scale the base fee (≈ 0.17). Entries show mean ± sample standard deviation over five learner seeds. Bold marks the lowest mean per column. Fees and query percentages are in Appendix E.2.

Method	$\times 1$	$\times 3$	$\times 5$	$\times 7.5$	$\times 12.5$	$\times 50$
RESPONSE-AWARE UCB	**3.711**	**3.984**	**4.024**	**4.079**	**4.117**	**4.174**
SleepingLinUCB	4.030	4.277	4.240	4.444	4.417	4.190
SharedLinUCB	4.022	4.201	4.180	4.242	4.202	4.182
D-LinUCB	4.078	4.371	4.447	4.452	4.485	4.552
LinTS	4.119 ± 0.049	4.167 ± 0.040	4.177 ± 0.036	4.233 ± 0.044	4.250 ± 0.065	4.182
NeuralUCB	4.185 ± 0.134	4.223 ± 0.108	4.190 ± 0.069	4.169 ± 0.045	4.189 ± 0.020	4.182
CUSUM-LinUCB	4.333 ± 0.039	4.331 ± 0.073	4.328 ± 0.030	4.362 ± 0.050	4.415 ± 0.051	4.666 ± 0.048
GLR-LinUCB	4.349 ± 0.056	4.489 ± 0.038	4.557 ± 0.021	4.574 ± 0.066	4.641 ± 0.049	5.113 ± 0.085

Extra experiments. Appendix E.1 gives full learning trajectories, and Appendix E.2 reports cost and query-rate curves. Appendix E.3 holds the cubic correction learner, pooling, and memory fixed while comparing joint and disjoint UCB routers. Joint response UCB uses our risk features with unit-weight regression; disjoint response UCB fits a separate risk model per expert. Each selects its confidence and risk regularization from 256 configurations per stream. Our router matches joint response UCB on three streams and improves on both alternatives on Delhi.

8 CONCLUSION

We studied online learning to defer when a purchased answer can be corrected before deployment. RESPONSE-AWARE UCB learns these corrections and accounts for their learning error when deciding which answers to buy. Under the stated response and risk models, the calibrated algorithm achieves sublinear pseudo-regret against the conditional response oracle; an extension quantifies the effect of approximation errors. On four test streams, the selected cubic policy has lower cost than seven direct-deferral baselines. With the correction learner held fixed, our router matches or improves on the tested UCB alternatives. Six-price comparisons measure how consultation cost and query rates vary with fees.

AI USE STATEMENT

We used large language model assistants to polish the writing, check mathematical arguments and proofs. We were assisted by AI to develop the nonlinear-correction example in Appendix C.3. We have also generated the synthetic data with AI and checked very carefully its soundness. We also used these tools to implement and debug methods and experimental code. We reviewed the AI-assisted material and take full responsibility for the final text, mathematical claims, code, and results.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.html.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 127–135, 2013. URL https://proceedings.mlr.press/v28/agrawal13.html.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, Heidelberg, 3 edition, 2006. doi: 10.1007/3-540-29587-9.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9:1823–1840, 2008.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(77):1–40, 2022. URL https://jmlr.org/papers/v23/20-1384.html.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441. URL https://web.stanford.edu/~boyd/cvxbook/.

Yuzhou Cao, Tianchi Cai, Lei Feng, Lihong Gu, Jinjie Gu, Bo An, Gang Niu, and Masashi Sugiyama. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in neural information processing systems*, 35:521–534, 2022.

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006. doi: 10.1017/CBO9780511546921.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans, 2022.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Algorithmic Learning Theory*, pp. 67–82, 2016.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, 92(2):277–315, 2024.

Nina L. Corvelo Benz and Manuel Gomez Rodriguez. Counterfactual inference of second opinions. In *Uncertainty in Artificial Intelligence*, pp. 453–463, 2022. URL https://proceedings.mlr.press/v180/corvelo-benz22a.html.

Giulia DeSalvo, Clara Mohri, Mehryar Mohri, and Yutao Zhong. Budgeted multiple-expert deferral, 2025. URL https://arxiv.org/abs/2510.26706.

Dang Hoang Duy, Yannis Montreuil, Maxime Meyer, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Online learning-to-defer with varying experts, 2026. URL https://arxiv.org/abs/2605.12340.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210, 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. Smoothness-adaptive contextual bandits. *Operations Research*, 70(6):3198–3216, 2022. doi: 10.1287/opre.2021.2215.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. URL https://hastie.su.domains/ElemStatLearn/.

Elad Hazan. Introduction to online convex optimization, 2023. URL https://arxiv.org/abs/1909.05207v3. Second edition.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35:34614–34625, 2022.

Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2478–2484, 2022. doi: 10.24963/ijcai.2022/344. URL https://doi.org/10.24963/ijcai.2022/344.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2007–2010, 2020. URL https://proceedings.mlr.press/v125/hu20a.html.

Rob J. Hyndman and Yangzhuoran Yang. tsdl: Time series data library—daily minimum temperatures in melbourne, australia, 1981–1990. https://pkg.yangzhuoranyang.com/tsdl/, 2018. Original source: Australian Bureau of Meteorology.

Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3959–3966, 2019. doi: 10.1609/aaai.v33i01.33013959. URL https://arxiv.org/abs/1711.07364.

Shalmali Joshi, Sonali Parbhoo, and Finale Doshi-Velez. Learning-to-defer for sequential medical decision-making under uncertainty. *arXiv preprint arXiv:2109.06312*, 2021.

Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4421–4434, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/234b941e88b755b7a72a1c1dd5022f30-Abstract.html.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 154–165, 2021. doi: 10.1145/3461702.3462516. URL https://doi.org/10.1145/3461702.3462516.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80(2–3):245–272, 2010. URL https://research.ibm.com/publications/regret-bounds-for-sleeping-experts-and-bandits.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010. doi: 10.1145/1772690.1772758.

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11746. URL https://ojs.aaai.org/index.php/AAAI/article/view/11746.

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Mitigating underfitting in learning to defer with consistent losses. In *International Conference on Artificial Intelligence and Statistics*, pp. 4816–4824, 2024. URL https://proceedings.mlr.press/v238/liu24h.html.

Yusha Liu, Yining Wang, and Aarti Singh. Smooth bandit optimization: Generalization to Hölder space. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2206–2214. PMLR, 2021. URL https://proceedings.mlr.press/v130/liu21f.html.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0b17d256cf1fe1cc084922a8c6b565b7-Abstract-Conference.html.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pp. 822–867, 2024a. URL https://proceedings.mlr.press/v237/mao24a.html.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Workshop on Combinatorial Image Analysis*, pp. 107–135. Springer, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34738–34759, 2024c. URL https://proceedings.mlr.press/v235/mao24d.html.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 4753–4761, 2024d. URL https://proceedings.mlr.press/v238/mao24a.html.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. No-regret reinforcement learning in smooth MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34760–34789. PMLR, 2024. URL https://proceedings.mlr.press/v235/maran24a.html.

Max Planck Institute for Biogeochemistry. Jena climate dataset 2009–2016. https://www.bgc-jena.mpg.de/wetter/, n.d. Weather Station Jena (Saaleaue); observations from 2009–2016. Accessed 22 September 2026.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://proceedings.mlr.press/v267/montreuil25a.html.

Yannis Montreuil, Yeo Shu Heng, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. A two-stage learning-to-defer approach for multi-task learning. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://proceedings.mlr.press/v267/montreuil25b.html.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Beyond augmented-action surrogates for multi-expert learning-to-defer. *arXiv preprint arXiv:2604.09414*, 2026a.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Why ask one when you can ask k? learning-to-defer to the top-k experts. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=mGbEv4kVoG.

Yannis Montreuil, Yeo Shu Heng, Lai Xing Ng, Axel Carlier, and Wei Tsang Ooi. Optimal query allocation in extractive QA with LLMs: A learning-to-defer framework with theoretical guarantees. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026c. URL https://openreview.net/forum?id=kEVupwepTq.

Yannis Montreuil, Leïna Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Learning-to-defer with expert-conditional advice, 2026d. URL https://arxiv.org/abs/2603.14324.

Yannis Montreuil, Lai Xing Ng, Axel Carlier, Wei Tsang Ooi, and Yu Letian. Adversarial robustness in one-stage learning-to-defer. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026e. URL https://openreview.net/forum?id=gF8tv1SaR2.

Yannis Montreuil, Letian Yu, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Learning-to-defer in non-stationary time series via switching state-space models, 2026f. URL https://arxiv.org/abs/2601.22538.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL https://proceedings.mlr.press/v206/mozannar23a.html.

Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. Probabilistic learning to defer: Handling missing expert annotations and controlling workload distribution. In *International Conference on Learning Representations*, volume 2025, pp. 48450–48470, 2025.

Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/571d3a9420bfd9219f65b643d0003bf4-Abstract.html.

Filippo Palomba, Andrea Pugnana, Jose Manuel Alvarez, and Salvatore Ruggieri. A causal framework for evaluating deferring systems. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://proceedings.mlr.press/v258/palomba25a.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Andrea Pugnana, Giovanni De Toni, Cesare Barbera, Roberto Pellungrini, Bruno Lepri, and Andrea Passerini. To ask or not to ask: Learning to require human feedback. *arXiv preprint arXiv:2510.08314*, 2025.

Naveen Raman and Michael Yee. Improving learning-to-defer algorithms through fine-tuning, 2021. URL https://arxiv.org/abs/2112.10768.

Florence Regol, Joseph Cotnareanu, Theodore Glavas, and Mark Coates. Is the acquisition worth the cost? surrogate losses for consistent two-stage classifiers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/085ea366002345cab8a1bf0f0ad1b210-Abstract-Conference.html.

Lev Reyzin, Aadirupa Saha, and Shuo Wu. Expert advice with costly observations. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5700–5715, 2026. URL https://proceedings.mlr.press/v337/reyzin26a.html.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/263fc48aae39f219b4c71d9d4bb4aed2-Abstract.html.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 280–287, 2014. URL https://proceedings.mlr.press/v32/seldin14.html.

Joshua Strong, Qianhui Men, and Alison Noble. Towards human-AI collaboration in healthcare: Guided deferral systems with large language models. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/forum?id=4c5rg9y4me.

Joshua Strong, Qianhui Men, and J Alison Noble. Trustworthy and practical ai for healthcare: A guided deferral system with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28413–28421, 2025.

Joshua Strong, Pramit Saha, Yasin Ibrahim, Cheng Ouyang, and Alison Noble. Identity-free deferral for unseen experts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4YG9ufFg58.

sumanthvrao. Daily climate time series data (delhi, 2013–2017). https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data, 2019. Kaggle dataset, version 3. Original source: Weather Underground; observations from 2013–2017.

Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, Rhode Island, 4 edition, 1975. URL https://bookstore.ams.org/coll-23.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3475–3483, 2024. URL https://proceedings.mlr.press/v238/tailor24a.html.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 11415–11434, 2023. URL https://proceedings.mlr.press/v206/verma23a.html.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11492–11502, 2020. URL https://proceedings.mlr.press/v119/zhou20a.html.

A PROBABILITY SPACE AND TIMING

A.1 INFORMATION AND TIMING

Figure 1 summarizes the timing described in section 4. The distinction between choosing a function and evaluating it explains how a prediction can depend on a purchased answer even though its response coefficients are fixed before purchase.

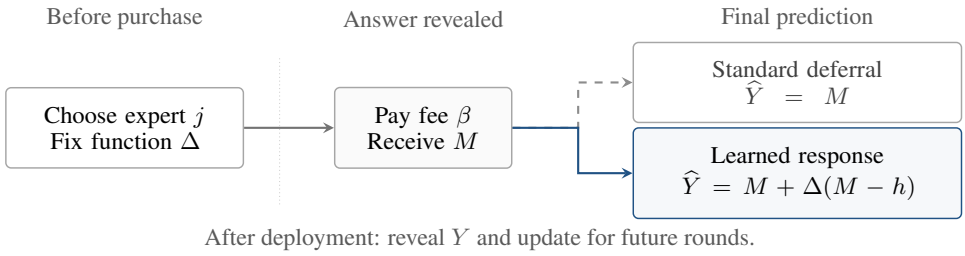

Figure 1: Consultation with a learned response, with round and expert subscripts suppressed. The solid path shows our queried action; the dashed branch shows standard deferral for comparison. The entire function Δ is fixed before observing M; its value is evaluated afterward. Free prediction, omitted here, uses h without paying a fee.

All random variables are defined on a common probability space admitting regular conditional laws for the finite-dimensional observations. We choose conditional versions simultaneously over the countable rounds and finite registry. The pre-query field $\mathcal{F}_t$ includes completed outcomes, purchased answers, algorithm state, current public features, h_t, availability and fees. The decision and response coefficients are $\mathcal{F}_t$-measurable. For a queried action, the intermediate field is $\mathcal{G}_t = \mathcal{F}_t \vee \sigma(M_{t,I_t})$; on a free round it is $\mathcal{F}_t$. The deployed prediction is $\mathcal{G}_t$-measurable, and Y_t is revealed afterward. The next pre-query field contains $\mathcal{G}_t$ and Y_t. Thus a purchased answer can enter subsequent response fits, while the current identity and response function are chosen before that answer is observed.

Conditional on $\mathcal{F}_t$, a common joint law specifies Y_t and all potential expert answers. The current action reveals coordinates of this vector without changing its law. Neither the expert answers nor the rounds need to be independent. An internal predictor may update causally on every outcome, provided its current h_t is fixed before routing and the stated model assumptions continue to hold. A returning identity keeps its learned state; the guarantee requires its conditional law on return to satisfy the same fixed parameter models.

