CONSISTENT LEARNING-TO-DEFER WITH EXPERT-CONDITIONAL ADVICE

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce learning-to-defer with advice, a new setting in which a system learns both which expert should handle an input and which information that expert should receive. The policy acts before observing advice and minimizes prediction loss plus expert and acquisition fees. Training may reveal all advice sources or only a subset on each record. We show that composing routing and advice losses can be inconsistent and that augmented training can incur avoidable estimation variance. We propose reference conditional classification, which jointly learns a fixed expert's error probability and each action's error conditional on that expert's success or failure. The reference outcome is observed on every labeled record, so records with missing advice still inform the estimates. We prove a consistency bound for bounded task losses, show that limited training preserves the full population objective, and establish variance reductions for the analyzed fits. On FEVER and Sensitive, full training has the lowest mean dev loss at every evaluated fee, while limited training improves on full training at most matched acquisition budgets.

1 INTRODUCTION

Many decision systems must choose both who should handle an input and what information that expert should receive. A retrieval system can supply relevant passages (Guu et al., 2020; Lewis et al., 2020), an agent can call a tool (Schick et al., 2023), and a fraud detector can request additional transaction attributes. The value of this information depends on the expert using it: a source may correct one expert's mistake while adding little to another's prediction. Acquiring advice also has a cost. That cost matters during training as well as deployment. Obtaining every source for every labeled example may be impractical, leaving the learner to choose an expert and advice from records that reveal only some of the available sources.

We introduce *learning-to-defer with advice* to study these decisions. Standard learning-to-defer (L2D) selects an expert while treating the information available to each expert as fixed (Madras et al., 2018; Mozannar & Sontag, 2020; Montreuil et al., 2026b). Our policy selects an expert and a source using the original input, before observing the advice. It minimizes the expected prediction loss plus the expert and acquisition fees. Because experts can benefit differently from the same source, the optimal expert is the one with the cheapest expected cost under its best source, including no advice. We consider both full training, which observes every source on each record, and limited training, which observes only a subset. In both regimes, the label and the predictions of the experts without advice are available. Their errors therefore remain observed training targets even when a paid source is missing.

Two familiar surrogate constructions raise different concerns in this setting. Predictor–rejector constructions (Cortes et al., 2016; Ni et al., 2019) motivate a product of routing and advice losses. We show that this product can be inconsistent: an unused expensive source can change the surrogate's ranking of experts. An augmented loss over expert–source pairs is Bayes consistent with a fixed cost bound. However, consistency does not determine how efficiently it uses the observed data. Recent work has identified underfitting in augmented L2D objectives (Liu et al., 2024; Montreuil et al., 2026a). We identify an additional statistical limitation. When action error rates are stable within the success and failure groups of an observed reference expert, ordinary augmented fitting can estimate costs with avoidable variance. With complete observations, we quantify the gain from sharing these conditional rates across inputs. With limited observations, a separate result shows that ordinary augmented importance weighting amplifies the variance associated with reference outcomes

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.

by the inverse acquisition probability. Its cost estimate can even fluctuate when the paid action's error is constant. These estimation effects motivate a training objective that uses the reference on every record.

We propose *reference conditional classification*. One fixed expert without advice supplies an always observed reference outcome. A single classification objective jointly learns the reference error probability and each action's error probability conditional on reference success or failure. Averaging the two conditional predictions recovers the expected action loss, to which we add the known fees. We extend the construction to arbitrary bounded task losses and prove a consistency bound without assuming that conditional errors are shared across inputs. When those rates are shared, fitting shared conditional heads pools observations across inputs and attains the full-observation variance reduction. Under limited advice, inverse-probability weighting preserves the full population objective, while every record still trains the reference. Experiments on FEVER retrieval and sensitive feature acquisition compare full training across advice fees and limited training under a fixed number of paid source observations.

Contributions. **(i)** We formulate the joint expert–advice decision and its Bayes rule. **(ii)** We prove inconsistency of product routing and advice losses, and quantify avoidable estimation error in augmented fitting under full and limited observations. **(iii)** We introduce reference conditional classification for bounded losses, with a consistency bound and an unbiased objective under limited advice. **(iv)** We evaluate both observation regimes on FEVER and Sensitive: reference conditional full has the lowest mean dev loss at every evaluated fee, and limited training improves on full training at most matched acquisition budgets.

2 Related Work

Learning-to-Defer (L2D) extends selective prediction and learning with abstention (Chow, 1970; Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2017) by routing inputs to external experts (Madras et al., 2018; Mozannar & Sontag, 2020; Verma et al., 2022). Research on its surrogates addresses consistency, probability calibration, and sample efficiency (Charusaie et al., 2022; Cao et al., 2024; Mao et al., 2024b; Montreuil et al., 2026b). Related predictor–rejector formulations jointly learn prediction and rejection scores (Cortes et al., 2016; Ni et al., 2019); our product construction applies this idea to expert selection and advice acquisition.

Recent work shows that augmented surrogates for L2D can underfit despite their consistency guarantees (Liu et al., 2024; Montreuil et al., 2026a). Montreuil et al. (2026a) separate class-probability and expert-utility estimation; our method estimates each action's error conditional on a shared reference expert's outcome.

Learning to Ask uses human feedback to enrich one predictor (Pugnana et al., 2025), while Regol et al. (2025) route between two classifiers, only one of which receives additional information. Both have a single enriched route. Our setting allows multiple experts and sources, with each source's value depending on the expert receiving it. Appendix F reviews further work on collaboration, deferral algorithms, and applications.

3 Preliminaries

We begin with the standard L2D problem with fixed experts; only the router is learned. An input $X \in \mathcal{X}$ has label $Y \in \mathcal{Y}$, with $(X, Y) \sim \mathcal{D}$. Consider $J \geq 1$ experts $e_j : \mathcal{X} \to \mathcal{Y}$, indexed by $[J] = \{1, \ldots, J\}$, and a task loss $\psi : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. Consulting expert j incurs a known fee $\beta_j \geq 0$ in addition to its prediction loss, giving the total cost $c_j(x, y) \coloneqq \psi(e_j(x), y) + \beta_j \in [0, C_0]$ for some finite $C_0 > 0$.

Learning-to-Defer. Let $s : \mathcal{X} \to \mathbb{R}^J$ be a score vector, with $s_j(x)$ the score for expert j. The router $r_s(x) \in \arg\max_{j \in [J]} s_j(x)$ selects an expert. For any router $r : \mathcal{X} \to [J]$, the deferral loss is the cost of the selected expert:

$$\ell_{\mathrm{def}}(r; x, y) \coloneqq \sum_{j \in [J]} c_j(x, y) \mathbf{1}\{r(x) = j\}. \tag{1}$$

Writing $\bar{c}_j(x) \coloneqq \mathbb{E}[c_j(X,Y) \mid X = x]$ for the conditional cost of expert j, any Bayes-optimal router satisfies $r^\star(x) \in \arg\min_{j\in[J]} \bar{c}_j(x)$ almost surely in X. The learning problem is therefore to recover these Bayes-optimal decisions from data.

Surrogate training and consistency. We train the router by minimizing a surrogate loss on its scores. A standard L2D construction uses cost-sensitive multiclass classification (Mozannar & Sontag, 2020; Mao et al., 2024b). For a score vector z indexed by a finite action set $\mathcal{I}$, let $\phi_{\mathrm{sm}}(z,i) \coloneqq -\log\big(\exp(z_i)/\sum_{a\in\mathcal{I}}\exp(z_a)\big)$ denote the log-softmax loss. The corresponding L2D surrogate is

$$\Phi_{\mathrm{L2D}}(s;x,y) \coloneqq \sum_{j\in[J]} \big(C_0 - c_j(x,y)\big)\phi_{\mathrm{sm}}(s(x),j). \tag{2}$$

Thus, lower-cost experts receive larger weights. Write $\mathcal{E}_{\ell_{\mathrm{def}}}(r) = \mathbb{E}[\ell_{\mathrm{def}}(r;X,Y)]$ and $\mathcal{E}_{\Phi_{\mathrm{L2D}}}(s) = \mathbb{E}[\Phi_{\mathrm{L2D}}(s;X,Y)]$ for the population risks. Their starred versions denote the infima over measurable routers and score functions, respectively.

A surrogate is *Bayes consistent* if every sequence whose surrogate population risk approaches its infimum has decoded decision risks approaching the Bayes risk (Bartlett et al., 2006; Steinwart, 2007). A consistency bound quantifies this implication: it controls the gap to the optimal decision risk by the gap to the optimal surrogate risk. The following bound applies to (2).

Theorem 1 (Consistency of the L2D surrogate)**.** *For every measurable finite-valued score function s with finite surrogate risk,*

$$\mathcal{E}_{\ell_{\mathrm{def}}}(r_s) - \mathcal{E}^\star_{\ell_{\mathrm{def}}} \le \Gamma\left(\mathcal{E}_{\Phi_{\mathrm{L2D}}}(s) - \mathcal{E}^\star_{\Phi_{\mathrm{L2D}}}\right), \tag{3}$$

with $\Gamma(u) = 2\sqrt{C_0 u}$. Consequently, Φ_{L2D} is consistent for the deferral loss.

Since $\Gamma(u) \to 0$ as $u \to 0$, vanishing surrogate excess guarantees vanishing decision excess.

4 Expert-Conditional Advice

4.1 Protocol and Bayes Decision

For each input, the system chooses an expert and an advice source before observing the advice. The selected expert then predicts using the input and, if acquired, that source's contents. Appendix A.1 illustrates the notation and training observations with a fraud-screening example.

Advice and costs. Let $A = (A^1, \ldots, A^K)$ collect the advice variables, with A^k valued in a measurable space $\mathcal{A}_k$. The joint law of (X, A, Y) has (X, Y) marginal $\mathcal{D}$. Source k reveals A^k when selected. We write $[K]_0 = \{0, \ldots, K\}$, where $k = 0$ means that no advice is acquired. For each $k \in [K]_0$, let $\widetilde{A}^{(k)}$ denote the advice vector with every unselected coordinate replaced by a missing symbol $\perp$; $\widetilde{A}^{(0)}$ masks every source. These vectors take values in $\widetilde{\mathcal{A}} = \prod_k(\mathcal{A}_k \cup \{\perp\})$; their formal construction is in Appendix A.2.

We extend each expert to a measurable map $e_j : \mathcal{X} \times \widetilde{\mathcal{A}} \to \mathcal{Y}$ and assume that $e_j(x, \widetilde{a}^{(0)})$ agrees with the no-advice expert in Section 3. Let $B > 0$ bound the task loss. For an expert–source pair (j,k), its task loss and total cost are

$$\begin{aligned} D_{j,k}(x,a,y) &= \psi\big(e_j(x,\widetilde{a}^{(k)}), y\big) \in [0,B], \\ c_{j,k}(x,a,y) &= D_{j,k}(x,a,y) + \beta_j + \gamma_k. \end{aligned} \tag{4}$$

The expert fees β_j and advice fees γ_k are finite and nonnegative, with $\gamma_0 = 0$. Because the task loss is bounded and there are finitely many fees, all expert–source costs lie in $[0, C]$ for some finite $C \ge C_0$. Moreover, $c_{j,0}(x,a,y) = c_j(x,y)$ when no advice is acquired.

Definition 1 (Deferral-advice policy and loss)**.** *A policy consists of measurable maps $r : \mathcal{X} \to [J]$ and $q : \mathcal{X} \times [J] \to [K]_0$, selecting an action from $\Pi = [J] \times [K]_0$. Here $q(x,j)$ is the source that would be used if expert j were selected. Its true loss is*

$$\ell_{\text{def-adv}}(r,q;x,a,y) = \sum_{j=1}^{J}\sum_{k=0}^{K} c_{j,k}(x,a,y)\mathbf{1}\{r(x) = j,\ q(x,j) = k\}. \tag{5}$$

Only the term for $(r(x), q(x, r(x)))$ is active. Equivalently, the executed pair is $\pi(x) = (r(x), q(x, r(x)))$. Write $\mathcal{E}_{\ell_{\text{def-adv}}}(\pi) = \mathbb{E}[c_{\pi(X)}(X, A, Y)]$ for its decision risk and $\mathcal{E}^{\star}_{\ell_{\text{def-adv}}}$ for the infimum over measurable pair policies. Thus q specifies a source for every expert, but the system observes and pays for advice only for the selected pair (Appendix Figure 3). Setting $K = 0$ recovers the deferral loss (1).

The Bayes-optimal policy. The policy chooses before observing any advice, so it compares expected costs conditional on the input. For each expert, it selects the source that minimizes expected task loss plus the source fee, including the option of no advice. It then compares these values across experts after adding each expert's consultation fee.

Lemma 2 (Bayes-optimal deferral-advice policy)**.** *One measurable Bayes-optimal policy is obtained by choosing, almost surely in X,*

$$\begin{aligned} q^{\star}(x, j) &\in \arg\min_{k \in [K]_0} \mathbb{E}[c_{j,k}(X, A, Y) \mid X = x], \\ r^{\star}(x) &\in \arg\min_{j \in [J]} \min_{k \in [K]_0} \mathbb{E}[c_{j,k}(X, A, Y) \mid X = x]. \end{aligned} \tag{6}$$

An expert with higher no-advice cost can therefore become optimal if advice reduces its expected task loss enough to cover the source fee and close the gap to the other experts. The proof is in Appendix A.5.

Since $k = 0$ is always available, advice cannot increase the L2D Bayes risk (Appendix A.7). For a given expert, a source is strictly preferable to no advice exactly when its expected reduction in task loss exceeds its fee (Lemma A.1).

What is observed during training. Every training record contains the original input and its label, so all fixed experts can be evaluated without advice. Full training also acquires every source and observes every expert–source loss. Limited training acquires only a subset of sources per record. Once a source is acquired, it can be supplied to each fixed expert, revealing the corresponding block of losses. Missing sources therefore leave blocks of action outcomes unobserved. We specify the acquisition probabilities in Subsection 4.4; deployment always selects a single expert–source pair before observing its advice.

4.2 Limitations of standard surrogate constructions

Prediction with rejection suggests composing losses for two decisions (Cortes et al., 2016; Ni et al., 2019). L2D suggests an augmented loss over expert–source pairs (Mozannar & Sontag, 2020). The product can fail to recover the Bayes decision. The augmented surrogate is Bayes consistent over unrestricted scores, but ordinary augmented cost estimates can amplify noise from missing advice.

4.2.1 A product surrogate for routing and advice

A natural construction uses one score to choose the expert and one advice score for each expert. Replacing the two indicators in (5) with smooth losses gives a product of routing and advice terms. Two experts and one source suffice to show inconsistency.

Definition 2 (Product router–query surrogate)**.** *Let $J = 2$ and $K = 1$. Choose either logistic factors $\phi_0(t) = \log(1 + e^{-t})$, $\phi_1(t) = \log(1 + e^{t})$ or exponential factors $\phi_0(t) = e^{-t}$, $\phi_1(t) = e^{t}$. The router score s_r selects expert 2 when $s_r(x) \geq 0$; expert j's advice score s_{q_j} selects the source when $s_{q_j}(x) \geq 0$. Define*

$$\Phi_{\text{prod}}(s_r, \mathbf{s}_q; x, a, y) = \sum_{j=1}^{2} \sum_{k=0}^{1} c_{j,k}(x, a, y)\, \phi_{j-1}(s_r(x)) \phi_k(s_{q_j}(x)), \qquad \mathbf{s}_q = (s_{q_1}, s_{q_2}). \tag{7}$$

Consider a deterministic problem with one input and freely varying scores. Write $c_{j,k}$ for each entry of its cost table. For an expert with no-advice and advice costs (u, v), the best achievable advice loss is the *profiled loss*

$$F(u, v) = \inf_{t \in \mathbb{R}} [u\phi_0(t) + v\phi_1(t)].$$

After profiling out the advice scores, the router minimizes $F(c_{1,0}, c_{1,1})\phi_0(t) + F(c_{2,0}, c_{2,1})\phi_1(t)$ over its score t. With logistic or exponential routing losses, this selects the expert with the smaller profiled loss. The Bayes rule instead compares the smallest cost available for each expert. For example, exponential losses give $F(u, v) = 2\sqrt{uv}$. Cost rows $(0.10, 0.90)$ and $(0.20, 0.20)$ have Bayes costs 0.10 and 0.20, but profiled losses 0.60 and 0.40. An exponential router therefore prefers expert 2. The first expert's query selects no advice, yet its unused expensive source raises its profiled loss.

Theorem 3 (Inconsistency of product routing and advice losses)**.** *For either choice of loss in Definition 2, the product surrogate is not Bayes consistent: for every $c_0 \in (0, C)$ there are $\delta \in (0, C - c_0)$, $\kappa > 0$, and a deterministic one-input problem with task-loss bound $B = C$, zero fees, and cost rows*

$$u_1 = (c_0, C), \qquad u_2 = (c_0 + \delta, c_0 + \delta), \tag{8}$$

such that expert 1 is Bayes optimal, while every score triple with surrogate excess below κ routes to expert 2 and incurs decision excess δ.

Appendix B.1 proves the result for a broader family of product losses and characterizes the row summaries that preserve Bayes routing on every cost table (Proposition B.5).

4.2.2 Augmented training with missing advice

Augmented training scores every expert–source pair using cost-sensitive cross-entropy (Mozannar & Sontag, 2020; Mao et al., 2024b).

Definition 3 (Augmented surrogate)**.** *For the pair-score vector $s(x) = (s_{j,k}(x))_{(j,k)\in\Pi}$ and the fixed cost bound C from (4), define*

$$\Phi^C_{\mathrm{aug}}(s; x, a, y) = \sum_{j,k} [C - c_{j,k}(x, a, y)]\phi_{\mathrm{sm}}(s(x), (j, k)). \tag{9}$$

The decoder chooses a pair with the largest score.

Cheaper actions receive larger weights. This loss is Bayes consistent over unrestricted scores. With known positive acquisition probabilities, inverse-probability weighting (Horvitz & Thompson, 1952) gives an unbiased loss when acquisition is independent of current labels and source outcomes conditional on the input and history. Consistency does not ensure efficient estimation. Recent work identifies underfitting in augmented L2D objectives (Liu et al., 2024; Montreuil et al., 2026a). We study an additional issue: ordinary augmented fitting can leave useful observations unused.

Fix a paid pair (j, k) at one input. Let $E, Z \in \{0, 1\}$ be the errors of an always observed no-advice expert and the paid pair, respectively. Let I indicate acquisition of source k, independently of (E, Z), with $\rho = \Pr(I = 1) > 0$. Consider n independent, identically distributed triplets (E_t, Z_t, I_t), with pair cost $c_t = Z_t + \beta_j + \gamma_k$. With freely varying scores, augmented training ranks pairs by their average weights. Importance weighting gives the implied cost estimate

$$\widehat{c}_{\mathrm{aug}} = C - \frac{1}{n}\sum_{t=1}^{n} \frac{I_t}{\rho}(C - c_t).$$

For example, if $\rho = 1/2$, an acquired record contributes $2(C - c_t)$ to the sum, while an unacquired record contributes zero.

Theorem 4 (Acquisition noise in augmented training)**.** *Under the sampling assumptions above, fix $C > 1 + \beta_j + \gamma_k$. For every $n \geq 1$, the estimate $\widehat{c}_{\mathrm{aug}}$ is unbiased for $\mathbb{E}[c_t]$ and satisfies*

$$n \operatorname{Var}(\widehat{c}_{\mathrm{aug}}) = \frac{\operatorname{Var}(Z)}{\rho} + \left(\frac{1}{\rho} - 1\right)\big(C - \mathbb{E}[c_t]\big)^2. \tag{10}$$

If $\rho < 1$, the second term is strictly positive even when the action error Z is constant.

The proof is in Appendix C.2. Acquiring one of K sources uniformly gives $1/\rho = K$: the action-error contribution to the estimator's variance grows by this factor. By the law of total variance, $\operatorname{Var}(Z)$ includes $\operatorname{Var}(\mathbb{E}[Z \mid E])$, the variation in action error explained by reference success or failure. The estimator amplifies this component even though E is observed on every record. Even when every

acquired record shows the same action error, the augmented estimate still fluctuates with the number of acquisitions. Appendix E.4 empirically verifies this variance decomposition (Figure 7). The next subsection uses the reference outcomes observed on every record to construct a classification loss for the joint decision.

4.3 Reference-Conditional Classification

The preceding subsection showed that augmented fitting can use paid observations inefficiently when a no-advice expert's outcome is known on every labeled record. We use that expert as a reference. In classification, we estimate an action's error separately when the reference succeeds and fails, then average the two estimates to obtain its task loss. The decoder adds the known fees. The training loss is binary cross-entropy, while the task may be multiclass classification or bounded regression. An *action* is an expert–source pair, including no advice.

4.3.1 Learning errors relative to a reference

Suppose an action has error probability 0.1 when the reference succeeds and 0.6 when it fails. If the reference fails with probability 0.2, the action's error probability is $0.8(0.1) + 0.2(0.6) = 0.2$. We learn these three probabilities jointly.

Index the $N = J(K+1)$ expert–source pairs by $i \in [N]$. We identify each index with its pair (j,k) when the expert and source matter. Let $i_0 = (j_0, 0)$ be the reference action. For now the task loss is 0–1 loss, so $Z_i = \mathbf{1}\{e_j(X, \widetilde{A}^{(k)}) \neq Y\}$ records whether action i makes an error. For $i \neq i_0$, the target probabilities are

$$\begin{aligned} p^\star(x) &= \Pr(Z_{i_0} = 1 \mid X = x), \\ a_i^\star(x) &= \Pr(Z_i = 1 \mid Z_{i_0} = 0, X = x), \\ b_i^\star(x) &= \Pr(Z_i = 1 \mid Z_{i_0} = 1, X = x), \\ \Pr(Z_i = 1 \mid X = x) &= (1 - p^\star(x))a_i^\star(x) + p^\star(x)b_i^\star(x). \end{aligned} \tag{11}$$

The final identity is the law of total probability; it requires no error independence. Conditional rates on zero-probability reference outcomes are arbitrary. Any no-advice expert can serve as the fixed reference; the experiments choose one using dev data.

4.3.2 Training the probabilities and choosing an action

Let $g(x)$ be the reference logit and $h_{i,0}(x), h_{i,1}(x)$ the conditional logits for action i. They use only x and may share a representation. For a logit z and target $t \in [0,1]$, write $\phi_{\log}(z,t) = \log(1+e^z) - tz$ for binary cross-entropy. On a labeled example, g predicts Z_{i_0}. The observed reference error Z_{i_0} selects the conditional head trained against Z_i, giving the action term $\phi_{\log}(h_{i,Z_{i_0}}, Z_i)$. We fit all logits jointly. Properness of binary cross-entropy (Gneiting & Raftery, 2007) gives the population targets in (11); Appendix D.1 states the full binary objective.

At deployment, the reference's correctness is unknown, so we average both conditional predictions. With $\sigma(v) = (1+e^{-v})^{-1}$, let $\widehat{p} = \sigma(g)$, $\widehat{a}_i = \sigma(h_{i,0})$, and $\widehat{b}_i = \sigma(h_{i,1})$. The predicted error is $\widehat{v}_{i_0} = \widehat{p}$ for the reference and $\widehat{v}_i = (1-\widehat{p})\widehat{a}_i + \widehat{p}\widehat{b}_i$ for another action. For a pair $i = (j,k)$, the decision rule adds its known fees: $\widehat{\pi}(x) \in \arg\min_{(j,k)\in\Pi}\{B\widehat{v}_{j,k}(x) + \beta_j + \gamma_k\}$, with $B = 1$ for classification. For bounded losses below, B restores the task-loss units. The selected source is then acquired for the selected expert. The fixed fees enter only this comparison.

With complete observations, shared conditional heads can pool information across inputs. For finitely many input values, when an action's true conditional error rates are constant across inputs, fitting two shared conditional probabilities and a reference probability per input attains the variance reduction in Theorem C.1 (Proposition C.3). The consistency bound below also allows input-dependent conditional rates.

4.3.3 Extending the construction to bounded task losses

For a task loss in $[0, B]$ with $B > 0$, write $D_i = D_{j,k}(X, A, Y)$ for action $i = (j,k)$ as in (4), and set $d_i = D_i/B \in [0,1]$. We use this normalized loss as a fractional Bernoulli target, aiming to reconstruct $\mathbb{E}[d_i \mid X = x]$.

The binary loss uses the joint outcome of the reference and the action. For fractional losses, imagine drawing one number uniformly from $[0, 1]$ and setting each auxiliary outcome to one when that number is below its normalized loss. Using the same draw for both outcomes gives the target weights

$$\begin{array}{c|cc} & \widetilde{Z}_i = 0 & \widetilde{Z}_i = 1 \\ \hline \widetilde{Z}_{i_0} = 0 & t^i_{00} = 1 - \max(d_{i_0}, d_i) & t^i_{01} = (d_i - d_{i_0})_+ \\ \widetilde{Z}_{i_0} = 1 & t^i_{10} = (d_{i_0} - d_i)_+ & t^i_{11} = \min(d_{i_0}, d_i). \end{array} \tag{12}$$

Here $(u)_+ = \max\{u, 0\}$, and the entries are the probabilities of $(\widetilde{Z}_{i_0}, \widetilde{Z}_i)$ given the realized losses. The last row and column sum to d_{i_0} and d_i; for binary losses it recovers the observed error pair. Among tables with these marginals, it minimizes disagreement, assigning weight $|d_i - d_{i_0}|$ to unequal outcomes (Levin & Peres, 2017, Proposition 4.7; Appendix D.3). We average binary cross-entropy under these weights, without sampling auxiliary outcomes. The shared draw ensures that equal losses give identical auxiliary outcomes. If the action loss exceeds the reference loss, their outcomes disagree only when the draw falls between the two normalized losses. Thus the conditional targets reflect the size and direction of the loss difference.