To define the conditional oracle, we need a measurable choice of response. For fixed t, j, its objective is a continuous quadratic on the compact ball $\mathbb{B}_q$, with coefficients given by measurable conditional moments. The measurable maximum theorem (Aliprantis & Border, 2006) therefore gives a measurable optimizer. Minimizing over a finite eligible set with a fixed tie rule gives a measurable oracle action. We use these versions throughout the proofs. Because rounds and identities are countable, the boundedness assumptions may be imposed on a common probability-one event.

The fee screen preserves the oracle value. Since $0 \le L_t^0 \le R^2$, every query with $\beta_{t,j} \ge R^2$ has conditional cost at least R^2 even if its prediction loss is zero. Excluding those actions preserves the oracle value, including boundary ties. The learner consequently incurs at most $4G^2 + R^2$ per round, and its pseudo-regret is nonnegative because its deployed coefficient belongs to the oracle's response ball.

A.2 NOTATION GUIDE

The main notation is collected below. Public quantities, including the features and response coefficients, are known given $\mathcal{F}_t$; the outcome and current expert answers are still hidden. Superscripts H, P distinguish the identity and shared response proposals. Superscripts $0, *$ distinguish the free-risk and optimal queried-risk regressions.

Table 3: Decisions, responses, and costs. All risks exclude consultation fees.

Symbol	Meaning
K, T	Registry capacity and interaction horizon.
$\mathcal{F}_t$	Information available before the round-t query.
$\mathcal{E}_t, \mathcal{A}_t$	Available identities and the eligible subset with fees below R^2.
$I_t, \beta_{t,j}$	Chosen action (0 means free prediction) and the fee for expert j.
$h_t, M_{t,j}, Y_t$	Internal prediction, potential expert answer, and outcome.
$d_{t,j}, e_{t,j}$	Disagreement $M_{t,j} - h_t$ and expert residual $Y_t - M_{t,j}$.
R, D, G	Bounds on $\lvert Y_t - h_t\rvert$, $\lvert d_{t,j}\rvert$, and $\lvert e_{t,j}\rvert$, with $G = R + D$.
$\phi_q, \mathbb{B}_q$	Dictionary in $\mathbb{R}^{q+1}$ and its coefficient unit ball.
$a^H_{t,j}, a^P_{t,j}, a_{t,j}$	Identity proposal, shared proposal, and their deployed mixture.
$\widehat{Y}_t, C_t$	Deployed prediction and realized loss plus fee.
$L^0_t, W^*_{t,j}$	Free risk and minimum queried risk over $\mathbb{B}_q$.
C^*_t, Reg_T	Oracle conditional cost and cumulative conditional pseudo-regret.

Table 4: Learning state and confidence quantities. Dimensions count feature coordinates.

Symbol	Meaning
$\ell_{t,j} \in \mathbb{R}^{d_\ell},\ s_{t,j} \in \mathbb{R}^{d_s}$	Public features for identity and shared responses.
$\Theta_{t,j}, \Psi_t$	Response matrices of sizes $d_\ell \times (q+1)$ and $d_s \times (q+1)$.
$w^H_{t,j}, w^P_{t,j}, \eta$	Proposal mixture weights and their exponential-weights rate.
$x^0_t \in \mathbb{R}^{d_0},\ c_{t,j} \in \mathbb{R}^{d_c}$	Public features for free and optimal queried risk.
$\Lambda_v, V^v_t, \widehat{\theta}^v_t$	Prior, pre-round design, and risk estimate, for $v \in \{0, *\}$.
λ_v, Q_v, B_v	Minimum prior eigenvalue, prior-norm bound, and noise-plus-prior radius.
$u_{t,j}, \omega_t, \tau$	Queried-risk width, regression sample weight, and its threshold.
$\alpha, S_{t,j}$	Confidence multiplier and acquisition score.
Z_t, ρ_t	Queried loss divided by $4G^2$ and response excess risk divided by G^2.
$N_{t,j}, N_t$	Queries through round t to identity j, and in total.
$A_t, E_t, \overline{E}_T$	Realized response-loss budget, conditional-risk budget, and deterministic horizon envelope.

In particular, w^H, w^P combine responses, while ω_t weights a risk-regression observation. Designs V^v_t use rounds strictly before t, whereas $N_{t,j}$ includes round t. We write Id_r for the $r \times r$ identity matrix to distinguish it from the chosen action I_t.

B IMPLEMENTATION AND EXPLICIT REGRET BOUND

This appendix gives the response budget and recursive updates for the identity model, followed by the connection between our assumptions and online-learning models, a continuous-context example, and the explicit finite-horizon regret bound. The proof of that bound appears in Appendix D.

Algorithm 1 RESPONSE-AWARE UCB: online updates

Require: Priors, response regularizers, $0 < \eta \leq 1/2$, score multiplier $\alpha \geq 0$, and weight threshold $\tau > 0$.

Initialize all regression sums and query counts to zero, designs to their priors, and each identity's pool weights to $(1/2, 1/2)$. Preserve all states during absence.
for $t = 1, \ldots, T$ **do**
  Observe $h_t, \mathcal{E}_t$, fees and public features; register new identities and restore returning states.
  Set $\mathcal{A}_t = \{j \in \mathcal{E}_t : \beta_{t,j} < R^2\}$. Compute the ridge forecasts in equation (15).
  For each $j \in \mathcal{A}_t$, freeze $a^H_{t,j}$ from equation (8), $a^P_{t,j}$ from equations (9) and (10), and $a_{t,j}$ from equation (11).
  Compute equation (17); set I_t to its first maximizer in the order $(0, \mathcal{A}_t)$, with experts ordered by registration.
  if $I_t = 0$ **then**
    Deploy $\widehat{Y}_t = h_t$.
  else
    Freeze ω_t by equation (16) with the chosen τ.
    Pay β_{t,I_t}, observe M_{t,I_t}, and deploy $\widehat{Y}_t = M_{t,I_t} + G a_{t,I_t}^\top \phi_q(M_{t,I_t} - h_t)$.
  end if
  Observe Y_t; update free ridge with $(x_t^0, (Y_t - h_t)^2/R^2)$.
  if $I_t \neq 0$ **then**
    Form Z_t from the deployed prediction; add $\omega_t c_{t,I_t} c_{t,I_t}^\top$ to the queried design and $\omega_t c_{t,I_t} Z_t$ to its cross-product.
    Add the purchased observation to the two response objectives equations (8) and (9); update the pool using the frozen proposal losses and increment N_{t,I_t}.
  end if
end for

B.1 THE RESPONSE-LEARNING BOUND

The bound E_t in equation (14) accounts for the loss accumulated while fitting and combining responses. Let $p_I = d_\ell(q+1)$ be the number of entries in each identity's response matrix. The constrained least-squares comparison in lemma D.1 contributes a regularization term and a logarithmic term for each identity. The exponential-weights comparison in lemma D.2 adds $4 \log 2/\eta$ per identity. Their sum is the pathwise bound

$$A_t = K\lambda_I + 8p_I \sum_{j=1}^{K} \log\left(1 + \frac{N_{t,j}}{p_I \lambda_I}\right) + \frac{4K \log 2}{\eta}. \tag{23}$$

This is a bound on the cumulative realized loss of the deployed response relative to the optimal identity response, in normalized units. It is a function of the observed query counts, rather than a deterministic constant. The normalization divides squared losses by G^2; the comparison uses the fixed optimal matrices in assumption 6.1. Here $N_{t,j}$ counts queries through round t, with all counts initially zero, and $N_t = \sum_j N_{t,j}$ counts all purchases. The calibrated router uses the deterministic horizon envelope below to fix its weighting threshold before interaction.

Lemma D.3 converts this realized-loss bound into a conditional excess-risk bound. Define

$$E_t = 2A_t + 64 \log(3/\delta). \tag{24}$$

With probability at least $1-\delta/3$, the inequality $\sum_{s\le t}\rho_s \le E_t$ holds simultaneously over all prefixes. Concavity of the logarithm and $\sum_j N_{t,j} \le T$ give the deterministic horizon envelope

$$\overline{E}_T = 2\left[K\lambda_I + 8Kp_I \log\left(1+\frac{T}{Kp_I\lambda_I}\right) + \frac{4K\log 2}{\eta}\right] + 64\log\frac{3}{\delta}, \qquad E_t \le \overline{E}_T \quad (t\le T). \tag{25}$$

This is the envelope used to set the weight threshold in equation (21). The factor 3 reserves the other two parts of the failure probability for the free and queried risk regressions.

B.2 Risk regressions and confidence calibration

In the benchmarks, the free, identity-response and shared-response features use the same bounded public state: $x_t^0 = \ell_{t,j} = s_{t,j}$ for every active expert j. The queried-risk fit shares information across experts while retaining separate identity coordinates. Its feature in the implementation is $c_{t,j} = 2^{-1/2}(s_{t,j}, \boldsymbol{e}_j \otimes \ell_{t,j})$, where the vectors are stacked, $\boldsymbol{e}_j \in \mathbb{R}^K$ is the registry basis vector, and $\otimes$ denotes the Kronecker product. Thus $d_c = d_s + Kd_\ell$. Write Id_r for the $r\times r$ identity matrix, reserving I_t for the selected action. For strengths $\lambda_s, \lambda_i > 0$, the block-diagonal prior $\Lambda_* = \mathrm{diag}(\lambda_s \mathrm{Id}_{d_s}, \lambda_i \mathrm{Id}_{Kd_\ell})$ controls shrinkage separately on shared and identity coordinates. Both priors are fixed before interaction. The full-history estimates in equation (15) use only observations available before round t. Let $\mathrm{clip}_{[a,b]}(v) = \min\{b, \max\{a, v\}\}$ denote scalar clipping. Use $\widehat{L}_t^0 = R^2\,\mathrm{clip}_{[0,1]}((x_t^0)^\top \widehat{\theta}_t^0)$ and the raw forecast $\widetilde{W}_{t,j} = 4G^2 c_{t,j}^\top \widehat{\theta}_t^*$. Clipping the free forecast cannot increase its distance from the true free risk in $[0, R^2]$. We retain the raw queried forecast when applying the confidence inequality that accounts for response-learning bias.

Let Q_v bound the corresponding parameter norm in assumption 6.2, measured by $\|\cdot\|_{\Lambda_v}$. For $v \in \{0, *\}$, put $\lambda_v = \lambda_{\min}(\Lambda_v) > 0$ and $d_* = d_c$. A valid horizon calibration is

$$B_v = \tfrac{1}{2}\sqrt{d_v \log(1 + T/(d_v\lambda_v)) + 2\log(3/\delta)} + Q_v, \qquad \alpha = 2B_*, \quad \tau = 4B_*/\overline{E}_T. \tag{26}$$

The explicit regret bound below uses this coupled calibration. One may take $Q_v = \sqrt{\lambda_{\max}(\Lambda_v)}S_v$ from a known Euclidean norm bound S_v. More precisely, the self-normalized radius uses $\frac{1}{2}\sqrt{\log(\det V_t^v / \det \Lambda_v) + 2\log(3/\delta)} + Q_v$. The horizon calibration bounds this radius throughout the analyzed rounds.

B.3 Computing the response fits

The identity response is obtained by solving a quadratic problem on a unit ball. For identity j, vectorize its response matrix row-wise and write $z_s = \ell_{s,j} \otimes \phi_q(d_{s,j}) \in \mathbb{R}^{p_I}$ and $y_s = e_{s,j}/G$ on rounds with $I_s = j$. Its pre-query design and cross-product vector are

$$V = \lambda_I \mathrm{Id}_{p_I} + \sum_{s<t:I_s=j} z_s z_s^\top, \qquad b = \sum_{s<t:I_s=j} z_s y_s.$$

The constrained objective is a positive-definite quadratic on the Euclidean unit ball. The quadratic optimality conditions (Boyd & Vandenberghe, 2004) imply that $V^{-1}b$ is the solution whenever it is feasible. Otherwise the solution is $(V+\mu \mathrm{Id}_{p_I})^{-1}b$, where $\mu > 0$ is the unique root of $\|(V+\mu \mathrm{Id}_{p_I})^{-1}b\|_2 = 1$. After diagonalization, the norm decreases monotonically with the multiplier, so a scalar root search finds the solution. We compute the constrained fit using 90 bisection iterations and check coefficient feasibility with tolerance 10^{-10}. The analysis uses exact minimizers.

The shared matrix Ψ_t solves the unconstrained ridge problem in equation (9); its proposal is projected by equation (10). This regression uses the product feature $s_{r,I_r} \otimes \phi_q(d_{r,I_r})$ with row-wise vectorization of Ψ_t. It incorporates the observed answer and outcome only after deployment. The proposal remains predictable and bounded even if the shared model is misspecified. The queried label is computed from the coefficient average fixed before querying; the later fit updates do not change that label.

B.4 Recursive implementation of the deployed policy

All quantities below use the learner's normalized units; the dataset-specific rescaling of observations, bounds, fees and reported costs is specified in Appendix F.1. In empirical runs, $\alpha_{\mathrm{emp}}, \tau_{\mathrm{emp}}$ denote

the selected values of α, τ. Set $p_P = d_s(q+1)$, the shared response dimension. A scalar-target ridge state consists of an inverse design $P \in \mathbb{R}^{p\times p}$ and a cross-product vector $b \in \mathbb{R}^p$. Given a feature $z \in \mathbb{R}^p$ and target $y \in \mathbb{R}$, define the update $\mathsf{U}(P, b; z, y)$ by

$$v = Pz, \qquad P^+ = P - \frac{vv^\top}{1 + z^\top v}, \qquad b^+ = b + zy. \tag{27}$$

The implementation symmetrizes P^+ after each update. The prediction is $z^\top Pb$ and its uncertainty width is $\sqrt{z^\top Pz}$. A negative squared width within the implementation tolerance 10^{-10} is set to zero; a larger negative value raises an error. These rank-one updates implement the full-history objectives above in exact arithmetic.