Definition 4 (Full reference-conditional surrogate)**.** *For normalized task losses $d_i = D_i/B$, target weights from (12), and reference weight $\alpha > 0$, define*

$$\Phi_{\mathrm{RC},\alpha} = \alpha\phi_{\log}(g, d_{i_0}) + \frac{1}{N}\sum_{i \neq i_0} \phi_i, \tag{13}$$

$$\phi_i = t^i_{00}\phi_{\log}(h_{i,0}, 0) + t^i_{01}\phi_{\log}(h_{i,0}, 1) + t^i_{10}\phi_{\log}(h_{i,1}, 0) + t^i_{11}\phi_{\log}(h_{i,1}, 1).$$

All logits are functions of X and are optimized together. The fees enter only the decision rule above.

When i corresponds to (j, k), we also write $\phi_{j,k}$ for ϕ_i. The first two terms of ϕ_i train $h_{i,0}$, and the last two train $h_{i,1}$. The population targets $p^\star, a^\star_i, b^\star_i$ are defined as in (11), with Z replaced by $\widetilde{Z}$. The weights use only D_{i_0} and D_i. The parameter α sets the weight of the shared reference term.

4.3.4 GUARANTEES

The table's action marginal gives $(1 - p^\star(x))a^\star_i(x) + p^\star(x)b^\star_i(x) = \mathbb{E}[D_i \mid X = x]/B$ for $i \neq i_0$, while $p^\star(x) = \mathbb{E}[D_{i_0} \mid X = x]/B$. At these population targets, the decoder compares each pair's conditional expected cost directly, avoiding the profiled row losses that cause the product failure. The next result bounds decision error away from the targets.

Theorem 5 (Consistency of reference-conditional classification)**.** *Let $N \geq 2$, let $\alpha > 0$, let every task loss lie in $[0, B]$ with $B > 0$, and fix the nonnegative expert and advice fees. Let $\mathcal{E}_{\Phi_{\mathrm{RC},\alpha}} = \mathbb{E}[\Phi_{\mathrm{RC},\alpha}]$. Its infimum $\mathcal{E}^\star_{\Phi_{\mathrm{RC},\alpha}}$ is over all measurable finite-valued logits. Every measurable finite-valued logit model with finite surrogate risk satisfies*

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\widehat{\pi}) - \mathcal{E}^\star_{\ell_{\text{def-adv}}} \leq B\sqrt{\max\{N, 2/\alpha\}\left(\mathcal{E}_{\Phi_{\mathrm{RC},\alpha}} - \mathcal{E}^\star_{\Phi_{\mathrm{RC},\alpha}}\right)}. \tag{14}$$

The proof is in Appendix D.2. Excess cross-entropy controls the reconstructed loss of each action. When the decoder selects the wrong pair, its regret is bounded by the estimation errors of that pair and the Bayes pair. Since fees are known exactly, these errors arise only from predicted task losses. The proof also handles boundary probabilities for which finite logits cannot attain the infimum.

The reference probability enters every action estimate, so a small α weakens the bound. For $\alpha \geq 2/N$, its coefficient is $B\sqrt{N}$. We select α on the dev set to account for finite-sample fit.

Acquisition variance. Under the sampling assumptions of Theorem 4, suppose both reference outcomes have positive probability. Joint empirical minimization over three free probabilities estimates the reference error from all records and the two conditional errors from the acquired reference groups (Proposition C.3). With the data law and acquisition probability fixed as $n \to \infty$, the reconstructed cost's $\sqrt{n}$-scaled estimation error has limiting variance

$$\mathrm{Var}(\mathbb{E}[Z \mid E]) + \frac{1}{\rho}\mathbb{E}[\mathrm{Var}(Z \mid E)].$$

The component explained by the reference is estimated from all records; only variation within the reference groups is amplified by $1/\rho$. The acquisition-count term in Theorem 4 is also removed (Proposition C.2).

4.4 Learning from Limited Advice

The full objective requires every source on every training example. With a fixed acquisition budget, we can instead observe a subset on each example and train the available action terms. We call these regimes *reference conditional full* and *reference conditional limited*, respectively.

An acquired source supplies losses for all J fixed experts. Every no-advice loss, including the reference loss D_{i_0}, is always available. Because $\phi_{j,k}$ uses only D_{i_0} and $D_{j,k}$, we can train it when source k is acquired. Each term depends on only one paid source, so changing the dependence between sources preserves the population objective when all (X, Y, A^k) marginals are fixed. Deployment still acquires one selected source.

Compensating for missing sources. If one of four sources is acquired uniformly, multiplying its observed block by four recovers that block's full-observation contribution in expectation. The always observed terms need no correction.

Let H be the acquisition history and I_k indicate acquisition of source k. Assume the known inclusion probability $\rho_k(X, H) = \Pr(I_k = 1 \mid X, H) > 0$ and that I_k is independent of the current (Y, A) conditional on (X, H).

Definition 5 (Limited reference-conditional surrogate)**.** *The loss observed under limited advice is*

$$\widetilde{\Phi}_{\mathrm{RC},\alpha} = \alpha\phi_{\log}(g, d_{i_0}) + \frac{1}{N}\left[\sum_{j:(j,0)\neq i_0} \phi_{j,0} + \sum_{k=1}^{K} \frac{I_k}{\rho_k(X, H)} \sum_{j=1}^{J} \phi_{j,k}\right]. \tag{15}$$

Inverse-probability weighting (Horvitz & Thompson, 1952) gives a rarely acquired source a larger weight. Acquisition can depend on the input and past records, but cannot depend on the current label or source outcomes. Uniformly selecting one source gives $\rho_k = 1/K$.

Proposition 6 (Limited advice targets the full classification risk)**.** *Suppose each source has known positive inclusion probability $\rho_k(X, H)$ and, conditional on (X, H), its acquisition indicator is independent of the current label and source outcomes. For every fixed finite logit model,*

$$\mathbb{E}[\widetilde{\Phi}_{\mathrm{RC},\alpha} \mid X, A, Y, H] = \Phi_{\mathrm{RC},\alpha}. \tag{16}$$

Whenever differentiation and conditional expectation may be interchanged, the same identity holds for the gradient. Consequently the population risk of (15) *obeys the consistency bound in Theorem 5.*

The proof is in Appendix D.4.

Only the paid conditional terms receive inverse-probability weights; every record trains the reference directly. This is the fitting rule behind the acquisition variance reduction above. Figure 2 evaluates how the full and limited methods use an equal budget of paid source observations.

5 Experiments

We evaluate all-source training across advice fees (Appendices E.2.1–E.2.2), nine matched acquisition budgets per task (Appendices E.3.1–E.3.2), and 30 variance settings with informative and uninformative references (Appendix E.4). Additional analyses cover expert-specific advice effects (Figure 4), advice use (Figure 5), cost components (Table 5), and paired-seed differences (Figure 6).

Setting. On FEVER (Thorne et al., 2018), Qwen3-4B-Instruct, Qwen3-8B-Instruct (Yang et al., 2025), and DeBERTa-v3-large (He et al., 2021) classify claims. The four paid sources provide BM25 (Robertson & Zaragoza, 2009) top-1, top-3, top-5, or reformulated top-5 passages. Sensitive feature acquisition (IEEE Computational Intelligence Society, 2019) uses balanced fraud/non-fraud splits and three experts: logistic regression, a gradient-boosted tree, and an MLP. Four nested sources reveal contact, payment, device, and identity features. Each policy chooses an expert and a source, including no advice, from the original input. Expert fees β_j stay fixed, while source k costs $\gamma_k = \lambda\gamma_k^{\mathrm{base}}$, with zero fee for no advice. We report dev prediction error plus fees. All methods use the same backbone architecture and hyperparameter-search budget for each task.

All-source training. With every source observed during training, reference conditional full has the lowest mean dev loss at every evaluated advice fee on both tasks (Figure 1). The comparison includes the augmented surrogate in Definition 3 and standard Learning-to-Defer, which has no advice action.

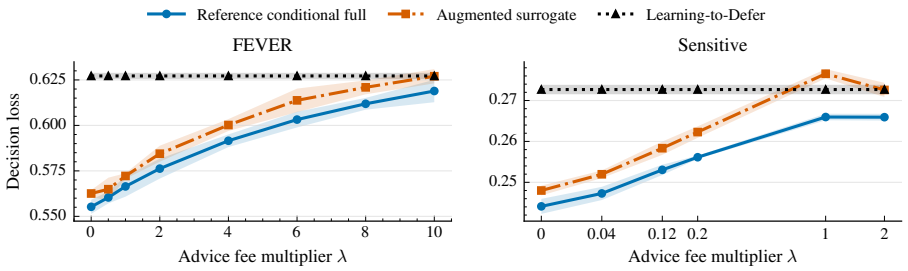

Figure 1: **All-source training.** Mean dev decision loss; bands show one sample standard deviation over four seeds. Sensitive shows $\lambda \leq 2$ on a $\log(1 + \lambda/0.04)$ scale (full sweep: Appendix Table 4). Lower is better.

Appendix Table 3 reports a mean of 0.576 for reference conditional full versus 0.584 for augmented training at $\lambda = 2$ on FEVER; Table 4 reports 0.253 versus 0.258 at $\lambda = 0.12$ on Sensitive. Figure 5 shows advice use, and Table 5 separates prediction error from fees.

Equal paid source observations. For Q paid source observations, full and augmented training observe all four sources on $Q/4$ records. Below the complete-data budget, limited training spreads the same Q observations across more records, giving the reference more labeled inputs. All methods receive the same number of optimizer updates (Appendix Table 6). At the largest budget, full and limited training use all sources and the same surrogate form. Their fitted models can still differ because their hyperparameters are selected separately (Appendix E.3).

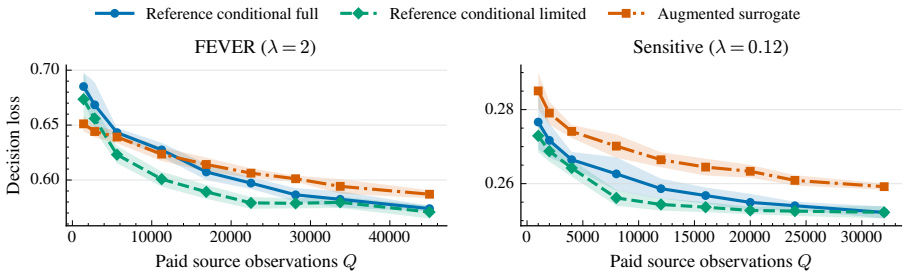

Figure 2: **Equal paid source observations.** Mean dev decision loss over eight seeds; bands show one sample standard deviation. FEVER uses $\lambda = 2$ and Sensitive uses $\lambda = 0.12$. Lower is better.

Appendix Tables 7 and 8 show lower mean loss for limited than full training at all nine FEVER budgets and the first eight Sensitive budgets. At $Q = 8{,}000$ on Sensitive, the means are 0.256 and 0.263; at the complete-table budget, both round to 0.252. Augmented training leads at the two smallest FEVER budgets; limited training leads at the other seven. Figure 6 gives paired seed differences.

Controlled variance experiment. With an informative reference, RC reduces estimation variance by 47.5% at acquisition probability $1/4$ relative to an acquired-only average, matching the predicted 48% reduction. An uninformative-reference control shows no comparable gain (Appendix E.4).

6 Conclusion

Reference conditional classification estimates each expert–source pair's expected task loss and adds the known fees when choosing a pair. It avoids the product surrogate's inconsistency, has a consistency bound for bounded task losses, and removes avoidable acquisition variance in the analyzed model. Limited training preserves the full population objective under the stated acquisition assumptions. On FEVER and Sensitive, full training leads the all-source comparisons, while limited training improves on full training at most matched acquisition budgets.

AI USE STATEMENT

We used LLM-based assistants to check mathematical arguments and draft proofs, implement and debug methods and experimental code. Proof of Theorem C.1 was AI assisted. For the controlled variance study in Appendix E.4, assistance included specifying the simulation and its parameters, writing code to generate synthetic data from the stated probability model. The simulation code was checked with unit tests and numerical consistency checks. The authors take full responsibility for the accuracy, originality, and attribution of all content, including AI-assisted text, proofs, code, and results.

REFERENCES

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 02 2006. doi: 10.1198/016214505000000907.

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, June 2008.

Nina L Corvelo Benz and Manuel Gomez Rodriguez. Counterfactual inference of second opinions. In *Uncertainty in Artificial Intelligence*, pp. 453–463. PMLR, 2022.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

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, January 1970. doi: 10.1109/TIT.1970.1054406.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles (eds.), *Algorithmic Learning Theory*, pp. 67–82, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

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.

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

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

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 Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2478–2484. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/344. URL https://doi.org/10.24963/ijcai.2022/344. Main Track.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

IEEE Computational Intelligence Society. Ieee-cis fraud detection. https://kaggle.com/competitions/ieee-fraud-detection, 2019. Kaggle competition.

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 M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4421–4434. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/234b941e88b755b7a72a1c1dd5022f30-Paper.pdf.

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*, AIES '21, pp. 154–165, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462516. URL https://doi.org/10.1145/3461702.3462516.

Nilanjana Laha, Aaron Sonabend-W, Rajarshi Mukherjee, and Tianxi Cai. Finding the optimal dynamic treatment regimes using smooth fisher consistent surrogate loss. *The Annals of Statistics*, 52(2):679–707, 2024. doi: 10.1214/24-AOS2363. URL https://arxiv.org/abs/2111.02826.