Initialize the free and queried states as $(P_0, b_0) = (\Lambda_0^{-1}, 0)$ and $(P_*, b_*) = (\Lambda_*^{-1}, 0)$, respectively, and the shared response state as $(P_P, b_P) = (\lambda_P^{-1}\mathrm{Id}_{p_P}, 0)$. The implemented free prior is $\Lambda_0 = \lambda_0 \mathrm{Id}_{d_0}$, with $\lambda_0 > 0$; the queried prior has the shared and identity blocks given above. For each newly registered identity j, initialize $(P_j, b_j) = (\lambda_I^{-1}\mathrm{Id}_{p_I}, 0)$, count $N_j = 0$, and cumulative proposal losses $H_j^H = H_j^P = 0$. New identities are registered in their supplied public order; existing registration ranks never change. Absence changes none of these states. The registry has capacity K.

Before querying, reshape the constrained solution from (P_j, b_j) row-wise into $\Theta_{t,j}$, and reshape $P_P b_P$ row-wise into Ψ_t. Use equations (8) and (10) for the proposals. Their weights are proportional to $\exp(-\eta H_j^u)$ for $u \in \{H, P\}$; in code, subtracting the largest log weight before exponentiation stabilizes this calculation. The pool contains the shared and identity proposals, both with initial weight $1/2$, and its coefficient is always deployed on a query. The queried-risk width is computed from P_* before acquisition. After selecting a queried identity, freeze its weight $\omega_t = \min\{1, \tau_{\mathrm{emp}}/\sqrt{c_{t,j}^\top P_* c_{t,j}}\}$, using one at zero width. Neither the answer nor the outcome enters this weight. Pool weights w^H, w^P and the regression weight ω_t have different roles. The queried forecast is raw, the free forecast is clipped, and there is no free-risk optimism term. Free prediction has priority on score ties, followed by the earliest registered eligible identity.

After observing the outcome, update $(P_0, b_0) \leftarrow \mathsf{U}(P_0, b_0; x_t^0, (Y_t - h_t)^2/R^2)$ on every round. If $j = I_t \neq 0$, retain the pre-query proposals and public features, and put $d = M_{t,j} - h_t$, $y = (Y_t - M_{t,j})/G$, and $\phi = \phi_q(d)$. Compute Z_t from the prediction already deployed, then apply

$$\begin{aligned} (P_*, b_*) &\leftarrow \mathsf{U}(P_*, b_*; \sqrt{\omega_t} c_{t,j}, \sqrt{\omega_t} Z_t), \\ (P_j, b_j) &\leftarrow \mathsf{U}(P_j, b_j; \ell_{t,j} \otimes \phi, y), \\ (P_P, b_P) &\leftarrow \mathsf{U}(P_P, b_P; s_{t,j} \otimes \phi, y), \qquad N_j \leftarrow N_j + 1. \end{aligned} \tag{28}$$

For each frozen proposal, update

$$H_j^u \leftarrow H_j^u + \mathcal{L}_t^u, \qquad \mathcal{L}_t^u = \tfrac{1}{4}\big(y - (a_{t,j}^u)^\top \phi\big)^2, \qquad u \in \{H, P\}. \tag{29}$$

Equivalently, multiply its current weight by $\exp(-\eta \mathcal{L}_t^u)$ and normalize the two weights. All targets and proposal losses use the observations and proposals already specified for the round. These state updates therefore give the same next-round mathematical state in any order. Other identities retain their fits, counts and weights. Free rounds update none of the queried or response states. There is no clipping of the final corrected prediction. Observed residuals and disagreements are checked against the declared bounds, permitting only one floating-point unit beyond the boundary; a violation raises an error.

Both feature and target are scaled by $\sqrt{\omega_t}$, so the queried cross-product increment is $\omega_t c_{t,j} Z_t$. The response fits retain their unweighted updates. The theoretical run uses equation (21) for these two constants and exact constrained minimization. The benchmark run uses the numerical solver specified above.

B.5 Model assumptions in context

Our assumptions connect two learned objects to the response oracle: the correction coefficients and the risk attained by an optimal correction. Their roles follow established online-regression and linear-bandit analyses.

Optimal-response realizability. Realizability connects a learned function class to an oracle decision rule. For example, Foster & Rakhlin (2020, Assumption 1) assume that one function in the regression class represents conditional mean costs, and their Assumption 2a compares online squared loss with a fixed function on adaptively chosen observations. Our assumption 6.1 instead represents a conditional risk minimizer: one feasible matrix per identity produces optimal coefficients across contexts. It lets us compare the online response fit with the oracle using the fixed-comparator principle of online convex optimization (Hazan, 2023, Chapter 3). The exact constrained least-squares bound used here is proved in lemma D.1. The optimal correction may be nonunique or lie on the coefficient-ball boundary, and need not recover the conditional mean residual. The shared model may be misspecified; its projected response supplies a feasible auxiliary proposal.

Linear risk models. Abbasi-Yadkori et al. (2011, Theorem 2) construct confidence ellipsoids for a fixed linear conditional mean under a parameter-norm bound and conditionally sub-Gaussian noise. Bounded features and parameter norms also appear in Agrawal & Goyal (2013, Section 2.1) and He et al. (2022, Assumption 2.1). These are the conditions used for our risk regressions, with the optimal queried risk as their latent target. Modeling that target separately from the coefficients specifies both what the learner should deploy and what it should expect to pay. For a general positive-definite prior, a Euclidean norm bound S_v gives $Q_v = \sqrt{\lambda_{\max}(\Lambda_v)}S_v$, as used in equation (26).

Bounded observations and adaptive contexts. Our residual and coefficient bounds put the normalized squared-loss labels in $[0, 1]$, the cost range also used by Foster & Rakhlin (2020, Section 1.1). Centering these labels conditionally gives $1/2$-sub-Gaussian noise by Hoeffding's lemma. Thus the noise condition for self-normalized regression follows from the protocol rather than a Gaussian answer model. Predictable features can depend on previous observations, as in Abbasi-Yadkori et al. (2011, Theorem 1) and Agrawal & Goyal (2013, Section 2.1); independent contexts are not needed. In our queried regression, the remaining conditional bias is the response-learning gap in equation (13). We bound its cumulative size and its contribution to the confidence radius in Appendix D.4.

B.6 A CONTINUOUS-CONTEXT EXAMPLE

The following construction satisfies both main model assumptions with continuously varying contexts. Let $x_t \in [-1, 1]$ be public and set $\ell_{t,j} = (1, x_t)^\top/\sqrt{2}$. Conditional on $\mathcal{F}_t$, let Y_t be uniform on $\{-0.3, 0.3\}$, with internal prediction $h_t = 0$ and expert answer $M_{t,j} = Y_t + 0.2 + 0.1x_t$. Take $q = 3$, $R = 0.3$, $D = 0.7$ and $G = 1$. The correction $\Delta(d) = -0.2 - 0.1x_t$ attains zero squared loss and is represented by

$$\Theta_j^* = -\sqrt{2(1 + 3D^2)} \begin{pmatrix} 0.2 & 0 & 0 & 0 \\ 0.1 & 0 & 0 & 0 \end{pmatrix}.$$

Its squared Frobenius norm is $2(1 + 3D^2)(0.2^2 + 0.1^2) = 0.247 < 1$, so it satisfies assumption 6.1 over the full interval of contexts. The free risk is 0.09 and optimal queried risk is zero. For example, choose $x_t^0 = 1$ and $s_{t,j} = \ell_{t,j}$, with the queried feature defined above. Then $\theta_0^* = 1$ and $\theta_* = 0$ give the normalized risk models in assumption 6.2, with finite norms for any fixed positive-definite priors. The optimal correction varies with context while the risks remain constant.

B.7 EXPLICIT FINITE-HORIZON BOUND

The regret bound combines uncertainty in the free-risk fit, uncertainty in the queried-risk fit, and the loss from learning responses. For an observation count $n \geq 0$, a positive integer dimension d_V, and a prior eigenvalue bound $\lambda > 0$, define the cumulative-width bound

$$\mathcal{W}(n, d_V, \lambda) = \sqrt{n(1 + \lambda^{-1})d_V \log(1 + n/(d_V\lambda))}, \qquad \mathcal{W}(0, d_V, \lambda) = 0. \tag{30}$$

Define also

$$\mathcal{H}(n) = (1 + \lambda_*^{-1})d_c \log(1 + n/(d_c\lambda_*)), \qquad \mathcal{H}(0) = 0. \tag{31}$$

Proposition B.1 (Explicit routing bound)**.** *Under the hypotheses and calibration of theorem 6.3, with probability at least* $1 - \delta$*,*

$$\begin{aligned} \mathrm{Reg}_T \leq\ & R^2 B_0 \mathcal{W}(T, d_0, \lambda_0) \\ & + G^2 \left[16B_* \sqrt{N_T \mathcal{H}(N_T)} + 4\overline{E}_T \mathcal{H}(N_T) + E_T\right]. \end{aligned} \tag{32}$$

This bound retains the separate contributions of free-risk estimation, queried-risk estimation, and response learning. In particular, the queried-risk terms depend on the number of purchases, not all T rounds. Appendix D restates and proves this bound together with theorem 6.3, including the reduction to its stated rate.

C Response value and approximation

A response can remove the part of an expert's error that its answer predicts. We first describe this reduction in prediction loss, then account for fees, and finally quantify the extra gain available from cubic terms.

C.1 Prediction gain and approximation

To quantify the prediction gain, fix a round t and expert j, and abbreviate $e = e_{t,j}$ and $d = d_{t,j}$. Write $\nu = \mathbb{E}[e \mid \mathcal{F}_t, d]$. For a square-integrable response $\Delta(d)$ whose map is fixed under $\mathcal{F}_t$, the squared-loss projection identity (Hastie et al., 2009) gives

$$\mathbb{E}[(e - \Delta(d))^2 \mid \mathcal{F}_t] = \mathbb{E}[\mathrm{Var}(e \mid \mathcal{F}_t, d) \mid \mathcal{F}_t] + \mathbb{E}[(\nu - \Delta(d))^2 \mid \mathcal{F}_t]. \tag{33}$$

The first term is irreducible after observing the purchased answer; the second measures how well the chosen correction approximates the conditional mean residual. The best unrestricted correction is therefore ν, up to conditional L_2 equality. For direct deployment, $\Delta = 0$, so the removable part of its risk is $\mathbb{E}[\nu^2 \mid \mathcal{F}_t]$.

Comparison with an unrestricted response. Restore the indices on $\nu_{t,j} = \mathbb{E}[e_{t,j} \mid \mathcal{F}_t, d_{t,j}]$. Define the best conditional approximation error $\mathfrak{A}_{t,j}$ by

$$\mathfrak{A}_{t,j} = \inf_{a \in \mathbb{B}_q} \mathbb{E}[(\nu_{t,j} - Ga^\top \phi_q(d_{t,j}))^2 \mid \mathcal{F}_t].$$

Let C_t^{meas} be the oracle value with the same free action, eligible experts and fees, but with any square-integrable response to the purchased answer. The conditional mean response attains each queried minimum, so a minimum over the finite action set is attained; denote an optimizing action by j_t^{meas}. Set $\mathfrak{A}_{t,0} = 0$ for the free action. By equation (33), the bounded class adds $\mathfrak{A}_{t,j}$ to the unrestricted risk of expert j. Evaluating it at j_t^{meas}, and using inclusion of the response classes, gives

$$0 \le C_t^* - C_t^{\mathrm{meas}} \le \mathfrak{A}_{t,j_t^{\mathrm{meas}}}. \tag{34}$$

Summing and using the definition of Reg_T gives

$$\sum_{t=1}^{T} \{\mathbb{E}[C_t \mid \mathcal{F}_t] - C_t^{\mathrm{meas}}\} \le \mathrm{Reg}_T + \sum_{t=1}^{T} \mathfrak{A}_{t,j_t^{\mathrm{meas}}}.$$

The bound separates online learning error within the chosen response class from its approximation error relative to the conditional mean residual.

C.2 When correction changes the acquisition decision

The projection identity quantifies the prediction error a response can remove. To account for acquisition, this gain must be compared with the expert's fee and the other available actions. Use simultaneous conditional versions as in Appendix A. For a fixed round define

$$L_{t,j}^{\mathrm{dir}} = \mathbb{E}[e_{t,j}^2 \mid \mathcal{F}_t], \qquad C_t^{\mathrm{dir}} = \min\{L_t^0, \min_{j \in \mathcal{A}_t}(L_{t,j}^{\mathrm{dir}} + \beta_{t,j})\},$$

and, for every eligible expert,

$$\Gamma_{t,j} = L_{t,j}^{\mathrm{dir}} - W_{t,j}^*, \qquad \kappa_{t,j} = L_{t,j}^{\mathrm{dir}} + \beta_{t,j} - C_t^{\mathrm{dir}}.$$

The gain $\Gamma_{t,j}$ measures the prediction loss removed by correction; $\kappa_{t,j}$ is the expert's direct-cost disadvantage relative to the best available direct action. Both are nonnegative. The following identity compares these quantities without response or risk realizability.

$$C_t^{\mathrm{dir}} - C_t^* = \max\big(\{0\} \cup \{\Gamma_{t,j} - \kappa_{t,j} : j \in \mathcal{A}_t\}\big). \tag{35}$$

Indeed, $W_{t,j}^* + \beta_{t,j} = C_t^{\mathrm{dir}} + \kappa_{t,j} - \Gamma_{t,j}$. Substituting into the minimum defining C_t^* gives the identity: the zero term covers a free direct optimum, and otherwise a queried direct optimum has $\kappa_{t,j} = 0$ and $\Gamma_{t,j} \ge 0$. An empty eligible set gives zero on both sides.

Correction strictly improves the oracle cost exactly when its prediction gain exceeds some expert's direct-cost disadvantage. Along the learner's history, the cumulative oracle advantage is $\sum_t (C_t^{\mathrm{dir}} - C_t^*)$; the learner's conditional cost advantage subtracts Reg_T from that sum. The comparison uses the same realized public history for both costs.