Nilanjana Laha, Nilson Chapagain, Victoria Cicherski, and Aaron Sonabend-W. On fisher consistency of surrogate losses for optimal dynamic treatment regimes with multiple categorical treatments per stage, 2025. URL https://arxiv.org/abs/2505.17285.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2 edition, 2017. doi: 10.1090/mbk/107. With contributions by Elizabeth L. Wilmer.

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

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. PMLR, 2024.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=GIlsH0T4b2.

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. PMLR, 2024a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *ISAIM*, 2024b.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024c.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4753–4761. PMLR, 02–04 May 2024d. URL https://proceedings.mlr.press/v238/mao24a.html.

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.

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.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Beyond augmented-action surrogates for multi-expert learning-to-defer, 2026a. URL https://arxiv.org/abs/2604.09414.

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, Hoang Duy Dang, Maxime Meyer, Lai Xing Ng, Axel Carlier, and Wei Tsang Ooi. Online learning-to-defer with varying experts. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026c. URL https://openreview.net/forum?id=1lix8ppUJ7.

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*, 2026d. URL https://openreview.net/forum?id=kEVupwepTq.

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. *arXiv preprint arXiv:2601.22538*, 2026f.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

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://api.semanticscholar.org/CorpusID:255941521.

Cuong C Nguyen, Thanh-Toan Do, and Gustavo Carneiro. Probabilistic learning to defer: Handling missing expert annotations and controlling workload distribution. In *The Thirteenth International Conference on Learning Representations*, 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://openreview.net/forum?id=mkkFubLdNW.

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.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009. doi: 10.1561/1500000019.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007. URL https://api.semanticscholar.org/CorpusID:16660598.

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.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3475–3483. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/tailor24a.html.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819, 2018.

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*, 2022. URL https://api.semanticscholar.org/CorpusID:253237048.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

APPENDIX

A Protocol and Bayes Decision **16**

A.1 A fraud-screening example . . . 16
A.2 Advice masking . . . 16
A.3 Notation used throughout . . . 17
A.4 Training and execution . . . 18
A.5 The Bayes comparison . . . 18
A.6 When advice is worth acquiring . . . 18
A.7 Comparison with no-advice deferral . . . 19
A.8 Sequential and pair policies . . . 19

B Inconsistency of Product Surrogates **19**

B.1 Proof of Theorem 3 . . . 22

C Fixed-Bound Augmented Training and Reference Pooling **23**

C.1 Complete observations: proof of Theorem C.1 . . . 24
C.2 Limited observations: proof of Theorem 4 . . . 25
C.3 Joint fitting: proof of Proposition C.3 . . . 26
C.4 A local consequence for choosing an action . . . 27

D Reference-Conditional Classification **28**

D.1 Training and deployment . . . 28
D.2 Proof of Theorem 5 . . . 29
D.3 Fractional targets for bounded losses . . . 30
D.4 Learning from observed source blocks . . . 31

E Experiments **31**

E.1 Common protocol . . . 31
E.2 All-source experiments . . . 33
E.3 Training under equal acquisition budgets . . . 34
E.4 Controlled acquisition-variance experiment . . . 37
E.5 Reproducibility . . . 40

F Additional Related Work **40**

Appendix A defines the protocol and proves the Bayes decision rule. Appendix B explains the product surrogate's inconsistency, and Appendix C derives the variance comparisons for augmented training. Appendix D gives the reference-conditional algorithm and its consistency proofs. Appendix E contains the experimental protocols and complete results.

A Protocol and Bayes Decision

A.1 A fraud-screening example

Consider a bank deciding whether a new transaction is fraudulent. The input x contains its amount, merchant, and time; the label $Y \in \{0, 1\}$ indicates fraud. There are $J = 2$ fixed prediction models: a transaction model e_1 and an account-security model e_2. Both can predict from x alone or use additional information. The bank can request $K = 2$ paid sources: the account's recent transactions A^1 and a device-activity report A^2. Thus $k = 0$ means no request, $k = 1$ requests transaction history, and $k = 2$ requests device information. This is an illustrative scenario, separate from the experimental protocol.

Choosing an expert and its advice. For a transaction from an unfamiliar merchant, history may help e_1 recognize a normal purchase, while device information may help e_2 detect account misuse. A policy could therefore specify $q(x, 1) = 1$ and $q(x, 2) = 2$. If the router chooses $r(x) = 2$, the executed pair is $\pi(x) = (2, 2)$. The bank requests only the device report, pays $\beta_2 + \gamma_2$, and obtains the prediction $e_2(x, \widetilde{a}^{(2)})$, where $\widetilde{a}^{(2)} = (\perp, a^2)$ masks the unrequested history. With 0–1 task loss, the realized cost is $c_{2,2}(x, a, y) = \mathbf{1}\{e_2(x, \widetilde{a}^{(2)}) \neq y\} + \beta_2 + \gamma_2$. The choice must be made before the report arrives. The Bayes rule therefore compares $\bar{c}_{j,k}(x) = \mathbb{E}[c_{j,k}(X, A, Y) \mid X = x]$ for all six pairs, including both no-advice options. For a given expert, better expected accuracy justifies a paid source only if the improvement covers its fee.

Learning from labeled transactions. Once the fraud label is known, both models can be evaluated without advice. Requesting one source also allows both models to be evaluated with that source. The observed task losses on a labeled record are therefore:

Training regime	Sources requested	Observed task losses, for $j \in \{1, 2\}$
Full	History and device report	$D_{j,0}, D_{j,1}, D_{j,2}$
Limited	History only	$D_{j,0}, D_{j,1}$
Limited	Device report only	$D_{j,0}, D_{j,2}$

For example, choosing one source uniformly, independently of the record's label and source contents, gives $\rho_1 = \rho_2 = 1/2$. The limited surrogate weights the acquired source's terms by two; every record contributes the no-advice terms.

Using the reference. Choose the transaction model without advice as $i_0 = (1, 0)$. Its error Z_{i_0} is known on every labeled record. For the device-assisted action $i = (2, 2)$, reference-conditional training learns its error probability when e_1 is correct, $a_i^\star(x)$, and when e_1 is wrong, $b_i^\star(x)$. These two targets use records with device reports; the reference target $p^\star(x)$ uses every record. All probabilities are learned jointly. For a new transaction, the label and reference error are unknown, so the model combines its predictions as $\widehat{v}_i(x) = (1 - \widehat{p}(x))\widehat{a}_i(x) + \widehat{p}(x)\widehat{b}_i(x)$. It adds $\beta_2 + \gamma_2$ to this predicted error and compares the result with the estimated costs of the other five pairs.

A.2 Advice masking

Each advice variable A^k takes values in a measurable space $\mathcal{A}_k$. Choose $\perp \notin \bigcup_k \mathcal{A}_k$ and give $\mathcal{A}_k \cup \{\perp\}$ its disjoint-union sigma-algebra. Define $\mathcal{A} = \prod_{k=1}^K \mathcal{A}_k$ and $\widetilde{\mathcal{A}} = \prod_{k=1}^K (\mathcal{A}_k \cup \{\perp\})$, each with its product sigma-algebra. The map m_k retains coordinate k and replaces every other coordinate by $\perp$; m_0 replaces them all. Coordinate projections and constant maps are measurable, so each m_k is measurable. Write $\widetilde{A}^{(k)} = m_k(A)$. For $K = 0$, the empty product is a singleton and only the no-advice action is available.

A.3 NOTATION USED THROUGHOUT

Table 1 records the symbols shared by the main text and proofs. Symbols introduced inside a single proof are local to that proof.

Table 1: Notation shared across the paper. A star marks a population target or an optimal value; a hat marks a fitted estimate or decoded decision.

Symbol	Meaning
$X, Y, A = (A^1, \dots, A^K)$	Input, label, and advice sources.
$J, K, [K]_0, \Pi$	Numbers of experts and paid sources; $[K]_0 = \{0, \dots, K\}$ includes no advice; $\Pi = [J] \times [K]_0$.
$e_j, \widetilde{A}^{(k)}, \psi$	Expert j, advice vector revealing only source k, and task loss.
B, C_0, C	Bounds on task loss, no-advice total cost, and all pair costs, respectively.
β_j, γ_k	Expert and source fees.
$D_{j,k}, c_{j,k}, \bar{c}_{j,k}(x)$	Task loss, total cost, and conditional expected total cost of a pair; subscript i abbreviates (j, k).
r, q, π, r_s	Router, expert-conditional source rule, pair policy, and router decoded from scores s.
$s(x), s_j(x)$	Score vector and its coordinate for expert j; pair scores are indexed by (j, k).
$\ell_{\mathrm{def}}, \ell_{\mathrm{def\text{-}adv}}$	Realized losses of the no-advice and expert–advice decisions.
$\phi_{\mathrm{sm}}, \phi_0, \phi_1$	Log-softmax loss and the chosen logistic or exponential product-loss factors.
$\phi_{q,k}, \phi_{r,j}$	Query and routing loss factors in the general product family (Appendix B).
$F(u, v), \nu$	Profiled query loss and the cost transform used in the general product family.
$\phi_{\log}, \phi_{\mathrm{Ber}}$	Binary cross-entropy in logit and probability form, respectively.
$\Phi_{\mathrm{L2D}}, \Phi_{\mathrm{prod}}$	Standard L2D and product router–query surrogates.
Φ^C_{aug}	Augmented surrogate with a fixed cost bound.
$\Phi_{\mathrm{RC},\alpha}, \widetilde{\Phi}_{\mathrm{RC},\alpha}$	Full and limited reference-conditional surrogates.
$\mathcal{E}_\ell, \mathcal{E}_\Phi$	Population risks of a true loss and a surrogate.
$N, i, i_0, Z_i, \widetilde{Z}_i$	Number and index of pairs, fixed no-advice reference, observed classification error, and auxiliary binary target for bounded losses.
$p^\star, a_i^\star, b_i^\star$	Population reference target probability and conditional action target probabilities; error probabilities under 0–1 loss.
d_i, t^i_{uv}, α	Normalized task loss, joint binary target weights, and reference-term weight.
$g, h_{i,0}, h_{i,1}$	Reference logit and conditional logits for action i.
$\widehat{p}, \widehat{a}_i, \widehat{b}_i, \widehat{v}_i$	Fitted reference and conditional probabilities, and reconstructed normalized task loss.
$I_k, \rho_k(X, H), H$	Source-acquisition indicator, its positive inclusion probability, and prior acquisition history.
$\widehat{c}_{\mathrm{aug}}, \widehat{c}_{\mathrm{RC}}$	Estimated mean action costs in the limited-observation variance comparison.
$\Delta(x), \widehat{\Delta}_{\mathrm{aug}}(x), \widehat{\Delta}_{\mathrm{RC}}(x)$	True and estimated cost differences between an action and the reference (Appendix C).
Q, λ	Paid source observation budget and advice-fee multiplier in the experiments.

For a loss L and a compatible policy or score model u, we write $\mathcal{E}_L(u) = \mathbb{E}[L(u; X, A, Y)]$; a star on the risk denotes its infimum over the class specified in the corresponding result. In particular, $\mathcal{E}^\star_{\Phi_{\mathrm{RC},\alpha}}$ is the infimum over measurable finite logits. The reference-conditional surrogate estimates task losses; its decoded decision risk also includes the fixed fees.

A.4 Training and execution

Training uses the label, every no-advice expert outcome, and the outcomes under each acquired source. All reference-conditional heads are optimized together and may share a representation (Algorithm 1). At deployment, the model uses only x to estimate task costs, adds the known fees, and selects one pair. Figure 3 shows the order of these decisions and observations. The reference loss is used to form the training targets; it is unknown at deployment.

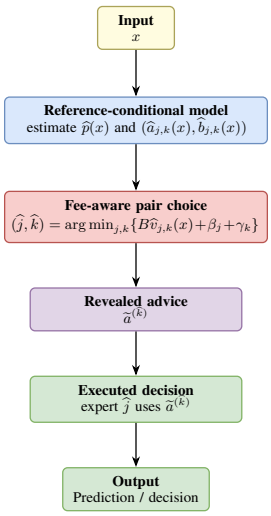

Figure 3: The deferral-advice protocol for the proposed decoder. The model uses only x to estimate every pair's marginal task cost. It adds the known fees, selects one pair, and only then reveals the selected source to the selected expert. The no-advice choice $\widehat{k} = 0$ reveals $\widetilde{a}^{(0)}$.