C.3 AN EXAMPLE OF NONLINEAR CORRECTION

A short example isolates the error that an affine response cannot remove. Let $S \sim \mathrm{Unif}[-1, 1]$, $h = 0$, and let one always-available expert return $M = S$ at fee $0 < \beta < 1/3$. For $b, \varepsilon \in (0, 1/4]$, take $Y = S + b + \varepsilon P_3(S)$, $D = 1$, $R = 1 + b + \varepsilon$ and $G = R + D$. The public context is constant, so the answer is the only source of information about the current draw.

Legendre orthogonality (Szegő, 1975) gives $\mathbb{E}[S^2] = 1/3$, $\mathbb{E}[P_3(S)^2] = 1/7$ and $\mathbb{E}[SP_3(S)] = 0$. Both S and $P_3(S)$ are centered. Thus an affine response removes the constant bias b but leaves risk $\varepsilon^2/7$, while a cubic response removes both components. Writing L_q^* for the optimal one-round cost with order q and L_{hard}^* for direct deferral,

$$L_3^* = \beta, \qquad L_1^* = \beta + \varepsilon^2/7, \qquad L_{\mathrm{hard}}^* = \beta + \varepsilon^2/7 + b^2. \tag{36}$$

The optimal corrections are feasible: their coefficient vectors are $a_1^* = (\sqrt{2}\, b/G, 0)^\top$ and $a_3^* = (2/G)(b, 0, 0, \varepsilon)^\top$, respectively, each with norm strictly below one. Free prediction has risk $1/3 + b^2 + \varepsilon^2/7$, so $\beta < 1/3$ makes a query preferable in every class. Pre-query randomization independent of S cannot improve these minima.

This example separates bias correction from nonlinear correction. It does not rely on nesting the normalized coefficient balls: the two coefficient vectors verify feasibility in their respective classes.

D PROOF OF THE ROUTING GUARANTEE

The proof has four steps. We bound realized response regret on purchased observations and convert it into conditional excess risk. That excess risk controls the bias in the queried-risk regression. Confidence bounds then support a one-round comparison with the oracle. All probability statements use the filtration in Appendix A. We analyze the identity-response model in assumption 6.1 throughout. The budgets A_t, E_t and query counts $N_{t,j}, N_t$ are those in Appendix B.1.

D.1 CONSTRAINED LEAST SQUARES ON SELECTED OBSERVATIONS

The first step compares the response fit with a fixed feasible parameter on exactly the observations used for fitting. We follow the stability analysis of regularized leaders (Hazan, 2023), retaining the prior and query counts in the bound. Within the following lemma, n indexes observations supplied to the fit, rather than calendar rounds.

Lemma D.1 (Deterministic constrained-regression comparison)**.** *Let $z_n \in \mathbb{R}^p$ and $y_n \in \mathbb{R}$ be a sequence of features and targets with $\|z_n\|_2 \le 1$ and $|y_n| \le 1$. Let $\mathcal{C} \subset \mathbb{R}^p$ be compact and convex, contain zero, and satisfy $|z_n^\top \theta| \le 1$ for every $\theta \in \mathcal{C}$ and n. Let $\Lambda \in \mathbb{R}^{p\times p}$ be positive definite. For exact minimizers*

$$\theta_n = \arg\min_{\theta\in\mathcal{C}} \{\|\theta\|_\Lambda^2 + \sum_{i<n} (y_i - z_i^\top \theta)^2\},$$

put $V_0 = \Lambda$ and $V_n = \Lambda + \sum_{i=1}^n z_i z_i^\top$ for $n \ge 1$. For every $\theta^* \in \mathcal{C}$ and every finite N, pathwise,

$$\sum_{n=1}^{N} \{(y_n - z_n^\top \theta_n)^2 - (y_n - z_n^\top \theta^*)^2\} \le \|\theta^*\|_\Lambda^2 + 8 \log \frac{\det V_N}{\det \Lambda}. \tag{37}$$

Proof. We separate the comparison into the cost of regularization and the change in loss caused by one update. Write $l_n(\theta) = (y_n - z_n^\top \theta)^2$ and $F_n(\theta) = \|\theta\|_\Lambda^2 + \sum_{i\le n} l_i(\theta)$.

Comparator cost. Optimality of θ_n for F_{n-1} gives

$$F_n(\theta_{n+1}) - F_{n-1}(\theta_n) \ge l_n(\theta_{n+1}).$$

Summing and using $F_0(\theta_1) = 0$, since zero is feasible, gives $\sum_{n\le N} l_n(\theta_{n+1}) \le F_N(\theta^*)$. It remains to bound the loss change $\sum_n \{l_n(\theta_n) - l_n(\theta_{n+1})\}$ caused by the updates.

Update stability. Put $\Delta_n = \theta_n - \theta_{n+1}$ and $p_n = z_n^\top \theta_n$. The variational inequalities for the two constrained minimizers are

$$\nabla F_{n-1}(\theta_n)^\top \Delta_n \le 0, \qquad \nabla F_n(\theta_{n+1})^\top \Delta_n \ge 0.$$

Since $\nabla F_n(\theta_n) = \nabla F_{n-1}(\theta_n) + 2(p_n - y_n) z_n$ and the Hessian of F_n is $2V_n$, subtraction gives

$$\|\Delta_n\|_{V_n}^2 \le (p_n - y_n) z_n^\top \Delta_n \le 2|z_n^\top \Delta_n|.$$

Cauchy–Schwarz implies $\|\Delta_n\|_{V_n} \le 2\sqrt{z_n^\top V_n^{-1} z_n}$, including $\Delta_n = 0$. Consequently

$$\begin{aligned} l_n(\theta_n) - l_n(\theta_{n+1}) &= 2(p_n - y_n) z_n^\top \Delta_n - (z_n^\top \Delta_n)^2 \\ &\le 4|z_n^\top \Delta_n| \le 8 z_n^\top V_n^{-1} z_n. \end{aligned}$$

With $x_n = z_n^\top V_{n-1}^{-1} z_n$, the matrix inversion lemma gives $z_n^\top V_n^{-1} z_n = x_n/(1 + x_n) \le \log(1 + x_n)$. The determinant lemma identifies the sum of these logarithms as $\log(\det V_N / \det \Lambda)$. Combining the inequalities proves equation (37). The argument is deterministic and requires no covariance assumption. □

Let $\lambda = \lambda_{\min}(\Lambda)$. The trace of $\Lambda^{-1/2}(V_N - \Lambda)\Lambda^{-1/2}$ is at most N/λ. Applying concavity of $\log$ to its p nonnegative eigenvalues shows

$$\log \frac{\det V_N}{\det \Lambda} \le p \log(1 + N/(p\lambda)). \tag{38}$$

For the identity product domain, apply lemma D.1 to each identity's subsequence, with dimension $p_I = d_\ell(q + 1)$ and prior $\lambda_I \mathrm{Id}_{p_I}$. Let N_j count the observations for identity j among the first N queries. Summing gives $K\lambda_I + 8p_I \sum_j \log(1 + N_j/(p_I \lambda_I))$. Including unqueried identities in the prior term only enlarges the bound: their observed loss difference is zero.

D.2 Pooling and conditional excess risk

Pooling adds a fixed comparison term per identity, even when the shared proposal changes or is misspecified. The comparison uses exponential weights and exp-concavity (Cesa-Bianchi & Lugosi, 2006; Hazan, 2023). We verify the normalization and learning rate needed for this argument.

Lemma D.2 (Pathwise pool comparison)**.** *Fix an identity j and let N_j be its number of queries. On its nth query, denote the normalized target by y_n, the answer feature by ϕ_n, the proposal coefficients by a_n^H, a_n^P, and their pooled coefficient by a_n. These are the quantities from equation (11) restricted to this identity's queried subsequence: $|y_n| \le 1$, $\phi_n \in \mathbb{R}^{q+1}$ has norm at most one, and $a_n^H, a_n^P \in \mathbb{B}_q$. For $0 < \eta \le 1/2$ and $u \in \{H, P\}$,*

$$\sum_{n=1}^{N_j} (y_n - a_n^\top \phi_n)^2 \le \sum_{n=1}^{N_j} (y_n - (a_n^u)^\top \phi_n)^2 + \frac{4\log 2}{\eta}. \tag{39}$$

The proposals may vary with the observed history.

Proof. For fixed $y \in [-1, 1]$ and a scalar prediction $p \in [-1, 1]$, put $l(p) = (y-p)^2/4$. Differentiating twice gives

$$\frac{d^2}{dp^2} e^{-\eta l(p)} = e^{-\eta l(p)} \{\eta^2 (p-y)^2/4 - \eta/2\} \le 0$$

when $0 < \eta \le 1/2$, since $(p-y)^2 \le 4$. For round n, set $l_n(p) = (y_n - p)^2/4$. Write $p_n^u = (a_n^u)^\top \phi_n$ and let w_n^u be the weights on this subsequence. For $p_n = a_n^\top \phi_n = \sum_u w_n^u p_n^u$, concavity implies $e^{-\eta l_n(p_n)} \ge \sum_u w_n^u e^{-\eta l_n(p_n^u)}$. Write the exponential-weights normalizer after n queries as $\mathcal{Z}_n = \frac{1}{2}\sum_u \exp(-\eta \sum_{i \le n} l_i(p_i^u))$. Then $\mathcal{Z}_0 = 1$ and $\sum_u w_n^u e^{-\eta l_n(p_n^u)} = \mathcal{Z}_n / \mathcal{Z}_{n-1}$. Taking logarithms and summing gives $\sum_n l_n(p_n) \le -\log \mathcal{Z}_{N_j}/\eta$. The contribution of either candidate to $\mathcal{Z}_{N_j}$ is at least $\frac{1}{2}\exp(-\eta \sum_n l_n(p_n^u))$; multiplying by four proves the claim. The proposals may change between queries: the calculation compares their realized losses on the same observations. □

We now compare the deployed response with the conditional optimum. Combine lemmas D.1 and D.2, using candidate H and the fixed matrix Θ_j^* for each identity. For a queried round, define

$$y_t = e_{t,I_t}/G, \quad p_t = a_{t,I_t}^\top \phi_q(d_{t,I_t}), \quad p_t^* = \ell_{t,I_t}^\top \Theta_{I_t}^* \phi_q(d_{t,I_t}),$$

and let $X_t = (y_t - p_t)^2 - (y_t - p_t^*)^2$. The coefficients in p_t and p_t^* are fixed under $\mathcal{F}_t$; their evaluations depend on the purchased answer. Set $X_t = 0$ on free rounds. For every prefix, pathwise,

$$\sum_{s \le t} X_s \le A_t, \tag{40}$$

where A_t is defined in equation (23).

Lemma D.3 (Conditional response-risk certificate)**.** *Under assumption 6.1, let $\rho_t = \mathbb{E}[X_t \mid \mathcal{F}_t]$, which is zero on free rounds. For each $\zeta \in (0, 1)$, with probability at least $1 - \zeta$, simultaneously for all prefixes,*

$$\sum_{s \le t} \rho_s \le 2A_t + 64\log(1/\zeta). \tag{41}$$

Proof. The pathwise comparison alone concerns observed losses. To obtain a conditional-risk bound, we show that the loss difference has a second moment controlled by its conditional mean, then use a supermartingale.

Second-moment control. On a query, write $v_t = p_t - p_t^*$. Conditional optimality on the convex coefficient ball implies its first-order variational inequality $\mathbb{E}[(p_t^* - y_t)v_t \mid \mathcal{F}_t] \ge 0$. This follows by taking the directional derivative toward the feasible, $\mathcal{F}_t$-measurable coefficient a_{t,I_t}; boundedness justifies exchanging this derivative with conditional expectation. Hence

$$\rho_t = \mathbb{E}[v_t^2 + 2(p_t^* - y_t)v_t \mid \mathcal{F}_t] \ge \mathbb{E}[v_t^2 \mid \mathcal{F}_t] \ge 0. \tag{42}$$

Because $|y_t|, |p_t|, |p_t^*| \le 1$, we have $|X_t| \le 4$ and $X_t^2 = v_t^2(p_t + p_t^* - 2y_t)^2 \le 16 v_t^2$. Thus $\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \le 16\rho_t$ and $0 \le \rho_t \le 4$. **Uniform concentration.** For $\lambda = 1/32$, use $e^u \le 1 + u + \frac{1}{2}u^2 e^{|u|}$ to

obtain

$$\mathbb{E}[e^{-\lambda X_t} \mid \mathcal{F}_t] \le 1 - \lambda\rho_t + 8\lambda^2 e^{1/8}\rho_t \le 1 - \lambda\rho_t/2 \le e^{-\rho_t/64},$$

where $8\lambda e^{1/8} < 1/2$ verifies the middle inequality. Define

$$\mathcal{M}_t = \exp\left\{\sum_{s\le t}\rho_s/64 - \sum_{s\le t} X_s/32\right\}, \qquad \mathcal{M}_0 = 1.$$

The preceding bound gives $\mathbb{E}[\mathcal{M}_t \mid \mathcal{F}_t] \le \mathcal{M}_{t-1}$. Thus $(\mathcal{M}_t)$ is a nonnegative supermartingale with respect to $(\mathcal{F}_{t+1})$. Ville's maximal inequality for nonnegative supermartingales (Howard et al., 2020) bounds its probability of ever exceeding $1/\zeta$ by ζ. On the complementary event, $\sum_{s\le t}\rho_s \le 2\sum_{s\le t} X_s + 64\log(1/\zeta)$ for every prefix. Apply equation (40). □

The conversion relies on the variational inequality for the constrained optimum. It does not require the oracle response to equal a conditional mean, and it covers zero answer variance, boundary optima and nonunique optima.

D.3 Proof of the response-error bound

Lemma 5.1 (Cumulative response error)**.** *Under the bounded protocol of section 4 and assumption 6.1, use the exact response fits and pooling updates of* Response-Aware UCB *with* $0 < \eta \le 1/2$*. For every* $\delta \in (0,1)$*, with probability at least* $1 - \delta/3$ *over the interaction, simultaneously for all* t*,*

$$\sum_{s\le t}\rho_s \le E_t, \tag{14}$$

where E_t *is the query-count bound in equation (24). For* $t \le T$*, it satisfies* $E_t \le \overline{E}_T$*, with the explicit horizon bound* $\overline{E}_T$ *given in equation (19).*

Proof. For the identity fits, lemmas D.1 and D.2 give the pathwise budget A_t in equation (23). Under assumption 6.1, the fixed comparator attains W^*_{t,I_t} on every queried round. Thus the conditional mean of its normalized loss difference X_t is exactly the gap ρ_t defined in section 5. Apply lemma D.3 with $\zeta = \delta/3$. With probability at least $1 - \delta/3$, simultaneously for all t, this gives $\sum_{s\le t}\rho_s \le 2A_t + 64\log(3/\delta) = E_t$. The equality is the definition in equation (24). Finally, concavity of $\log$ and $\sum_j N_{t,j} \le T$ give $A_t \le K\lambda_I + 8Kp_I\log(1 + T/(Kp_I\lambda_I)) + 4K\log 2/\eta$. Substitution yields $E_t \le \overline{E}_T$ for every $t \le T$. □

D.4 Risk confidence with response-learning bias

We next bound the errors in the two risk estimates. The free labels have centered noise; the queried labels also contain response-learning bias.

Lemma D.4 (Simultaneous risk confidence)**.** *Under the hypotheses and calibration of theorem 6.3, with probability at least* $1 - \delta$*, the response bound* $\sum_{s\le t}\rho_s \le E_t$ *and the free-risk bound below hold simultaneously for every* $t \le T$*. The queried-risk bound also holds for every* $j \in \mathcal{A}_t$ *on each such round:*

$$|\widehat{L}^0_t - L^0_t| \le R^2 B_0 \|x^0_t\|_{(V^0_t)^{-1}}. \tag{43}$$

$$|\widetilde{W}_{t,j} - W^*_{t,j}| \le 8G^2 B_* u_{t,j} \tag{44}$$

Proof. We allocate failure probability $\delta/3$ to each of the response bound, free-risk noise, and queried-risk noise. The weights are fixed before the answer arrives, which preserves the conditional noise bounds.

Regression noise. For the noise terms, we use the following standard consequence of self-normalization (Abbasi-Yadkori et al., 2011). Let $x_s \in \mathbb{R}^{d_V}$ be $\mathcal{F}_s$-measurable, let $\Lambda \in \mathbb{R}^{d_V \times d_V}$ be fixed and positive definite, and let ϵ_s be $\mathcal{F}_{s+1}$-measurable noise satisfying, for some $\sigma \geq 0$,

$$\mathbb{E}[\epsilon_s \mid \mathcal{F}_s] = 0, \qquad \mathbb{E}[e^{\gamma \epsilon_s} \mid \mathcal{F}_s] \leq e^{\gamma^2 \sigma^2 / 2} \quad \text{for every } \gamma \in \mathbb{R}.$$

For $\zeta \in (0, 1)$, with probability at least $1 - \zeta$ simultaneously for all prefixes,

$$\left\| \sum_{s<t} x_s \epsilon_s \right\|_{V_t^{-1}} \leq \sigma \sqrt{\log \frac{\det V_t}{\det \Lambda} + 2 \log(1/\zeta)}, \quad V_t = \Lambda + \sum_{s<t} x_s x_s^\top. \tag{45}$$

The result extends to the fixed positive-definite prior by whitening x_s by $\Lambda^{-1/2}$. The assumptions concern conditional moment-generating functions and predictability; neither independent covariates nor independent rounds are needed. For a variable in an interval of length one, its centered version is conditionally $1/2$-sub-Gaussian by Hoeffding's lemma (Hoeffding, 1963).

Free-risk confidence. Apply equation (45) with failure probability $\delta/3$ to the centered free-risk labels. The uncentered labels lie in $[0, 1]$, and their conditional means are $(x_s^0)^\top \theta_0^*$ by assumption 6.2. The ridge identity, nonexpansiveness of clipping onto $[0, 1]$, and $\|\Lambda_0 \theta_0^*\|_{(V_t^0)^{-1}} \leq \|\theta_0^*\|_{\Lambda_0}$ imply simultaneously equation (43). Here equation (38) bounds the determinant ratio by the expression in equation (26) for all $t \leq T$.

Queried-risk confidence. For the queried labels, put zero covariates and zero noise on free rounds. On a query, write $c_s = c_{s,I_s}$. This feature is predictable, and equation (13) gives

$$Z_s = c_s^\top \theta_* + \rho_s/4 + \epsilon_s, \qquad \mathbb{E}[\epsilon_s \mid \mathcal{F}_s] = 0,$$

where ϵ_s is conditionally $1/2$-sub-Gaussian. Apply equation (45) again, at failure probability $\delta/3$, to $x'_s = \sqrt{\omega_s} c_s$ and $\epsilon'_s = \sqrt{\omega_s} \epsilon_s$ on queries and zero on free rounds. Both ω_s and c_s are $\mathcal{F}_s$-measurable. Since $0 < \omega_s \leq 1$, the transformed noise remains conditionally $1/2$-sub-Gaussian. The resulting martingale sum is $\sum_s \omega_s c_s \epsilon_s$, and the determinant ratio is at most $(1 + T/(d_c \lambda_*))^{d_c}$. Including the prior norm bounds the martingale and regularization terms by B_* in the $(V_t^*)^{-1}$ norm, simultaneously over $t \leq T$.

The remaining term is the accumulated response-learning bias, $b_t = \sum_{s<t: I_s \neq 0} \omega_s c_s \rho_s / 4$. Indeed, the queried ridge error decomposes as

$$\widehat{\theta}_t^* - \theta_* = (V_t^*)^{-1} \left\{ \sum_{s<t: I_s \neq 0} \omega_s c_s \epsilon_s + b_t - \Lambda_* \theta_* \right\}.$$

On the response event of lemma D.3, $\sum_{s<t} \rho_s \leq \overline{E}_T$. The triangle inequality, $\rho_s \geq 0$, and $V_t^* \succeq V_s^*$ for $s < t$ give

$$\begin{aligned} \|b_t\|_{(V_t^*)^{-1}} &\leq \tfrac{1}{4} \sum_{s<t: I_s \neq 0} \omega_s \rho_s \|c_s\|_{(V_t^*)^{-1}} \\ &\leq \tfrac{1}{4} \sum_{s<t: I_s \neq 0} \omega_s \rho_s u_{s,I_s} \leq \tfrac{1}{4} \tau \overline{E}_T = B_*. \end{aligned} \tag{46}$$

The last inequality uses $\omega_s u_{s,I_s} \leq \tau$, including zero width. When there are no past queries, $b_t = 0$. Cauchy–Schwarz and multiplication by $4G^2$ now give equation (44) for every eligible expert and every $t \leq T$ on the common event. The two risk-confidence events and the response conversion event have total failure probability at most δ. A vector-norm confidence event controls every current feature, so no further union bound over experts is needed. The noise inequality is established before intersecting with the response event; its conditional moment-generating-function assumption is not asserted after conditioning on that event. □

D.5 PROOF OF THE ROUTING THEOREM

Theorem 6.3 (Regret of response-aware routing). *Assume the bounded non-interventional protocol of section 4 and assumptions 6.1 and 6.2. Fix $T \geq 1$, $\delta \in (0,1)$, feature maps and positive-definite priors before interaction; public feature vectors may depend on history. Run* RESPONSE-AWARE UCB *with exact constrained response fits, full-history weighted queried-risk estimates, $0 < \eta \leq 1/2$ and equation (21). With probability at least $1-\delta$ over the joint interaction law,*

$$\mathrm{Reg}_T \leq C_{\mathrm{prob}} \left\{ \sqrt{T} \log \frac{T+1}{\delta} + \log \frac{T+1}{\delta} \right\}, \tag{22}$$

where C_{prob} is finite and depends only on the residual bounds, dimensions, registry capacity, regularizers, priors, parameter-norm bounds and pool rate, not on T or δ.

Proposition B.1 (Explicit routing bound). *Under the hypotheses and calibration of theorem 6.3, with probability at least $1-\delta$,*

$$\begin{aligned} \mathrm{Reg}_T \leq\ & R^2 B_0 \mathcal{W}(T, d_0, \lambda_0) \\ & + G^2 \left[16 B_* \sqrt{N_T \mathcal{H}(N_T)} + 4 \overline{E}_T \mathcal{H}(N_T) + E_T \right]. \end{aligned} \tag{32}$$

Proof of theorem 6.3 and proposition B.1. By lemma D.4, the response and risk-estimation bounds hold on a common event with probability at least $1-\delta$. We work on that event and first compare costs on one round, then sum the result using bounds on regression widths.

One-round comparison. For a queried round write $u_t = u_{t,I_t}$, and set $u_t = 0$ on free rounds. Define a lower bound on optimal queried risk and the free-risk estimation error by

$$\underline{W}_{t,j} = \widetilde{W}_{t,j} - 8G^2 B_* u_{t,j}, \qquad \epsilon_t^0 = \widehat{L}_t^0 - L_t^0.$$

With $\alpha = 2B_*$, equation (44) gives $\underline{W}_{t,j} \leq W_{t,j}^*$. The score is $S_{t,j} = \widehat{L}_t^0 - \underline{W}_{t,j} - \beta_{t,j}$. If an oracle action $j^* \neq 0$ queries, then $S_{t,j^*} \geq L_t^0 - C_t^* + \epsilon_t^0$. For the selected queried action j, its actual conditional cost is $W_{t,j}^* + G^2 \rho_t + \beta_{t,j}$ and

$$W_{t,j}^* - \underline{W}_{t,j} \leq 16 G^2 B_* u_{t,j}. \tag{47}$$

We compare the selected action with an oracle action, separating the four possible pairs of free and queried choices. If both actions query, score maximality gives $\underline{W}_{t,j} + \beta_{t,j} \leq \underline{W}_{t,j^*} + \beta_{t,j^*} \leq C_t^*$, and hence

$$\mathbb{E}[C_t \mid \mathcal{F}_t] - C_t^* \leq W_{t,j}^* - \underline{W}_{t,j} + G^2 \rho_t;$$

the free-risk error cancels. If the selected action queries and the oracle is free, $S_{t,j} \geq 0$ gives the same bound plus ϵ_t^0. If the selected action is free and the oracle queries, $0 \geq S_{t,j^*} \geq L_t^0 - C_t^* + \epsilon_t^0$, so the regret is at most $-\epsilon_t^0$. If both are free the regret is zero. Let $\mathbf{1}\{\cdot\}$ denote the indicator of an event. In all cases,

$$\mathbb{E}[C_t \mid \mathcal{F}_t] - C_t^* \leq |\epsilon_t^0| + \mathbf{1}\{I_t \neq 0\} G^2 \{16 B_* u_t + \rho_t\}. \tag{48}$$

This derivation also handles an empty eligible set.

Summing regression widths. To sum the uncertainty terms, we bound cumulative regression widths by the number of observations. Let $x_i \in \mathbb{R}^{d_V}$ be features of norm at most one, where d_V is a positive integer. Let $\Lambda \succeq \lambda \mathrm{Id}_{d_V}$ with $\lambda > 0$ and $V_n = \Lambda + \sum_{i<n} x_i x_i^\top$. Then $v_n = x_n^\top V_n^{-1} x_n \leq \lambda^{-1}$ and $v_n \leq (1+\lambda^{-1}) \log(1+v_n)$, using $\log(1+v) \geq v/(1+v)$ for $v \geq 0$. The determinant lemma and equation (38) give

$$\sum_{n=1}^{N} v_n \leq (1+\lambda^{-1}) d_V \log(1 + N/(d_V \lambda)).$$

Cauchy–Schwarz therefore gives $\sum_{n=1}^{N} \sqrt{v_n} \leq \mathcal{W}(N, d_V, \lambda)$, also valid for $N = 0$. Apply this bound to the free-risk features on all T rounds. The free-risk regression receives an outcome on every

round, including rounds on which an expert is queried. For the queried-risk design, the determinant lemma and $\omega_s u_s^2 \le \lambda_*^{-1}$ give

$$\sum_{s:I_s\neq 0} \omega_s u_s^2 \le (1+\lambda_*^{-1}) \log \frac{\det V_{T+1}^*}{\det \Lambda_*} \le \mathcal{H}(N_T).$$

The whitened update matrix has trace at most N_T/λ_*, which gives the last inequality. On rounds with $\omega_s = 1$, Cauchy–Schwarz bounds the sum of widths by $\sqrt{N_T \mathcal{H}(N_T)}$. On the remaining queried rounds, $\omega_s u_s^2 = \tau u_s$, so

$$\sum_{s:I_s\neq 0} u_s \le \sqrt{N_T \mathcal{H}(N_T)} + \mathcal{H}(N_T)/\tau.$$

This also holds when $N_T = 0$. Sum equation (48), use equation (43), $\sum_t \rho_t \le E_T$, and substitute $\tau = 4B_*/\overline{E}_T$. This proves equation (32) and hence proposition B.1.