A.5 The Bayes comparison

Lemma 2 (Bayes-optimal deferral-advice policy)**.** *One measurable Bayes-optimal policy is obtained by choosing, almost surely in X,*

$$\begin{aligned} q^\star(x, j) &\in \arg\min_{k\in[K]_0} \mathbb{E}[c_{j,k}(X, A, Y) \mid X = x], \\ r^\star(x) &\in \arg\min_{j\in[J]} \min_{k\in[K]_0} \mathbb{E}[c_{j,k}(X, A, Y) \mid X = x]. \end{aligned} \tag{6}$$

Proof. Because the protocol pays only for the executed pair, the Bayes rule compares the conditional expected costs of the available pairs. Write $\bar{c}_{j,k}(x) = \mathbb{E}[c_{j,k}(X, A, Y) \mid X = x]$. The finitely many bounded costs admit measurable conditional expectations. For every policy, conditioning on X fixes the executed pair, so

$$\mathcal{E}_{\ell_{\text{def-adv}}}(r, q) = \mathbb{E}[\bar{c}_{r(X),q(X,r(X))}(X)] \geq \mathbb{E} \min_{j,k} \bar{c}_{j,k}(X).$$

A finite minimum is measurable, and its finite argmin admits a measurable selector. Minimizing first over k for each expert and then over j attains this lower bound and gives exactly (6). In particular,

$$\mathcal{E}^\star_{\ell_{\text{def-adv}}} = \mathbb{E} \min_{j,k} \bar{c}_{j,k}(X).$$

□

A.6 When advice is worth acquiring

Lemma A.1 (Advice acquisition condition)**.** *For almost every x, a source $k \in [K]$ is strictly preferable to no advice for expert j if and only if*

$$\mathbb{E}[D_{j,0}(X, A, Y) \mid X = x] - \mathbb{E}[D_{j,k}(X, A, Y) \mid X = x] > \gamma_k. \tag{17}$$

Proof. For a fixed expert, β_j cancels from the comparison between source k and no advice, and $\gamma_0 = 0$. Their expected cost difference is $\gamma_k - \mathbb{E}[D_{j,0} - D_{j,k} \mid X = x]$, which is negative exactly when (17) holds. □

A.7 Comparison with no-advice deferral

Lemma A.2 (Bayes risk with and without advice)**.** *The Bayes risks satisfy*

$$\mathcal{E}^\star_{\ell_{\text{def-adv}}} \le \mathcal{E}^\star_{\ell_{\text{def}}}. \tag{18}$$

Proof. Every no-advice action remains available. Consequently, almost surely,

$$\min_{j,k} \bar{c}_{j,k}(X) \le \min_{j} \bar{c}_{j,0}(X).$$

Taking expectations proves the claim. Since the difference between the right and left sides is nonnegative and bounded, equality holds exactly when the two conditional minima agree almost surely. Advice strictly improves Bayes risk precisely when it lowers the best available conditional cost on a set of positive probability; improving an expert that is never selected need not suffice. □

A.8 Sequential and pair policies

Proposition A.3 (Sequential–composite equivalence)**.** *For every sequential policy (r, q), the pair policy $\pi(x) = (r(x), q(x, r(x)))$ incurs the same realized loss as (r, q) under $\ell_{\text{def-adv}}$. Conversely, every measurable pair policy $\pi : \mathcal{X} \to \Pi$ can be implemented by a sequential policy with the same realized loss. Consequently, the sequential and composite formulations induce the same population risks and therefore share the same Bayes-optimal decisions.*

Proof. The realized loss depends only on the selected expert and its advice. We construct a policy in each representation that selects the same pair.

Let (r, q) be any measurable sequential policy and define the measurable pair policy $\pi(x) = (r(x), q(x, r(x)))$. Equation (5) gives

$$\ell_{\text{def-adv}}(r, q;\, x, a, y) = c_{r(x),\, q(x,r(x))}(x, a, y) = c_{\pi_1(x),\, \pi_2(x)}(x, a, y) = \ell_{\text{def-adv}}(\pi;\, x, a, y),$$

where π_1, π_2 are the expert and source coordinates. Conversely, for any measurable pair policy π, set $r(x) = \pi_1(x)$ and $q(x, j) = \pi_2(x)$ for every j. These maps are measurable and select exactly $\pi(x)$, so the same loss identity holds. Taking expectations gives the same achievable risks, infima, and Bayes-optimal decisions. □

B Inconsistency of Product Surrogates

Product surrogates: assumptions and supporting lemmas

After query optimization, the router compares the profiled losses $F(c_{j,0}, c_{j,1})$. We show that these values increase with either cost entry, then construct a table where their ranking reverses the ranking of the row minima. We begin with the logistic example and state the conditions under which the argument extends to the full product family.

For the logistic construction below, use the factors $\phi_0(t) = \log(1 + e^{-t})$ and $\phi_1(t) = \log(1 + e^t)$ from Definition 2.

We work in the binary setting $J = 2$, $K = 1$ introduced in Subsection 4.2.1. To extend the product loss in (7), allow a cost transform ν, query losses $\phi_{q,0}, \phi_{q,1}$, and routing losses $\phi_{r,1}, \phi_{r,2}$:

$$\Phi^\nu_{\text{prod}} = \sum_{j=1}^{2} \sum_{k=0}^{1} \nu(c_{j,k}) \phi_{r,j}(s_r) \phi_{q,k}(s_{q_j}).$$

The main-text loss Φ_{prod} is the special case $\nu(c) = c$, $\phi_{r,j} = \phi_{j-1}$, and $\phi_{q,k} = \phi_k$.

Let $s_r : \mathcal{X} \to \mathbb{R}$ be a binary router score and, for each expert $j \in \{1, 2\}$, let $s_{q_j} : \mathcal{X} \to \mathbb{R}$ be a binary query score. Write $\mathbf{s}_q = (s_{q_1}, s_{q_2})$. As in the main text, the decoded decisions are

$$r(x) = 1 + \mathbf{1}\{s_r(x) \geq 0\}, \qquad q(x, j) = \mathbf{1}\{s_{q_j}(x) \geq 0\}.$$

At a fixed realization, write $\Phi^{\nu}_{\mathrm{prod}}(t_r, t_1, t_2; x, a, y)$ for this general product loss evaluated at the three scalar score values.

Regularity conditions. The argument allows monotone cost transforms, positive continuous query loss factors, and different routing loss factors. We impose the following conditions; the query objectives may be nonconvex and may have several minimizers.

(N1) *Cost transform:* $\nu : [0, C] \to [0, \infty)$ is continuous and strictly increasing.

(Q1) *Query loss factors:* $\phi_{q,0}, \phi_{q,1} : \mathbb{R} \to (0, \infty)$ are continuous and strictly positive. Neither differentiability nor monotonicity is required.

(Q2) *Well-posedness:* for every $(u, v) \in (0, C]^2$, $t \mapsto \nu(u)\, \phi_{q,0}(t) + \nu(v)\, \phi_{q,1}(t)$ is coercive. No convexity or uniqueness of the query minimizer is required.

(P1) *Router weights:* $\phi_{r,1}, \phi_{r,2} \in C^2(\mathbb{R})$ with $\phi_{r,j} > 0$, $\phi'_{r,1} < 0$, $\phi'_{r,2} > 0$, $\phi''_{r,j} > 0$.

(P2) *Routing ratio:* $\chi(t) \coloneqq -\phi'_{r,1}(t)/\phi'_{r,2}(t)$ is a strictly decreasing bijection $\mathbb{R} \to (0, +\infty)$ with $\chi(0) = 1$.

Examples covered by the conditions. The logistic construction takes $\nu(c) = c$, $\phi_{q,k} = \phi_k$, and logistic router weights. It satisfies the conditions for strictly positive costs, as used in the counterexample. For logistic routing factors, $\chi(t) = e^{-t}$; for exponential routing factors, $\chi(t) = e^{-2t}$. Both satisfy (P1)–(P2).

The query factors may also be multiplied by a positive function or evaluated at a transformed score. For any continuous $G : \mathbb{R} \to (0, \infty)$ and $U : \mathbb{R} \to \mathbb{R}$, set $\phi_{q,k}(t) = G(t)\phi_k(U(t))$ whenever (Q2) holds. The proof needs no monotonicity or invertibility of U. For exponential loss factors $\phi_{q,0}(t) = e^{-t}$ and $\phi_{q,1}(t) = e^{t}$, a positive-cost query objective tends to infinity as $t \to -\infty$ through its first term and as $t \to +\infty$ through its second. Thus it is coercive. The same conclusion holds for logistic loss factors, since $\log(1 + e^{-t}) \to \infty$ as $t \to -\infty$ and $\log(1 + e^{t}) \to \infty$ as $t \to +\infty$.

The profiled loss of an expert. For a cost row (u, v), let $F(u, v)$ be the smallest attainable query loss. The next three lemmas establish its existence, its role in routing, and its strict dependence on both row entries.

Definition B.1 (Profiled query loss)**.** *For $(u, v) \in [0, C]^2$, define*

$$F(u, v) \;\coloneqq\; \inf_{t \in \mathbb{R}} \big[\nu(u)\, \phi_{q,0}(t) + \nu(v)\, \phi_{q,1}(t)\big]. \tag{19}$$

Write $\phi_q(t; u, v) = \nu(u)\phi_{q,0}(t) + \nu(v)\phi_{q,1}(t)$ for the query objective whose minimum defines $F(u, v)$.

Existence of query minimizers

Lemma B.1 (Existence of a query minimizer)**.** *Under (N1)–(Q2), for every $(u, v) \in (0, C]^2$, the infimum in (19) is attained, and its minimizer set is nonempty and compact. Write $t^\star(u, v)$ for any such minimizer.*

Proof. Fix $(u, v) \in (0, C]^2$. By (Q1)–(Q2), $\phi_q(\cdot; u, v)$ is continuous and coercive on $\mathbb{R}$. Its nonempty sublevel set at $\phi_q(0; u, v)$ is compact, so the minimum is attained. The minimizer set is a closed subset of that compact set. □

Profiling the query losses

Lemma B.2 (Routing objective after query minimization)**.** *Under (N1)–(Q2) and (P1), for every $t_r \in \mathbb{R}$ and every realization (x, a, y) whose four costs lie in $(0, C]$,*

$$\begin{aligned} \inf_{t_1, t_2 \in \mathbb{R}} \Phi^{\nu}_{\mathrm{prod}}(t_r, t_1, t_2; x, a, y) = \phi_{r,1}(t_r)\, F(c_{1,0}(x, a, y), c_{1,1}(x, a, y)) \\ + \phi_{r,2}(t_r)\, F(c_{2,0}(x, a, y), c_{2,1}(x, a, y)). \end{aligned} \tag{20}$$

The profiled query minimizers are $t_j^\star = t^\star(c_{j,0}(x,a,y), c_{j,1}(x,a,y))$, hence they are independent of the router score t_r.

Proof. Fix (t_r, x, a, y) and suppress the data arguments of the costs. The product loss is $\sum_{j=1}^{2} \phi_{r,j}(t_r)\phi_q(t_j; c_{j,0}, c_{j,1})$. The query scores occur in separate summands, each multiplied by a positive constant. Thus

$$\inf_{t_1,t_2} \sum_{j=1}^{2} \phi_{r,j}(t_r)\phi_q(t_j; c_{j,0}, c_{j,1}) = \sum_{j=1}^{2} \phi_{r,j}(t_r) \inf_{t_j} \phi_q(t_j; c_{j,0}, c_{j,1}).$$

The inner infima are the profiled losses in (19), proving (20). By Lemma B.1, each is attained. Multiplication by a positive routing factor leaves its minimizing scores unchanged. □

Continuity and dependence on both costs

Lemma B.3 (Continuity and strict monotonicity of the profiled loss)**.** *Under (N1)–(Q2), the map $F : (0, C]^2 \to (0, \infty)$ is continuous and strictly increasing in each coordinate on $(0, C]^2$.*

Proof. Write $\phi_q(t; u, v) = \nu(u)\phi_{q,0}(t) + \nu(v)\phi_{q,1}(t)$. Fix $(u_0, v_0) \in (0, C]^2$. Since both coefficients at this point are positive and continuous, in a compact relative neighborhood of (u_0, v_0), $\phi_q(t; u, v) \geq \frac{1}{2}\phi_q(t; u_0, v_0)$ for every t. The values $\phi_q(0; u, v)$ are bounded on a sufficiently small neighborhood. Coercivity of $\phi_q(\cdot; u_0, v_0)$ therefore places all nearby minimizers in one compact interval. Joint continuity of $\phi_q(t; u, v)$ on this interval implies continuity of its minimum: uniform continuity gives $|F(u, v) - F(u_0, v_0)| \leq \sup_t |\phi_q(t; u, v) - \phi_q(t; u_0, v_0)|$ there. The attained minimum is positive because all its factors are positive.

For $0 < v_1 < v_2 \leq C$, choose a minimizer t_2 for (u, v_2). Then

$$F(u, v_2) = \phi_q(t_2; u, v_1) + (\nu(v_2) - \nu(v_1))\phi_{q,1}(t_2) > F(u, v_1).$$

The same argument using $\phi_{q,0}$ proves strict increase in the first coordinate. This argument needs neither differentiability nor a unique minimizer. □

Increasing either cost raises the profiled loss, including when the other cost remains the row minimum. The next lemma shows how the ordering of these losses determines the selected expert.

The resulting router decision

Lemma B.4 (Router minimizer)**.** *Under (P1)–(P2), for any $w_1, w_2 > 0$ the function $t \mapsto w_1\phi_{r,1}(t) + w_2\phi_{r,2}(t)$ has a unique minimizer $t_r^\star(w_1, w_2)$ satisfying*

$$w_1 > w_2 \implies t_r^\star > 0, \qquad w_1 = w_2 \implies t_r^\star = 0, \qquad w_1 < w_2 \implies t_r^\star < 0.$$

When the profiled losses differ, the router selects the expert with the smaller value.

Proof. Strict convexity follows from (P1), since

$$\frac{d^2}{dt^2}\left[w_1\phi_{r,1}(t) + w_2\phi_{r,2}(t)\right] = w_1\phi''_{r,1}(t) + w_2\phi''_{r,2}(t) > 0.$$

Hence there is at most one minimizer. Differentiating the objective and setting the derivative to zero yields

$$w_1\phi'_{r,1}(t_r^\star) + w_2\phi'_{r,2}(t_r^\star) = 0,$$

or equivalently

$$-\frac{\phi'_{r,1}(t_r^\star)}{\phi'_{r,2}(t_r^\star)} = \frac{w_2}{w_1}.$$

By (P2), the left-hand side is the strictly decreasing bijection $\chi : \mathbb{R} \to (0, \infty)$. Since $w_2/w_1 > 0$, the stationary equation has a unique solution. Strict convexity makes it the global minimizer. Since $\chi(0) = 1$, we obtain

$$\frac{w_2}{w_1} < 1 \iff t_r^\star(w_1, w_2) > 0,$$
$$\frac{w_2}{w_1} = 1 \iff t_r^\star(w_1, w_2) = 0,$$
$$\frac{w_2}{w_1} > 1 \iff t_r^\star(w_1, w_2) < 0.$$

The sign of the minimizing score gives the stated router decision. □

COST SUMMARIES THAT PRESERVE THE BEST EXPERT

Proposition B.5 (Cost summaries that preserve Bayes routing)**.** *Let $m \geq 1$ be an integer and $G : (0, C]^m \to \mathbb{R}$ be continuous. Ranking every finite collection of cost rows by G selects only experts with minimum row cost if and only if $G(u) = h(\min_k u_k)$ for a continuous strictly increasing function h.*

Proof. Suppose G always selects only rows whose minimum entry is smallest. For any rows u, v with $\min_k u_k < \min_k v_k$, applying this property to the two-row collection forces $G(u) < G(v)$; equality would leave the second row among the minimizers. Define $h(a) = G(a\mathbf{1})$ for $a \in (0, C]$. Then h is continuous and strictly increasing. If $a = \min_k u_k < C$, then for every sufficiently small $\epsilon > 0$,

$$h(a - \epsilon) < G(u) < h(a + \epsilon).$$

Continuity and $\epsilon \downarrow 0$ give $G(u) = h(a)$. If $a = C$, then $u = C\mathbf{1}$ and the same identity holds directly. Conversely, a profile of the form $h(\min_k u_k)$ with strictly increasing h ranks rows exactly by their minimum entry. □

B.1 PROOF OF THEOREM 3

Theorem 3 (Inconsistency of product routing and advice losses)**.** *For either choice of loss in Definition 2, the product surrogate is not Bayes consistent: for every $c_0 \in (0, C)$ there are $\delta \in (0, C - c_0)$, $\kappa > 0$, and a deterministic one-input problem with task-loss bound $B = C$, zero fees, and cost rows*

$$u_1 = (c_0, C), \qquad u_2 = (c_0 + \delta, c_0 + \delta), \tag{8}$$

such that expert 1 is Bayes optimal, while every score triple with surrogate excess below κ routes to expert 2 and incurs decision excess δ.

Proof. **Reversing the ordering of the two experts.** The regularity lemmas above show that the product profile F is continuous and strictly increasing in each coordinate. Holding one coordinate at the unique row minimum and changing another coordinate changes F, so it cannot equal $h(\min_k u_k)$. By Proposition B.5, it cannot preserve Bayes routing on every cost table. We construct the table in the theorem and show that the wrong decision persists for all scores sufficiently close to the surrogate infimum.

Fix $c_0 \in (0, C)$. Lemma B.3 gives $F(c_0, C) > F(c_0, c_0)$. Continuity then gives a $\delta \in (0, C - c_0)$ such that

$$F(c_0, C) > F(c_0 + \delta, c_0 + \delta). \tag{21}$$

For the rows in (8), the Bayes costs are c_0 and $c_0 + \delta$. Bayes therefore selects expert 1.

By Lemma B.2, minimizing over the two query scores reduces the pointwise objective to

$$\widetilde{\Phi}^{\nu}_{\mathrm{prod}}(t) = F(c_0, C)\, \phi_{r,1}(t) + F(c_0 + \delta, c_0 + \delta)\, \phi_{r,2}(t).$$

Define

$$w_1 := F(c_0, C), \qquad w_2 := F(c_0 + \delta, c_0 + \delta).$$

By (21), we have $w_1 > w_2$. Lemma B.4 then implies that the profiled router objective has a unique minimizer $t_r^\star(w_1, w_2) > 0$. Under the decoding rule

$$r(x) = 1 + \mathbf{1}\{s_r(x) \geq 0\},$$

this router minimizer decodes to expert 2.

A positive gap for every score that selects expert 1. Let $h(t) = w_1\phi_{r,1}(t) + w_2\phi_{r,2}(t)$ and $\kappa = h(0) - h(t_r^\star) > 0$. Query minima are attained, so the full surrogate infimum equals $h(t_r^\star)$. Strict convexity of h and $t_r^\star > 0$ imply $h(t) \geq h(0)$ for $t \leq 0$. For every score triple with router score $t_r \leq 0$, therefore,

$$\Phi^\nu_{\mathrm{prod}}(t_r, t_1, t_2) - \inf \Phi^\nu_{\mathrm{prod}} \geq h(0) - h(t_r^\star) = \kappa.$$

Consequently, every triple with surrogate excess less than κ has $t_r > 0$ and selects expert 2. Either advice action of expert 2 costs exactly $c_0 + \delta$, whereas the Bayes cost is c_0, giving decision excess δ. This applies to every minimizing sequence eventually, as well as every exact minimizer.

Realizing the costs with fixed experts. Take $\mathcal{Y} = [0, C]$, let (X, A, Y) be deterministic with $Y = 0$, and use absolute task loss. Set $A^1 = 1$ and define each fixed expert's outputs on the masked and revealed advice inputs to equal the corresponding two entries of (8). With zero expert and advice fees, the four task costs are exactly the displayed table and are bounded by $B = C$. The population and pointwise objectives then coincide. In particular, no interchange of ν with conditional expectation is used. If the two query objectives are additionally strictly convex, the full surrogate minimizer is unique, as it is for the logistic construction. □

Locally correct logistic queries can still misroute. For $\nu(c) = c$ and $\phi_{q,k} = \phi_k$, direct differentiation gives

$$t^\star(u, v) = \log(u/v), \qquad F(u, v) = (u + v)\log(u + v) - u\log u - v\log v.$$

With the specified sign decoder, $t^\star$ selects an action of minimum row cost. Both query heads are therefore correct on the cost table in Theorem 3. Yet its profiles satisfy $w_1 = F(c_0, C) > F(c_0 + \delta, c_0 + \delta) = w_2$. For logistic router loss factors the unique router score is $\log(w_1/w_2) > 0$, selecting the wrong expert even though each query head chooses a cheapest source for its expert.

Relation to sequential-surrogate theory. The predictor–rejector formulation of Cortes et al. (2016) separates the prediction and rejection scores. The multiplicative pairwise-comparison surrogate in Ni et al. (2019, Eq. (7)) weights a pairwise classification loss by a rejection loss factor. Equation (7) applies this construction to the two indicators that select an expert and that expert's advice.

Laha et al. (2024, Section 3.1) show that optimizing a later decision's surrogate can replace the maximum reward by a different summary and change the earlier decision. Laha et al. (2025, Theorems 4.1–4.2) give conditions under which surrogates based on nonnegative stagewise rewards preserve the optimal decisions. Our theorem covers the cost-weighted products in (7), allows nonconvex query objectives, and gives a positive surrogate gap below which every solution chooses the more costly expert in the constructed problem. Their reward results do not transfer by substitution alone: for score factors $a_i(s)$, replacing v_i by $C - v_i$ in $\sum_i v_i a_i(s)$ introduces $C\sum_i a_i(s)$, which need not be constant in the scores.

C Fixed-Bound Augmented Training and Reference Pooling

All comparisons in this section keep the fees and the deterministic bound C fixed. We compare the cost estimates obtained by the specified augmented and reference-conditional fits. With complete observations, shared conditional rates allow the fit to use records from several input cells. With limited observations, the reference frequency can also be estimated from records whose advice is missing. Throughout the statistical comparisons the task-loss bound is $B = 1$, i_0 is the reference action, and $E = Z_{i_0} \in \{0, 1\}$ is its observed error. Probability minimizers below are taken on the closed simplex or closed probability cube. Values zero and one can be approached by finite logits but need not be attained by them.

C.1 Complete observations: proof of Theorem C.1

We first compare cost estimates with all sources observed. Let $E = Z_{i_0}$ be the binary reference error and $Z = Z_i$ the binary error of another action. Let the input take finitely many values, called cells, and assume that Z is independent of X conditional on E. Thus its two conditional error rates $a = \Pr(Z = 1 \mid E = 0)$ and $b = \Pr(Z = 1 \mid E = 1)$ are shared across cells, while the reference error probability $p_x = \Pr(E = 1 \mid X = x)$ can vary freely. The action's error rate can therefore vary across inputs only through the reference's error probability. This assumption is used only for this statistical comparison.

In a sample of n independent, identically distributed records, write $n_{x,r}$ for the count with $(X, E) = (x, r)$, $n_x = \sum_r n_{x,r}$, $n_{\cdot,r} = \sum_x n_{x,r}$, and $\widehat{p}_x = n_{x,1}/n_x$. Let $\overline{Z}_x$ average action errors in cell x, and $\overline{Z}_{E=r}$ average them over every record with $E = r$. An empty reference group has zero weight in every observed cell; set its mean to zero. For the action $i = (j, k)$ and reference $i_0 = (j_0, 0)$, the true cost difference is $\Delta(x) = a + (b - a - 1)p_x + \beta_j + \gamma_k - \beta_{j_0}$. Augmented empirical minimization with free scores in each cell compares the empirical action costs, allowing limiting scores when an optimal probability is zero. Pooling the conditional errors across cells gives another estimate of the same cost difference:

$$\widehat{\Delta}_{\mathrm{aug}}(x) = \overline{Z}_x - \widehat{p}_x + \beta_j + \gamma_k - \beta_{j_0}, \qquad \widehat{\Delta}_{\mathrm{RC}}(x) = \sum_{r=0}^{1} \frac{n_{x,r}}{n_x} \overline{Z}_{E=r} - \widehat{p}_x + \beta_j + \gamma_k - \beta_{j_0}.$$

The pooled estimate uses the reference frequency in the target cell and the action's conditional error rates from all cells.

Theorem C.1 (Avoidable estimation error with complete observations)**.** *Under the preceding independent sampling and shared conditional error assumptions, condition on the design $\mathcal{G} = \sigma((X_t, E_t)_{t=1}^n)$. For every observed cell x,*

$$\begin{aligned} &\mathbb{E}[(\widehat{\Delta}_{\mathrm{aug}}(x) - \Delta(x))^2 \mid \mathcal{G}] - \mathbb{E}[(\widehat{\Delta}_{\mathrm{RC}}(x) - \Delta(x))^2 \mid \mathcal{G}] \\ &\qquad = \sum_{r:n_{x,r}>0} \frac{n_{x,r}}{n_x^2} \left(1 - \frac{n_{x,r}}{n_{\cdot,r}}\right) v_r \ \geq 0, \qquad v_0 = a(1-a), \quad v_1 = b(1-b). \end{aligned} \tag{22}$$

The improvement is strict if a reference group has positive action-error variance and contains records both inside and outside cell x. Both estimates have the same conditional mean.

Proof. Write $v_r = \mathrm{Var}(Z \mid E = r)$ and $T_r = \sum_{t:E_t=r} Z_t$. Given $\mathcal{G}$, the Z_t are independent, with the same conditional law within each reference group, because the records are independent and $Z \perp X \mid E$. Conditional on their group sum, that group's variables remain exchangeable. Hence, whenever $E_t = r$ and $n_{\cdot,r} > 0$,

$$\mathbb{E}[Z_t \mid \mathcal{G}, T_0, T_1] = T_r/n_{\cdot,r} = \overline{Z}_{E=r}.$$

Summing this identity over the records in cell x, and noting that its reference proportion $\widehat{p}_x = n_{x,1}/n_x$ and its fees are fixed given $\mathcal{G}$, proves

$$\widehat{\Delta}_{\mathrm{RC}}(x) = \mathbb{E}[\widehat{\Delta}_{\mathrm{aug}}(x) \mid \mathcal{G}, T_0, T_1]. \tag{23}$$

If a global reference group is empty, then $n_{x,r} = 0$ and its term is omitted; no division by zero is needed.