Horizon dependence. For $\alpha = 2B_*$ and fixed problem parameters, $E_T \le C_1 \log((T+1)/\delta)$ and $B_v \le C_2\sqrt{\log((T+1)/\delta)}$. The width envelope is at most $C_3\sqrt{T\log(T+1)}$, and the additive response term is at most $C_4 \log((T+1)/\delta)\log(T+1)$ for a fixed finite C_4. Using $\log(T+1) \le \sqrt{T}$ gives equation (22), with constants depending only on the quantities listed in the theorem. This proves theorem 6.3. □

D.6 Approximate response and risk models

The exact models in assumptions 6.1 and 6.2 connect observations across rounds. Here fixed models may approximate the optimal responses and risks; the comparator remains the same conditional oracle. The response proof must account for the fixed comparator's approximation error. The risk proof must also handle signed model errors, separately from the nonnegative bias caused by response learning.

Approximation assumptions. Fix $T \ge 1$, $\delta \in (0,1)$ and deterministic tolerances $\varepsilon_R, \varepsilon_0, \varepsilon_* \ge 0$ before interaction. For each identity, let $\overline{\Theta}_j \in \mathbb{R}^{d_\ell \times (q+1)}$ be a fixed matrix with $\|\overline{\Theta}_j\|_F \le 1$, and write $\overline{a}_{t,j} = \overline{\Theta}_j^\top \ell_{t,j}$. Assume, almost surely for every $t \le T$ and $j \in \mathcal{A}_t$,

$$0 \le \epsilon_{t,j} := \mathbb{E}\Big[\big(e_{t,j}/G - \overline{a}_{t,j}^\top \phi_q(d_{t,j})\big)^2 \mid \mathcal{F}_t\Big] - W_{t,j}^*/G^2 \le \varepsilon_R. \tag{49}$$

Nonnegativity follows from feasibility. The error measures excess risk over the conditional ball optimum, rather than distance between coefficient vectors. This distinction allows boundary optima and singular answer laws. Also assume fixed parameters $\overline{\theta}_0 \in \mathbb{R}^{d_0}$ and $\overline{\theta}_* \in \mathbb{R}^{d_c}$ satisfy $\|\overline{\theta}_0\|_{\Lambda_0} \le Q_0$ and $\|\overline{\theta}_*\|_{\Lambda_*} \le Q_*$, with

$$\big|L_t^0/R^2 - (x_t^0)^\top \overline{\theta}_0\big| \le \varepsilon_0, \qquad \big|W_{t,j}^*/(4G^2) - c_{t,j}^\top \overline{\theta}_*\big| \le \varepsilon_*. \tag{50}$$

The free-risk inequality holds almost surely on every round, including when $\mathcal{A}_t$ is empty; the queried-risk inequality holds almost surely for every eligible expert on every round. These parameters and matrices are fixed before the future observations are realized. Their values need not be known to the learner. Calibration uses the supplied bounds $\varepsilon_R, \varepsilon_*$; the free-risk tolerance ε_0 enters the guarantee but not the calibration.

Let A_t be the pathwise response budget in equation (23), and define its deterministic envelope and the cumulative response approximation error by

$$\begin{aligned} \overline{A}_T &= K\lambda_I + 8Kp_I \log\left(1 + \frac{T}{Kp_I\lambda_I}\right) + \frac{4K\log 2}{\eta}, \\ \mathcal{D}_t &= \sum_{s\le t: I_s\neq 0} \epsilon_{s,I_s}, \qquad p_I = d_\ell(q+1). \end{aligned} \tag{51}$$

The response certificate and its horizon envelope are

$$E_t^{\mathrm{app}} = 2A_t + 3\mathcal{D}_t + 128\log(3/\delta), \qquad \overline{E}_T^{\mathrm{app}} = 2\overline{A}_T + 3T\varepsilon_R + 128\log(3/\delta). \tag{52}$$

Unlike A_t, the quantity $\mathcal{D}_t$ is unobserved. The algorithm uses the deterministic envelope $\overline{E}_T^{\mathrm{app}}$, which is at least E_t^{app} for every $t \le T$. With B_v as in equation (20), set

$$B_*^{\mathrm{app}} = B_* + \varepsilon_* \sqrt{T}, \qquad \alpha = 2B_*^{\mathrm{app}}, \qquad \tau = \frac{4B_*^{\mathrm{app}}}{\overline{E}_T^{\mathrm{app}}}. \tag{53}$$

Proposition D.5 (Routing under approximate models)**.** *Assume the bounded non-interventional protocol of section 4 and the approximation assumptions above, including the fixed feasible matrices, parameter-norm bounds and equations (49) and (50). Choose feature maps and positive-definite priors before interaction, with public feature vectors of norm at most one. Run the* RESPONSE-AWARE UCB *updates with exact constrained response fits, full-history risk regressions,* $0 < \eta \le 1/2$*, and equation (53). With probability at least* $1 - \delta$ *over the joint interaction law,*

$$\begin{aligned} \mathrm{Reg}_T \le{}& R^2 \left[(B_0 + \varepsilon_0 \sqrt{T}) \mathcal{W}(T, d_0, \lambda_0) + T\varepsilon_0 \right] \\ &+ G^2 \left[16 B_*^{\mathrm{app}} \sqrt{N_T \mathcal{H}(N_T)} + 4 \overline{E}_T^{\mathrm{app}} \mathcal{H}(N_T) + E_T^{\mathrm{app}} + 8T\varepsilon_* \right], \end{aligned} \tag{54}$$

with $\mathcal{W}, \mathcal{H}$ *defined in equations (30) and (31) and* Reg_T *measured against the original oracle in equation (7).*

Proof. We first adapt the conditional response certificate, then bound the two risk estimates and repeat the one-round routing comparison.

Response approximation. On a queried round, abbreviate $y_t = e_{t,I_t}/G$, $p_t = a_{t,I_t}^\top \phi_q(d_{t,I_t})$ and $\overline{p}_t = \overline{a}_{t,I_t}^\top \phi_q(d_{t,I_t})$. Let $p_t^* = (a_{t,I_t}^*)^\top \phi_q(d_{t,I_t})$ for a measurable conditional minimizer over $\mathbb{B}_q$. The three coefficients are $\mathcal{F}_t$-measurable, and the predictions and target belong to $[-1, 1]$. Define

$$X_t = (y_t - p_t)^2 - (y_t - \overline{p}_t)^2, \qquad \epsilon_t = \epsilon_{t,I_t}, \qquad \rho_t = \mathbb{E}[(y_t - p_t)^2 \mid \mathcal{F}_t] - W_{t,I_t}^*/G^2,$$

and set all three quantities to zero on free rounds. Both ρ_t and ϵ_t are nonnegative, while $\mathbb{E}[X_t \mid \mathcal{F}_t] = \rho_t - \epsilon_t$ can have either sign. The pathwise regression and pool bounds compare to any fixed feasible matrices, so lemmas D.1 and D.2 still give $\sum_{s \le t} X_s \le A_t$ for every prefix.

Conditional optimality of a_{t,I_t}^* gives the variational inequality

$$\mathbb{E}[(p_t^* - y_t)(p_t - p_t^*) \mid \mathcal{F}_t] \ge 0,$$

and the same inequality with p_t replaced by $\overline{p}_t$. Expanding the two excess risks therefore gives

$$\mathbb{E}[(p_t - p_t^*)^2 \mid \mathcal{F}_t] \le \rho_t, \qquad \mathbb{E}[(\overline{p}_t - p_t^*)^2 \mid \mathcal{F}_t] \le \epsilon_t.$$

Since $|X_t| \le 4$ and $X_t^2 = (p_t - \overline{p}_t)^2 (p_t + \overline{p}_t - 2y_t)^2$,

$$\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \le 16 \mathbb{E}[(p_t - \overline{p}_t)^2 \mid \mathcal{F}_t] \le 32(\rho_t + \epsilon_t). \tag{55}$$

For $\lambda = 1/64$, the exponential Taylor bound used in lemma D.3 now gives

$$\begin{aligned} \mathbb{E}[e^{-\lambda X_t} \mid \mathcal{F}_t] &\le 1 - \lambda(\rho_t - \epsilon_t) + 16\lambda^2 e^{4\lambda} (\rho_t + \epsilon_t) \\ &\le 1 - \lambda \rho_t / 2 + 3\lambda \epsilon_t / 2 \le \exp(-\rho_t/128 + 3\epsilon_t/128). \end{aligned}$$

The middle inequality uses $16\lambda e^{4\lambda} < 1/2$. Thus

$$\exp\left\{ \sum_{s \le t} \rho_s/128 - 3 \sum_{s \le t} \epsilon_s/128 - \sum_{s \le t} X_s/64 \right\}$$

is a nonnegative supermartingale starting at one, with respect to $(\mathcal{F}_{t+1})$. Ville's inequality, at failure probability $\delta/3$, implies simultaneously for every prefix $t \le T$

$$\sum_{s \le t} \rho_s \le 2 \sum_{s \le t} X_s + 3\mathcal{D}_t + 128 \log(3/\delta) \le E_t^{\mathrm{app}}. \tag{56}$$

Risk approximation. Write the two signed model residuals as $r_t^0 = L_t^0/R^2 - (x_t^0)^\top \overline{\theta}_0$ and $r_{t,j}^* = W_{t,j}^*/(4G^2) - c_{t,j}^\top \overline{\theta}_*$. On a query, the normalized label satisfies

$$Z_t = c_{t,I_t}^\top \overline{\theta}_* + r_{t,I_t}^* + \rho_t/4 + \epsilon_t^{\text{noise}}, \qquad \mathbb{E}[\epsilon_t^{\text{noise}} \mid \mathcal{F}_t] = 0.$$

Because $Z_t \in [0,1]$, the centered noise is conditionally $1/2$-sub-Gaussian. Applying equation (45) to the predictably weighted features bounds the noise and regularization terms by B_* in the $(V_t^*)^{-1}$ norm, with failure probability $\delta/3$.

The signed model errors admit a separate bound. Let U have columns $\sqrt{\omega_s} c_{s,I_s}$ over past queries, and let v have entries $\sqrt{\omega_s} r_{s,I_s}^*$. Since $V_t^* = \Lambda_* + UU^\top$, $\|(V_t^*)^{-1/2}U\|_{\text{op}} \le 1$, and hence

$$\left\| \sum_{s<t:I_s\neq 0} \omega_s c_{s,I_s} r_{s,I_s}^* \right\|_{(V_t^*)^{-1}} \le \|v\|_2 \le \varepsilon_* \sqrt{T}. \tag{57}$$

Here $\|\cdot\|_{\text{op}}$ denotes the Euclidean operator norm; its bound follows from $UU^\top \preceq V_t^*$. On the response event, design monotonicity and $\omega_s u_{s,I_s} \le \tau$ also give

$$\left\| \tfrac{1}{4} \sum_{s<t:I_s\neq 0} \omega_s c_{s,I_s} \rho_s \right\|_{(V_t^*)^{-1}} \le \tfrac{\tau}{4} \sum_{s<t} \rho_s \le \tfrac{\tau}{4} \overline{E}_T^{\text{app}} = B_*^{\text{app}}.$$

The noise, regularization and signed model-error terms total at most B_*^{app}, so evaluating the ridge error at a current feature and including its own residual gives

$$|\widetilde{W}_{t,j} - W_{t,j}^*| \le 8G^2 B_*^{\text{app}} u_{t,j} + 4G^2\varepsilon_* \quad (t \le T,\ j \in \mathcal{A}_t). \tag{58}$$

The vector-norm event controls all eligible experts simultaneously.

For the free regression, the same matrix argument with columns x_s^0 and residual vector $(r_s^0)_{s<t}$ gives

$$\left\| \sum_{s<t} x_s^0 r_s^0 \right\|_{(V_t^0)^{-1}} \le \sqrt{\sum_{s<t} (r_s^0)^2} \le \varepsilon_0 \sqrt{T}.$$

Self-normalization at failure probability $\delta/3$, the prior-norm bound and clipping toward the true normalized risk in $[0,1]$ imply

$$|\widehat{L}_t^0 - L_t^0| \le R^2 \left[(B_0 + \varepsilon_0\sqrt{T}) \|x_t^0\|_{(V_t^0)^{-1}} + \varepsilon_0 \right]. \tag{59}$$

The response event and the two noise events hold together with probability at least $1-\delta$. The conditional noise inequalities are established before this intersection, as in the exact-model proof.

Routing comparison. Work on that common event and put $\underline{W}_{t,j} = \widetilde{W}_{t,j} - 8G^2 B_*^{\text{app}} u_{t,j}$. Equation (58) gives $\underline{W}_{t,j} \le W_{t,j}^* + 4G^2\varepsilon_*$ and $W_{t,j}^* - \underline{W}_{t,j} \le 16G^2 B_*^{\text{app}} u_{t,j} + 4G^2\varepsilon_*$. If both learner and oracle query, score maximality compares their lower risk estimates plus fees. Each expert contributes at most $4G^2\varepsilon_*$ of model error, giving $8G^2\varepsilon_*$ in total. If only one queries, the comparison incurs at most $4G^2\varepsilon_*$ and the absolute free-risk error; if both are free, regret is zero. Therefore

$$\begin{aligned} \mathbb{E}[C_t \mid \mathcal{F}_t] - C_t^* \le{}& |\widehat{L}_t^0 - L_t^0| + 8G^2\varepsilon_* \\ &+ \mathbf{1}\{I_t \neq 0\} G^2 \{16 B_*^{\text{app}} u_{t,I_t} + \rho_t\}. \end{aligned} \tag{60}$$

The width summations in the exact-model proof hold for any $\tau > 0$, including equation (53):

$$\sum_{t:I_t\neq 0} u_{t,I_t} \le \sqrt{N_T \mathcal{H}(N_T)} + \mathcal{H}(N_T)/\tau, \qquad \sum_{t=1}^{T} \|x_t^0\|_{(V_t^0)^{-1}} \le \mathcal{W}(T, d_0, \lambda_0).$$

Summing equation (60), using equations (56) and (59), and substituting $\tau = 4B_*^{\text{app}}/\overline{E}_T^{\text{app}}$ proves equation (54). Empty query sequences and empty eligible sets are covered by the same bounds. □

Dependence on approximation error. For fixed problem parameters and $\delta = (T+1)^{-2}$, the right-hand side of equation (54) is at most a constant times

$$\sqrt{T}\log(T+1) + T\varepsilon_R \log(T+1) + T(\varepsilon_0 + \varepsilon_*)\sqrt{\log(T+1)}.$$

The constant depends on the same fixed problem quantities as in theorem 6.3, and not on the tolerances or T. At zero approximation error this recovers the original rate; theorem 6.3 gives the sharper response-concentration constant. For fixed positive errors, the bound's approximation terms are linear in T up to logarithmic factors. The conditional laws may vary subject to the stated approximation bounds. If a known deterministic bound $\overline{\mathcal{D}}_T$ on $\mathcal{D}_T$ is available, it can replace $T\varepsilon_R$ in the horizon envelope. Such a bound must hold almost surely for the analyzed interaction; a bound holding only with high probability requires its failure probability to be included.