Independence of different records conditional on $\mathcal{G}$ gives

$$\mathrm{Var}(\widehat{\Delta}_{\mathrm{aug}}(x) \mid \mathcal{G}) = \sum_{r=0}^{1} \frac{n_{x,r}}{n_x^2} v_r,$$

$$\mathrm{Var}(\widehat{\Delta}_{\mathrm{RC}}(x) \mid \mathcal{G}) = \sum_{r:n_{\cdot,r}>0} \frac{n_{x,r}^2}{n_x^2 n_{\cdot,r}} v_r.$$

For the true population cost contrast, their common conditional bias is

$$\mathbb{E}[\widehat{\Delta}_M(x) \mid \mathcal{G}] - \Delta(x) = (b - a - 1)(\widehat{p}_x - p_x), \qquad M \in \{\mathrm{aug}, \mathrm{RC}\}.$$

The equal squared biases cancel. Subtracting the conditional variances therefore gives (22). Each summand is nonnegative because $n_{x,r} \le n_{\cdot,r}$, and it is positive exactly when $v_r > 0$ and $0 < n_{x,r} < n_{\cdot,r}$ for some r.

Finally, let $\overline{c}_{i',x}$ be the empirical mean cost of action $i' \in \Pi$ in cell x. With free scores in each cell, the empirical augmented weights are $\widehat{w}_{i'} = C - \overline{c}_{i',x} \ge 0$. If their sum is positive, the minimizing simplex probability is proportional to these weights, so larger probabilities correspond to smaller empirical costs. If every weight is zero, all empirical costs equal C and all actions are tied. This identifies $\widehat{\Delta}_{\mathrm{aug}}$ with the augmented empirical cost comparison, with a common tie rule. □

Convex estimation losses. Identity (23) also proves, by conditional Jensen's inequality, that for every fixed scalar target Δ and convex loss L with finite expectations,

$$\mathbb{E}[L(\widehat{\Delta}_{\mathrm{RC}}(x) - \Delta) \mid \mathcal{G}] \le \mathbb{E}[L(\widehat{\Delta}_{\mathrm{aug}}(x) - \Delta) \mid \mathcal{G}].$$

C.2 Limited observations: proof of Theorem 4

Theorem 4 (Acquisition noise in augmented training)**.** *Under the sampling assumptions above, fix $C > 1 + \beta_j + \gamma_k$. For every $n \ge 1$, the estimate $\widehat{c}_{\mathrm{aug}}$ is unbiased for $\mathbb{E}[c_t]$ and satisfies*

$$n \operatorname{Var}(\widehat{c}_{\mathrm{aug}}) = \frac{\operatorname{Var}(Z)}{\rho} + \left(\frac{1}{\rho} - 1\right)\big(C - \mathbb{E}[c_t]\big)^2. \tag{10}$$

If $\rho < 1$, the second term is strictly positive even when the action error Z is constant.

Proof. Write $U = C - c_t = C - \beta_j - \gamma_k - Z_t$. Because I_t is independent of U and has mean ρ,

$$\mathbb{E}\left[\frac{I_t U}{\rho}\right] = \mathbb{E}U, \qquad \operatorname{Var}\left(\frac{I_t U}{\rho}\right) = \frac{\operatorname{Var}(U)}{\rho} + \left(\frac{1}{\rho} - 1\right)(\mathbb{E}U)^2.$$

The records are independent, so the variance of their average is this variance divided by n. Also, $\operatorname{Var}(U) = \operatorname{Var}(Z)$, $\mathbb{E}U = C - \mathbb{E}[c_t]$. Substitution proves (10); the first identity proves unbiasedness. Since $C > 1 + \beta_j + \gamma_k \ge \mathbb{E}[c_t]$, the final term is positive for $\rho < 1$, including when Z is constant. For $p = \Pr(E = 1) \in (0, 1)$, write $a = \Pr(Z = 1 \mid E = 0)$ and $b = \Pr(Z = 1 \mid E = 1)$. The law of total variance gives $\operatorname{Var}(Z) = (1-p)a(1-a) + pb(1-b) + p(1-p)(b-a)^2$. The last term is the variation in action error explained by reference success or failure. □

The next proposition compares this ordinary augmented estimate with the reference-conditional estimate introduced in Subsection 4.3.

Proposition C.2 (Variance removed by reference conditioning with limited advice)**.** *Under the sampling assumptions of Theorem 4, write $p = \Pr(E = 1) \in (0, 1)$, $a = \Pr(Z = 1 \mid E = 0)$, and $b = \Pr(Z = 1 \mid E = 1)$. Let $\widehat{p}$ be the reference error frequency on all n records and let $\widehat{a}, \widehat{b}$ be the action error averages among acquired records with $E = 0$ and $E = 1$, respectively. Set $\widehat{c}_{\mathrm{RC}} = \beta_j + \gamma_k + (1 - \widehat{p})\widehat{a} + \widehat{p}\widehat{b}$; empty group means may be assigned any bounded value. If τ^2_{aug} and τ^2_{RC} are the limiting variances of the $\sqrt{n}$-scaled cost-estimation errors, then*

$$\tau^2_{\mathrm{aug}} - \tau^2_{\mathrm{RC}} = \left(\frac{1}{\rho} - 1\right)\left[p(1-p)(b-a)^2 + (C - \mathbb{E}[c_t])^2\right].$$

The same difference holds for estimated cost gaps from the always observed reference. Even if a fixed centering constant in the augmented estimate is chosen using the true action mean, its excess limiting variance is $(\rho^{-1} - 1)p(1-p)(b-a)^2$.

Proof. All limits in this proof take $n \to \infty$ with the distribution, fees, and acquisition probability $\rho > 0$ fixed. Set $\theta = b - a$, $m(E) = a + \theta E$, $\epsilon = Z - m(E)$, and $W = I/\rho$. Thus $\mathbb{E}[\epsilon \mid E] = 0$ and $\mu = a + \theta p$. For an arbitrary fixed centering constant ζ, define

$$\widehat{\mu}_\zeta = \zeta + \frac{1}{n}\sum_{t=1}^{n} W_t(Z_t - \zeta), \qquad \widehat{\Delta}_\zeta = \widehat{\mu}_\zeta - \widehat{p} + \beta_j + \gamma_k - \beta_{j_0}.$$

For the action (j, k) in the main theorem, $c_t = Z_t + \beta_j + \gamma_k$. Its augmented cost estimate is $\beta_j + \gamma_k + \widehat{\mu}_\zeta$ with $\zeta = C - \beta_j - \gamma_k$. Subtracting the observed reference cost estimate gives $\widehat{\Delta}_\zeta$. Indeed, the reference coefficient is $C - \beta_{j_0} - \widehat{p}$ and the nonreference coefficient is $n^{-1} \sum_t W_t(C - \beta_j - \gamma_k - Z_t)$. The reference coefficient minus the nonreference coefficient is $\widehat{\Delta}_\zeta$.

For the acquired reference-group means, the strong law gives denominators converging after division by n to $\rho(1-p) > 0$ and $\rho p > 0$. They are eventually positive almost surely. The multivariate central limit theorem applies to the bounded empirical moments defining $\widehat{p}, \widehat{a}, \widehat{b}$. The ratio delta method yields influence functions

$$E - p, \qquad \frac{W(1-E)}{1-p}(Z - a), \qquad \frac{WE}{p}(Z - b),$$

respectively. Differentiating the reconstruction $(1-p)a + pb - p + \beta_j + \gamma_k - \beta_{j_0}$ gives the contrast influence function

$$\mathrm{IF}_{\mathrm{RC}} = (\theta - 1)(E - p) + W\epsilon. \tag{24}$$

Arbitrary bounded definitions of empty-group means do not change this expansion, since the empty-group probabilities tend to zero.

The cost difference estimated using the fixed centering constant ζ has influence function

$$\mathrm{IF}_\zeta = W(Z - \zeta) - (\mu - \zeta) - (E - p).$$

Subtracting (24) gives

$$\xi := \mathrm{IF}_\zeta - \mathrm{IF}_{\mathrm{RC}} = (W - 1)\{\theta(E - p) + \mu - \zeta\}.$$

Independence of I from (E, Z) implies $\mathbb{E}[\xi] = 0$. It also makes ξ uncorrelated with $(\theta - 1)(E - p)$ because $\mathbb{E}(W - 1) = 0$. Its covariance with $W\epsilon$ vanishes because $\mathbb{E}[\epsilon \mid E] = 0$. Hence

$$\begin{aligned}\mathrm{Var}(\mathrm{IF}_\zeta) - \mathrm{Var}(\mathrm{IF}_{\mathrm{RC}}) &= \mathbb{E}[\xi^2] \\ &= \left(\frac{1}{\rho} - 1\right)\{p(1-p)\theta^2 + (\mu - \zeta)^2\}.\end{aligned}$$

Substituting $\zeta = C - \beta_j - \gamma_k$ gives $\mu - \zeta = \mathbb{E}[c_t] - C$, proving the claimed asymptotic variance difference for contrasts. For the cost estimates themselves, both influence functions add back $E - p$, which is uncorrelated with ξ. Their variance difference is therefore identical. The oracle constant $\zeta = \mu$ removes the squared term and leaves the reference term. The argument permits zero limiting variance, with a degenerate normal limit. When $\rho = 1$, the estimators agree: any empty reference group has zero reconstruction weight. The displayed variance gap is then zero. □

The variance comparison is asymptotic. Ordinary importance weighting divides the sum of observed weights by $n\rho$. Averaging only over acquired records would divide by their random count instead, giving a different estimator whose action rankings can differ.

C.3 Joint fitting: proof of Proposition C.3

Proposition C.3 (Classification fitting realizes the conditional estimates)**.** *In the model of Theorem C.1, fit one free reference probability per input cell and two shared conditional probabilities for the action. Joint empirical minimization of* (25) *reconstructs $\widehat{\Delta}_{\mathrm{RC}}(x)$ in every observed cell. At the fixed input of Theorem 4, jointly fitting the reference on all records and the conditional heads on acquired records reconstructs $\widehat{c}_{\mathrm{RC}} - \widehat{p} - \beta_{j_0}$ whenever both acquired reference groups are nonempty. Both claims hold for every positive reference weight, with minimizers taken over the closed probability cube; boundary values are limits of finite logits.*

Proof. Write binary cross-entropy in probability form as $\phi_{\mathrm{Ber}}(u, z) = -z \log u - (1 - z) \log(1 - u)$ for $z \in [0, 1]$, with the usual extended-value conventions at $u = 0, 1$. For any nonempty set of m targets, its summed loss is

$$\sum_{t=1}^{m} \phi_{\mathrm{Ber}}(u, z_t) = m\{-\overline{z} \log u - (1 - \overline{z}) \log(1 - u)\}, \qquad \overline{z} = \frac{1}{m} \sum_{t=1}^{m} z_t.$$

Subtracting its value at $u = \bar{z}$ gives $m\,\mathrm{KL}(\mathrm{Ber}(\bar{z})\|\,\mathrm{Ber}(u))$. The unique probability minimizer is therefore the target mean, including when that mean is zero or one. Positive constant multipliers do not change this conclusion.

For complete observations, let g_x be the reference logit in cell x and let $h_{i,0}, h_{i,1}$ be the two shared conditional logits for the action. Their part of the joint objective is

$$\sum_{t=1}^{n}\left[\alpha\phi_{\log}(g_{X_t}, E_t) + \frac{1}{N}\{(1-E_t)\phi_{\log}(h_{i,0}, Z_t) + E_t\phi_{\log}(h_{i,1}, Z_t)\}\right], \qquad \alpha > 0.$$

The other action heads have separate terms. In probability coordinates, the preceding minimization identity therefore gives the reference estimate $\widehat{p}_x$ for $n_x > 0$ and conditional estimates $\widehat{a} = \overline{Z}_{E=0}$, $\widehat{b} = \overline{Z}_{E=1}$ in nonempty reference groups. Boundary values are attained in the closed probability cube and approached by finite logits. The reconstruction equals $\widehat{\Delta}_{\mathrm{RC}}(x)$. A globally empty reference group has zero reconstruction weight in each observed cell, so its arbitrary probability does not affect this identity. All three probability estimates are fitted jointly from the same sample.

At the fixed input with limited observations, replace the conditional summand by $(I_t/(N\rho))\{(1-E_t)\phi_{\log}(h_{i,0}, Z_t) + E_t\phi_{\log}(h_{i,1}, Z_t)\}$. The reference mean still uses all n records; in each acquired reference group, the multiplier $1/\rho$ is constant. Thus the conditional minimizers are precisely the acquired-group means $\widehat{a}, \widehat{b}$ whenever the groups are nonempty. If a group is empty, its conditional probability is unrestricted by the objective; choose the same bounded convention as in Proposition C.2. This yields $\widehat{c}_{\mathrm{RC}} - \widehat{p} - \beta_{j_0}$ exactly, and all such conventions have the same limiting law. The corresponding full-data and limited-data variance comparisons now follow from Theorem C.1 and Proposition C.2. □

The variance reduction with complete observations requires fitting conditional probabilities shared across cells. If both probabilities are freely fitted inside each observed cell, their reconstruction obeys the exact identity

$$(1-\widehat{p}_x)\widehat{a}_x + \widehat{p}_x\widehat{b}_x = \overline{Z}_x.$$

Such a model and augmented fitting with free scores in each cell have identical empirical cost comparisons there. Neither objective determines predictions at empty input cells. If the true conditional means vary across cells but the conditional heads are pooled globally, the population reconstruction bias at cell x is

$$(1-p_x)\{\mathbb{E}[Z \mid E=0] - \mathbb{E}[Z \mid X=x, E=0]\} + p_x\{\mathbb{E}[Z \mid E=1] - \mathbb{E}[Z \mid X=x, E=1]\}.$$

The assumption that the true conditional rates are shared across cells makes this bias zero. It is an additional modeling assumption.