E ADDITIONAL EXPERIMENTS

We report (i) learning curves on the four benchmark streams (Appendix E.1); (ii) cost and query rates at six consultation prices (Appendix E.2); and (iii) routing comparisons with a common response learner (Appendix E.3). Appendix F describes the data and expert construction.

Common protocol. We rank methods by mean squared prediction error plus the consultation fees actually paid. A free prediction incurs no fee. All methods receive the same public information, availability schedule, and fee for each expert; only the selected expert's answer is revealed. Reported real-data costs are in squared-temperature units. We select hyperparameters by fee-inclusive cost on validation data and report results on separate test data. This protocol applies to the four benchmarks, every consultation price, and the matched-response comparisons. Hyperparameters remain fixed during test evaluation; the methods continue their prescribed online updates after each round's feedback. The benchmark search budget is 256 configurations per method and dataset.

Repetitions and reporting. Benchmark means and sample standard deviations use five learner seeds on each fixed stream. Deterministic methods have identical replays; we omit zero standard deviations. Bold values indicate the lowest displayed cost within a comparison, including ties. Exact configurations, seeds, and trajectories are retained with the experimental records.

E.1 LEARNING ON THE FOUR BENCHMARK STREAMS

In brief. RESPONSE-AWARE UCB has the lowest final mean total cost among the eight methods on all four streams. The trajectories show how these endpoint comparisons develop as the methods learn from selected answers.

Data. We use a synthetic stream, daily Melbourne temperatures (Hyndman & Yang, 2018), four-hourly Jena observations (Max Planck Institute for Biogeochemistry, n.d.), and daily Delhi observations (sumanthvrao, 2019). The synthetic process alternates between regimes favoring free prediction, direct deferral, linear correction, and nonlinear correction. The real-data splits are chronological: Melbourne uses a 1200-observation fitting prefix that is also replayed for offline warmup; Jena uses 2000 fitting and 2000 warmup observations; Delhi uses 600 fitting and 250 warmup observations. Jena standardization uses its full fitting-and-warmup prefix. No scored observation enters these preprocessing fits.

Experts. Synthetic and Melbourne each have twelve simulated experts in four families, with four experts available per round. Melbourne's private sensors combine real temperatures with controlled bias and noise. Jena has twelve frozen models from random forests, extra trees, gradient boosting, and radial-basis kernel ridge, with four available per round. Delhi has six frozen tree-based models, with four to six available. An internal ridge predictor updates after each revealed real-data outcome. Predetermined availability schedules create absence and return, allowing the response learner to reuse an identity's earlier observations. The seven external baselines deploy purchased answers unchanged; their linear, sampling, neural, discounted, and restart-based implementations follow the methods cited in section 7.

Metric and fees. Every point in figure 2 is the running mean total cost from the first scored round, not a moving-window average. Per-query fees are 0.07, approximately 0.30, approximately 0.17, and 0.02 on synthetic, Melbourne, Jena, and Delhi, respectively. Lines average the five learner seeds; bands show one sample standard deviation.

Results. The final costs are 0.08092, 4.49069, 3.72844, and 2.59240, respectively, compared with 0.08316, 4.61400, 4.02217, and 2.63729 for the best external baseline on each stream. Table 1 gives the complete endpoint comparison. The plots retain the full scored intervals and every baseline.

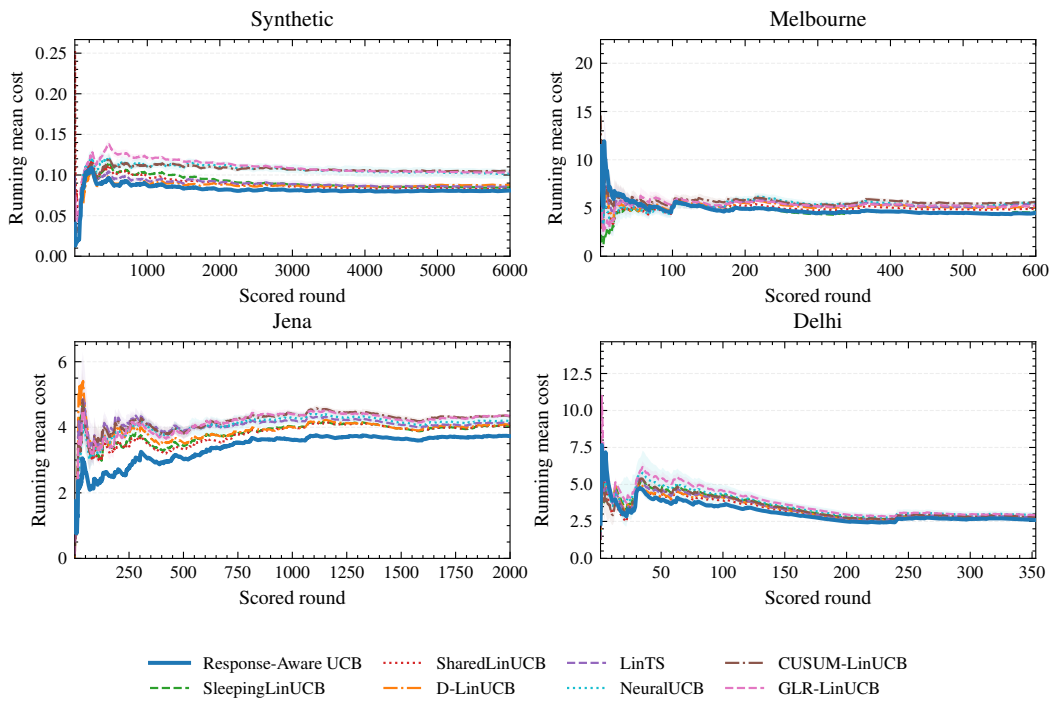

Figure 2: Fee-inclusive test learning curves on all four streams. Every scored round is included, with linear cost axes starting at zero. Running means start at the first scored round. Shading shows one learner-seed standard deviation where nonzero.

E.2 Cost-sensitive consultation

In brief. Across the six price-specific configurations, Response-Aware UCB reduces its query fraction from 96% to 0.1% while retaining lower mean total cost than the external baselines under equal hyperparameter-search budgets.

Data. We use the same Jena validation and test blocks as the benchmark. Every price is evaluated by replaying warmup and scoring, so a new fee can change both the purchased observations and the learned responses.

Experts and comparisons. We retain the twelve frozen Jena experts, their four-family availability schedule, and the same internal predictor. We compare with all seven external baselines. For each price, we select every method's configuration on the validation block and evaluate it on the test block, using equal search budgets for Response-Aware UCB and all seven external baselines.

Metrics and fees. A multiplier m changes the common per-query fee to $\beta(m) \approx 0.17m$; the unrounded base fee is used in every replay. The two metrics are mean total cost and the fraction of scored rounds with a query. Their joint interpretation matters: avoiding a query saves its fee but can increase prediction error. The six multipliers $1, 3, 5, 7.5, 12.5, 50$ span frequent consultation to near-free prediction. Reported standard deviations describe learner-seed variability.

Results at six prices. Table 5 presents the external comparisons, with cost and query percentage in separate panels. Response-Aware UCB has lower cost than every external baseline at each price. Its query rates are 96%, 79.4%, 9.7%, 8.15%, 1.85%, and 0.10%. The base-price configuration in this study is selected separately from the benchmark configuration.

Table 5: Jena test performance against external baselines at six prices. Panels share the same methods and six prices; fees and total costs are in squared-temperature units. Entries show mean and nonzero sample standard deviation. Bold marks minimum displayed total cost in each column; query percentages are not ranked. The configuration-selection protocol is stated above.

(a) Mean total cost

Fee multiplier	×1	×3	×5	×7.5	×12.5	×50
Fee per query	0.17	0.51	0.85	1.28	2.13	8.52
RESPONSE-AWARE UCB	**3.711**	**3.984**	**4.024**	**4.079**	**4.117**	**4.174**
SleepingLinUCB	4.030	4.277	4.240	4.444	4.417	4.190
SharedLinUCB	4.022	4.201	4.180	4.242	4.202	4.182
D-LinUCB	4.078	4.371	4.447	4.452	4.485	4.552
LinTS	4.119 ± 0.049	4.167 ± 0.040	4.177 ± 0.036	4.233 ± 0.044	4.250 ± 0.065	4.182
NeuralUCB	4.185 ± 0.134	4.223 ± 0.108	4.190 ± 0.069	4.169 ± 0.045	4.189 ± 0.020	4.182
CUSUM-LinUCB	4.333 ± 0.039	4.331 ± 0.073	4.328 ± 0.030	4.362 ± 0.050	4.415 ± 0.051	4.666 ± 0.048
GLR-LinUCB	4.349 ± 0.056	4.489 ± 0.038	4.557 ± 0.021	4.574 ± 0.066	4.641 ± 0.049	5.113 ± 0.085

(b) Query percentage

Fee multiplier	×1	×3	×5	×7.5	×12.5	×50
Fee per query	0.17	0.51	0.85	1.28	2.13	8.52
RESPONSE-AWARE UCB	96.00	79.40	9.70	8.15	1.85	0.10
SleepingLinUCB	35.65	34.25	18.00	23.90	12.05	0.10
SharedLinUCB	21.50	14.40	6.10	5.05	0.65	0.00
D-LinUCB	49.45	32.90	29.20	26.75	16.75	2.65
LinTS	22.07 ± 3.04	13.38 ± 4.83	7.10 ± 1.22	6.49 ± 4.13	2.10 ± 1.63	0.00
NeuralUCB	18.07 ± 6.33	7.37 ± 2.00	5.78 ± 1.72	1.90 ± 1.26	0.82 ± 0.80	0.00
CUSUM-LinUCB	39.31 ± 2.51	25.38 ± 1.66	15.65 ± 0.74	10.03 ± 0.36	6.50 ± 0.41	4.82 ± 0.42
GLR-LinUCB	55.35 ± 2.94	47.87 ± 2.46	39.26 ± 3.03	30.60 ± 2.10	18.50 ± 1.29	10.08 ± 0.70

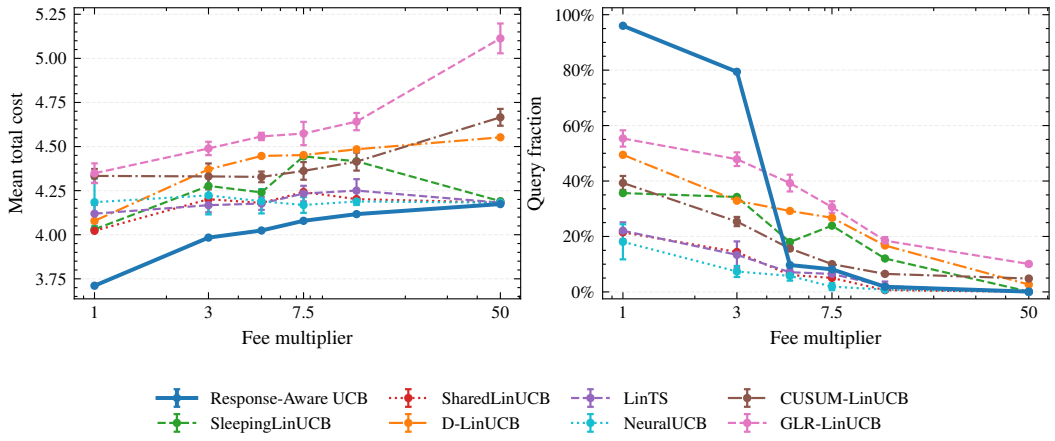

Figure 3: Six-price test comparison with external baselines: total cost and query fraction. The fee axis is logarithmic. Error bars show one sample standard deviation when nonzero; lines join the measured prices without smoothing. The cost axis has a nonzero origin.

E.3 ROUTING WITH THE SAME RESPONSE LEARNER

In brief. With the response architecture held fixed, RESPONSE-AWARE UCB attains the lowest cost on all four streams, tying joint response UCB on three and improving on both UCB alternatives on Delhi.

Data. We use all four benchmark streams with the same fitting, warmup, validation, and test intervals. Public information and preprocessing are unchanged from Appendix E.1.

Experts and routers. Every querying method uses the same experts and cubic correction learner, including shared and identity responses, pooling, and retained memory. Joint response UCB retains our shared and identity risk features but uses ordinary unit-weight ridge regression. Disjoint response UCB instead fits a separate ordinary risk regression for each expert. Both alternatives select their confidence coefficient and risk regularization on validation data using the same 256-configuration budget per dataset; response settings remain fixed. Cyclic response queries the next available identity in registry order; free prediction never queries. Each router learns from its own purchases, so its fitted responses reflect its acquisition decisions.

Metric and fees. We report mean squared error plus the same benchmark consultation fees: 0.07, approximately 0.30, approximately 0.17, and 0.02. The risk-regression target is the loss of the deployed, corrected prediction, not the loss of the expert answer before correction.

Results. Table 6 compares the routers across all four streams. On Delhi, RESPONSE-AWARE UCB costs 2.59240, compared with 2.62820 for joint response UCB and 2.62465 for disjoint response UCB. Joint response UCB matches the benchmark costs on the other three streams. Both cyclic querying and never querying have higher cost than RESPONSE-AWARE UCB on every stream.

Table 6: Matched-response routing: mean total test cost on all four streams. All querying methods can correct the selected expert's answer. Bold values mark the minimum per dataset.

Router	Synthetic	Melbourne	Jena	Delhi
RESPONSE-AWARE UCB	**0.08092**	**4.49069**	**3.72844**	**2.59240**
Joint response UCB	**0.08092**	**4.49069**	**3.72844**	2.62820
Disjoint response UCB	0.08358	4.62449	3.74382	2.62465
Cyclic response	0.15557	5.66927	3.97404	2.84056
Free prediction	0.10949	5.40136	4.18186	2.72163

F DATA AND EXPERT CONSTRUCTION

The benchmarks use two types of expert answers. Synthetic and Melbourne use simulated signals whose bias, noise, and nonlinear distortion we can control. Jena and Delhi use predictions from models fitted to earlier observations and then frozen. This distinction lets us study correction both when the answer mechanism is known and when errors come from trained predictors. Within each dataset, all methods use the same prepared stream and availability schedule. Appendix E.1 gives the evaluation protocol and consultation fees; we describe the inputs and experts here. Full generator and model configurations are retained with the experimental records.

F.1 SHARED PREPROCESSING

Information available to the learner. All routers receive the same public features, internal prediction, availability, and fees. The features describe the current context and, on real data, the internal predictor's past errors. They never contain the current outcome or an unpurchased expert answer. Melbourne and Delhi use only lagged measurements and calendar features. Jena additionally uses weather covariates observed at the target timestamp, before temperature is revealed.

Feature scaling. We standardize continuous covariates using the dataset's preprocessing prefix; calendar sine–cosine terms and the linear trend remain as constructed. Melbourne uses its fitting block, which is also replayed for offline warmup. Jena uses its full fitting-and-warmup prefix, with the transform fixed before replay. No scored observation enters these fits. For the real-data streams, the regression state is

$$x_t^{\mathrm{ctx}} = (p+3)^{-1/2}\big(1, \mathrm{clip}_{[-3,3]}(\widetilde{x}_t)^\top/3, v_{t-1}, \bar{v}_{t-1}\big)^\top,$$

where $\widetilde{x}_t \in \mathbb{R}^p$ contains the preprocessed public features. The two error coordinates are bounded logarithmic transforms of the internal predictor's previous squared error and its exponentially weighted average. Both use only past outcomes. Each coordinate inside the parentheses lies in $[-1,1]$, so $\|x_t^{\mathrm{ctx}}\|_2 \le 1$. In RESPONSE-AWARE UCB, we use $x_t^0 = s_{t,j} = \ell_{t,j} = x_t^{\mathrm{ctx}}$; expert identity enters through the registry block of $c_{t,j}$. The baselines receive the same context vector without that expansion.

Prediction units. On real data, we divide outcomes and predictions by the internal predictor's fitting root mean squared error s, and fees by s^2. We convert reported costs back to squared-temperature units; synthetic data use $s = 1$. The physical residual and disagreement bounds $(R_{\mathrm{phys}}, D_{\mathrm{phys}})$ are $(1.2, 1.3)$, $(15, 20)$, $(15, 25)$, and $(8, 14)$ for synthetic, Melbourne, Jena, and Delhi, respectively. The learner uses $R = R_{\mathrm{phys}}/s$, $D = D_{\mathrm{phys}}/s$, and $G = R + D$. Internal and frozen-model predictions are clipped to $[-50, 60]$ on real data; corrected predictions receive no additional clipping. Bounds are checked without truncating observations.

F.2 SYNTHETIC

The synthetic stream separates four situations: free prediction is adequate, an expert can be used directly, an expert needs rescaling, or an expert needs nonlinear correction. These regimes recur every 240 rounds, with 60 rounds per regime. The public context contains a constant and four sine–cosine pairs of this phase, normalized by $\sqrt{5}$.

The internal prediction h_t is a fixed periodic signal. In regime r, the outcome is $Y_t = h_t + \mu_r(U_t) + \epsilon_t$, where U_t is a symmetric seven-point latent signal on $[-1, 1]$ and ϵ_t is mean-zero noise on $\{-0.2, 0, 0.2\}$. The regime functions are

$$\mu_0(u) = 0, \qquad \mu_1(u) = u, \qquad \mu_2(u) = 0.65u,$$
$$\mu_3(u) = 0.12u + 0.30\tanh(3u) + 0.50(u^2 - \mathbb{E}U_t^2).$$

The last regime includes a hyperbolic-tangent term rather than using a cubic polynomial as its generating function.

Twelve experts form four families with three identities each. Their answers are noisy, biased linear measurements of U_t added to h_t. Three families specialize in the informative regimes; the fourth gives a noisier signal across regimes. Shared and identity-specific noise make expert answers dependent

without making them identical. One identity per family is available at a time, for 150 rounds followed by 300 rounds of absence. The generated outcomes and all potential answers are fixed across learner seeds.

F.3 MELBOURNE

We use daily minimum temperatures from Melbourne (Hyndman & Yang, 2018), with 365 initial observations reserved to form lags. Public features are temperature lags $1, 2, 7, 14, 30, 365$, weekday and annual sine–cosine pairs, the second annual harmonic, and a linear trend. An internal ridge predictor is fitted before the online replay and subsequently updated from revealed outcomes.

The twelve experts are simulated sensors, not trained forecasters. They measure the current temperature through four response types: identity, two linear rescalings, and a saturating hyperbolic-tangent transform. Specifically, the transforms of the clipped standardized residual $u_t = \mathrm{clip}_{[-5,5]}((Y_t - h_t)/s)$ are u_t, $0.65u_t$, $\tanh(1.8u_t)/1.8$, and $0.8u_t$. Sensor reliability varies with season and recent temperature change, measured by the absolute difference between standardized lag-one and lag-seven features relative to its fitting-prefix distribution. Lower reliability attenuates the signal and increases noise. Each family contains three identities with small gain and bias differences; noise has both shared and identity-specific components.

This construction combines real temperature dynamics with controlled measurement errors. Only the purchased sensor answer is revealed. One identity per family is available each day; identities rotate after 25 days, giving each a 50-day absence before its next active period.

F.4 JENA

We sample the Jena climate series (Max Planck Institute for Biogeochemistry, n.d.) every four hours and predict temperature. Its 13 public features are pressure, humidity, wind speed, temperature lags $1, 6, 42, 180$, and hour, weekday, and annual sine–cosine pairs. Pressure, humidity, and wind speed are contemporaneous measurements; all temperature inputs are lagged.

The twelve frozen experts belong to four families: random forests, extra trees, gradient-boosted trees, and RBF kernel ridge. Each family contains three models with different depths or kernel widths, giving distinct predictors without changing the prediction task. These models are fitted on the initial 2000 observations using scikit-learn (Pedregosa et al., 2011). They remain frozen while the common internal ridge predictor updates after each outcome.

One identity per family is available at a time. The active identity changes every 84 rounds, corresponding to 14 days of availability followed by 28 days of absence. The schedule is fixed independently of the router's purchases.

F.5 DELHI

We predict daily mean temperature in the Delhi climate series (sumanthvrao, 2019). The 16 public features are one-day-lagged humidity, wind speed, and pressure; temperature lags $1, 2, 7, 14, 30, 365$; weekday and annual sine–cosine pairs; the second annual harmonic; and a linear trend. Unlike Jena, this stream uses no contemporaneous weather measurements.

Six frozen scikit-learn models (Pedregosa et al., 2011) are fitted on the initial 600 transformed observations: two random forests, two extra-trees models, and two gradient-boosted models. Different tree depths distinguish the two experts within each family. The common internal ridge predictor is initialized on the same prefix and updates after each revealed outcome.

Four to six experts are available per round. Two are always available; the others follow a fixed schedule of absences, with one arriving during warmup. Identities persist across the stream, and the availability schedule is shared by every method.

G Extended related work

Learning to defer. L2D extends selective prediction by replacing rejection with a decision to rely on an external expert. Selective prediction studies the error–rejection tradeoff (Chow, 1970), convex rejection losses (Bartlett & Wegkamp, 2008), and abstention with deep networks (Geifman & El-Yaniv, 2017). Predictor–rejector formulations learn prediction and rejection scores (Cortes et al., 2016; 2024; Ni et al., 2019; Mao et al., 2024a); other formulations accommodate arbitrary task losses (Cao et al., 2022) or score rejection as an additional action (Mao et al., 2024d). Deferral instead accounts for the receiving expert's errors (Madras et al., 2018; Mozannar & Sontag, 2020), with multiple-expert methods comparing input-dependent performance (Verma et al., 2023; Mao et al., 2024b) or selecting a set of experts (Montreuil et al., 2026b). This dependence on expert capabilities motivates work on fairness and accuracy in routing (Keswani et al., 2021), complementary human–AI teams (Hemmer et al., 2022), sample-efficient training with expert predictions (Charusaie et al., 2022), and fine-tuning for individual experts (Raman & Yee, 2021). Joint predictor–router training has been studied through exact optimization for linear models (Mozannar et al., 2023), softmax parameterizations for probability estimation (Cao et al., 2023), and losses that mitigate classifier underfitting (Liu et al., 2024). Two-stage methods instead fit the predictor before learning the router (Mao et al., 2023), while Montreuil et al. (2026a) separate classifier and expert-utility heads during training. The observations available for learning are another distinction: budgeted deferral queries a subset of experts on each labeled example (DeSalvo et al., 2025), and probabilistic deferral handles missing expert annotations while controlling workload allocation (Nguyen et al., 2025). Other methods accommodate experts unseen during training through meta-learning (Tailor et al., 2024) or identity-free deferral (Strong et al., 2026). Online routing with varying experts and selected-action feedback (Duy et al., 2026) and deferral in non-stationary time series (Montreuil et al., 2026f) address changes during deployment. Beyond classification, L2D has been developed for regression (Mao et al., 2024c), multi-task learning (Montreuil et al., 2025b), and adversarial robustness in both two-stage and one-stage systems (Montreuil et al., 2025a; 2026e). Applications include sequential medical decisions (Joshi et al., 2021), guided deferral with language models in healthcare (Strong et al., 2024; 2025), and query allocation in extractive question answering (Montreuil et al., 2026c). Causal approaches study counterfactual second opinions (Corvelo Benz & Gomez Rodriguez, 2022) and the evaluation of deferring systems (Palomba et al., 2025). Expert input can also inform a prediction rather than replace it: Kerrigan et al. (2021) combine human predictions with model probabilities, and Pugnana et al. (2025) learn when to request human feedback for an enriched predictor. Related acquisition methods route to a classifier with additional features (Regol et al., 2025) or supply advice to a selected expert (Montreuil et al., 2026d). Our response instead transforms the purchased answer. We learn both acquisition and responses online, fitting the latter only on purchased answers. The router must therefore distinguish the loss of its current response from the risk achievable with the best feasible response.

Online routing and bandits. The choice of an expert is an action-selection problem with partial feedback. Linear contextual bandits estimate conditional action values and use confidence bounds to balance prediction and exploration (Li et al., 2010; Abbasi-Yadkori et al., 2011); Thompson sampling uses parameter randomization (Agrawal & Goyal, 2013), and NeuralUCB uses neural gradient features for its confidence model (Zhou et al., 2020). Regression-oracle methods provide another route from prediction to contextual-bandit decisions (Foster & Rakhlin, 2020). Sleeping bandits restrict each round's choice to the currently available actions (Kleinberg et al., 2010). These ideas underlie SleepingLinUCB, SharedLinUCB, LinTS, and NeuralUCB in our comparison. We also include mechanisms for changing action values: discounted linear regression in D-LinUCB (Russac et al., 2019), and detector-triggered restarts in CUSUM-LinUCB and GLR-LinUCB, adapted from the non-contextual methods of Liu et al. (2018) and Besson et al. (2022). The seven external baselines follow the same availability and feedback protocol and deploy purchased answers unchanged; separate matched-response comparisons retain our correction learner and vary the router. Our confidence weighting follows He et al. (2022), who study corrupted linear bandits. In our setting, response learning generates the discrepancy from the linear target; bounding this discrepancy lets us calibrate acquisition.

Paid information and aggregation. Learning to buy information is related to, but distinct from, choosing an expert whose answer is the final prediction. Janisch et al. (2019) sequentially acquire feature values before classification, learning a policy from episodes sampled from labeled data.

Paid-observation models also control which advice is obtained (Seldin et al., 2014; Reyzin et al., 2026), with different prediction and feedback structures. Stacked generalization likewise learns how to use other predictors' outputs (Wolpert, 1992; Hastie et al., 2009). In our setting, however, the outputs available for fitting are selected by the router, and the resulting fitting errors affect future purchases. A purchased answer and the revealed outcome suffice to evaluate both response proposals and combine them by exponential weighting (Cesa-Bianchi & Lugosi, 2006).

Polynomial responses. Polynomial features allow nonlinear prediction with coefficients learned by linear methods. In online decision problems, Hu et al. (2020) estimate contextual rewards by local polynomial regression, while Liu et al. (2021) use local Taylor approximations with misspecified linear-bandit algorithms. Smoothness-adaptive contextual bandits address the choice of approximation scale (Gur et al., 2022), and Maran et al. (2024) use Legendre features to represent smooth state-action functions in reinforcement learning. Unlike these context- or action-based approximations, our dictionary transforms a purchased answer. The regret proof uses bounded features and linear coefficients, without smoothness or distributional orthogonality assumptions.