C.4 A LOCAL CONSEQUENCE FOR CHOOSING AN ACTION

A smaller mean squared estimation error does not by itself guarantee fewer incorrect action choices. We next compare decisions between two actions whose expected costs differ by order $n^{-1/2}$, with complete observations.

Proposition C.4 (Two-action decision risk near a tie)**.** *Assume the scalar complete-observation setting with $m \geq 2$ input cells, $\omega_y = \Pr(X = y) > 0$ and $p_y \in (0,1)$ for every cell, and $a, b \in (0,1)$. Only the reference and one further action $i = (j,k)$ are available. Put $\theta = b - a$, $v_0 = a(1-a)$, $v_1 = b(1-b)$, and $\overline{p} = \sum_y \omega_y p_y$. Fix the fees and suppose the true contrast at a target cell x is zero at an interior probability p_x^{tie}:*

$$a + (\theta - 1)p_x^{\mathrm{tie}} + \beta_j + \gamma_k - \beta_{j_0} = 0.$$

For fixed $u \neq 0$, vary only this probability to $p_{x,n} = p_x^{\mathrm{tie}} + u/\sqrt{n}$ for sufficiently large n; the other parameters remain fixed. Let $\mathrm{Reg}_{M,n}(x)$ be the expected excess cost of choosing the action with smaller estimated cost at cell x, averaged over its training sample, for $M \in \{\mathrm{aug}, \mathrm{RC}\}$. Then

$$\lim_{n\to\infty} \sqrt{n}\ \mathrm{Reg}_{M,n}(x) = |(\theta-1)u|\, F_{\mathrm{N}}\left(-\frac{|(\theta-1)u|}{\tau_M(x)}\right),$$

where F_{N} is the standard normal distribution function. The following variances are evaluated at the limiting law, with $p_x = p_x^{\mathrm{tie}}$:

$$\tau^2_{\mathrm{aug}}(x) = \frac{(\theta-1)^2 p_x(1-p_x) + (1-p_x)v_0 + p_x v_1}{\omega_x},$$
$$\tau^2_{\mathrm{RC}}(x) = \frac{(\theta-1)^2 p_x(1-p_x)}{\omega_x} + \frac{(1-p_x)^2 v_0}{1-\overline{p}} + \frac{p_x^2 v_1}{\overline{p}}.$$

Under these assumptions, $\tau^2_{\mathrm{aug}}(x) > \tau^2_{\mathrm{RC}}(x) > 0$, and the augmented scaled excess-cost limit is strictly larger.

Proof. Put $\epsilon = Z - a - \theta E$. Its conditional mean given (X, E) is zero and its conditional variance is v_E. The ratio delta method for the empirical cell mean gives the augmented contrast influence function

$$\mathrm{IF}_{\mathrm{aug,x}} = \frac{\mathbf{1}\{X = x\}}{\omega_x}\{(\theta-1)(E - p_x) + \epsilon\}.$$

Differentiating $a + (b - a - 1)p_x + \beta_j + \gamma_k - \beta_{j_0}$ with respect to the cell reference probability and the two global group means gives

$$\mathrm{IF}_{\mathrm{RC,x}} = (\theta-1)\frac{\mathbf{1}\{X = x\}}{\omega_x}(E - p_x) + \frac{(1-p_x)(1-E)}{1-\overline{p}}\epsilon + \frac{p_x E}{\overline{p}}\epsilon.$$

The reference term is uncorrelated with each residual term by $\mathbb{E}[\epsilon \mid X, E] = 0$; the two residual group terms have zero product. Their second moments give the two asserted variances. Subtracting them yields

$$\tau^2_{\mathrm{aug}}(x) - \tau^2_{\mathrm{RC}}(x) = \frac{v_0(1-p_x)\sum_{y\neq x}\omega_y(1-p_y)}{\omega_x(1-\overline{p})} + \frac{v_1 p_x \sum_{y\neq x}\omega_y p_y}{\omega_x \overline{p}} > 0.$$

All denominators are positive, and both numerators are positive under the interior assumptions and the presence of another cell.

Under the local alternatives, each row of the empirical-moment triangular array consists of independent, identically distributed bounded vectors. Their covariances converge to that of the limiting law, and boundedness implies the Lindeberg condition for the centered array scaled by $n^{-1/2}$. The multivariate triangular-array central limit theorem therefore applies. Its ratio denominators converge to strictly positive limits, so the delta method gives

$$\sqrt{n}\{\widehat{\Delta}_M(x) - \Delta_n(x)\} \implies \mathcal{N}(0, \tau^2_M(x)), \qquad \Delta_n(x) = (\theta-1)u/\sqrt{n}.$$

Define predictions on empty cells or reference groups arbitrarily; their probabilities tend to zero and do not change this conclusion. Since $a, b \in (0,1)$, $\theta < 1$ and $\kappa = (\theta-1)u \neq 0$. The probability of selecting the more costly action consequently tends to $F_{\mathrm{N}}(-|\kappa|/\tau_M(x))$. The continuous normal limit also implies that empirical ties have asymptotically vanishing probability. With two actions, the expected excess cost is exactly $|\Delta_n(x)|$ times that error probability, proving the limit. Finally, $F_{\mathrm{N}}(-|\kappa|/\tau)$ is strictly increasing in $\tau > 0$, which proves the strict comparison. □

D Reference-Conditional Classification

Under the acquisition assumptions of Proposition 6, the full and limited surrogates share the same population objective. We first describe their implementation and prove the consistency bound for the full objective. The proof expresses excess cross-entropy as relative entropy, which controls the estimated task losses and hence the selected pair's cost. We then justify the fractional targets and prove the equality of the two population objectives. The variance comparisons are proved in Appendix C.

D.1 Training and deployment

For binary task losses, the full objective uses the observed reference error to select which conditional logit is trained for each action. With $\phi_{\log}$ as defined in Subsection 4.3, it becomes

$$\Phi_{\mathrm{RC},\alpha} = \alpha\phi_{\log}(g, Z_{i_0}) + \frac{1}{N}\sum_{i\neq i_0}\left[(1 - Z_{i_0})\phi_{\log}(h_{i,0}, Z_i) + Z_{i_0}\phi_{\log}(h_{i,1}, Z_i)\right], \qquad \alpha > 0. \quad (25)$$

Algorithm 1 gives the training and deployment procedure for both observation regimes.

Algorithm 1 Reference-conditional classification

Require: Labeled inputs $(x_t, y_t)_{t=1}^n$, fixed experts $(e_j)_{j=1}^J$, $K \geq 1$ sources, a no-advice reference i_0, loss bound $B > 0$, weight $\alpha > 0$.

1: Choose the full regime or inclusion probabilities $\rho_{t,k} > 0$.
2: **for** $t = 1, \ldots, n$ **do**
3: Draw the paid-source set M_t using only (x_t, H_t) and independent randomization, with the stated inclusion probabilities.
4: Observe all no-advice losses and acquire only $(a_t^k)_{k \in M_t}$.
5: Compute $d_{t,j,k} = \psi(e_j(x_t, \widetilde{a}_t^{(k)}), y_t)/B$ for $k \in M_t \cup \{0\}$ and form the targets in (12).
6: **end for**
7: Jointly fit all logits by minimizing the empirical version of (15); use $I_k = \rho_k = 1$ for full advice.
8: **Deployment:** compute the normalized task-loss estimates $\widehat{v}_i(x)$ from Section 4.3.
9: $(\widehat{j}, \widehat{k}) \leftarrow \arg\min_{j \in [J],\, k \in [K]_0} \{B\widehat{v}_{j,k}(x) + \beta_j + \gamma_k\}$.
10: Acquire $a^{\widehat{k}}$ if $\widehat{k} > 0$; return $e_{\widehat{j}}(x, \widetilde{a}^{(\widehat{k})})$.

D.2 Proof of Theorem 5

Theorem 5 (Consistency of reference-conditional classification)**.** *Let $N \geq 2$, let $\alpha > 0$, let every task loss lie in $[0, B]$ with $B > 0$, and fix the nonnegative expert and advice fees. Let $\mathcal{E}_{\Phi_{\mathrm{RC},\alpha}} = \mathbb{E}[\Phi_{\mathrm{RC},\alpha}]$. Its infimum $\mathcal{E}^\star_{\Phi_{\mathrm{RC},\alpha}}$ is over all measurable finite-valued logits. Every measurable finite-valued logit model with finite surrogate risk satisfies*

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\widehat{\pi}) - \mathcal{E}^\star_{\ell_{\text{def-adv}}} \leq B\sqrt{\max\{N, 2/\alpha\}\left(\mathcal{E}_{\Phi_{\mathrm{RC},\alpha}} - \mathcal{E}^\star_{\Phi_{\mathrm{RC},\alpha}}\right)}. \tag{14}$$

Proof. **Representing bounded losses by binary outcomes.** Fix measurable versions of the conditional probabilities below given $X = x$. Introduce one auxiliary variable $U \sim \mathrm{Unif}[0, 1]$, independent of the data, and define $\widetilde{Z}_i = \mathbf{1}\{U \leq D_i/B\}$. This variable is used only for the proof. Conditional on the realized normalized losses, the four probabilities of $(\widetilde{Z}_{i_0}, \widetilde{Z}_i)$ are exactly the targets in equation (12). In particular,

$$\mathbb{E}[\widetilde{Z}_i \mid X = x] = \mathbb{E}[D_i \mid X = x]/B.$$

Let $p^\star(x) = \Pr(\widetilde{Z}_{i_0} = 1 \mid X = x)$. For $i \neq i_0$, choose versions

$$a_i^\star(x) = \Pr(\widetilde{Z}_i = 1 \mid \widetilde{Z}_{i_0} = 0, X = x), \qquad b_i^\star(x) = \Pr(\widetilde{Z}_i = 1 \mid \widetilde{Z}_{i_0} = 1, X = x),$$

with arbitrary values when the conditioning event has probability zero.

Expressing the surrogate excess. For a candidate model, write $\widehat{p} = \sigma(g)$, $\widehat{a}_i = \sigma(h_{i,0})$, and $\widehat{b}_i = \sigma(h_{i,1})$. Define

$$k_0(x) = \mathrm{KL}(\mathrm{Ber}(p^\star) \| \,\mathrm{Ber}(\widehat{p})),$$
$$k_i(x) = (1 - p^\star)\,\mathrm{KL}(\mathrm{Ber}(a_i^\star) \| \,\mathrm{Ber}(\widehat{a}_i)) + p^\star\,\mathrm{KL}(\mathrm{Ber}(b_i^\star) \| \,\mathrm{Ber}(\widehat{b}_i)).$$

Taking the conditional expectation of (13) and using the Bernoulli cross-entropy decomposition shows that its gap above the corresponding conditional entropy is

$$e_\alpha(x) = \alpha k_0(x) + \frac{1}{N}\sum_{i \neq i_0} k_i(x). \tag{26}$$

We next verify that this entropy lower bound is the unrestricted population infimum, including boundary probabilities. For $\epsilon \in (0, 1)$, replace every true Bernoulli parameter u by $u_\epsilon = (1-\epsilon)u + \epsilon/2$. Its logit is finite and measurable, and

$$\mathrm{KL}(\mathrm{Ber}(u) \| \,\mathrm{Ber}(u_\epsilon)) \leq -\log(1 - \epsilon).$$

This bound is uniform in X and tends to zero. Summing the finitely many weighted terms and taking expectations shows that finite logits approach the entropy lower bound. Consequently

$$\mathcal{E}_{\Phi_{\mathrm{RC},\alpha}} - \mathcal{E}^{\star}_{\Phi_{\mathrm{RC},\alpha}} = \mathbb{E}[e_\alpha(X)]. \tag{27}$$

Controlling each estimated task loss. For $i \neq i_0$, let P_i be the conditional joint law of $(\widetilde{Z}_{i_0}, \widetilde{Z}_i)$ and let Q_i be the law with reference marginal $\widehat{p}$ and conditional probabilities $\widehat{a}_i, \widehat{b}_i$. The chain rule for relative entropy gives

$$\mathrm{KL}(P_i \| Q_i) = k_0 + k_i.$$

Mapping the joint law to its $\widetilde{Z}_i$ marginal cannot increase relative entropy. Bernoulli Pinsker's inequality therefore yields

$$\left| \frac{\mathbb{E}[D_i \mid X = x]}{B} - \widehat{v}_i(x) \right| \leq \sqrt{\frac{k_0(x) + k_i(x)}{2}}. \tag{28}$$

For the reference, the same bound holds with $k_{i_0} = 0$.

Controlling the selected pair's cost. Let $\widehat{i} = \widehat{\pi}(x)$ and let $i^\star$ be a Bayes action. The conditional excess decision cost is $\mathrm{reg}(x) = \bar{c}_{\widehat{i}}(x) - \bar{c}_{i^\star}(x)$. It is zero when $\widehat{i} = i^\star$. Otherwise, the decoder minimizes estimated task loss plus the exact fees, so

$$\begin{aligned}
\mathrm{reg}(x) &\leq B \left(\left| \widehat{v}_{\widehat{i}} - \frac{\mathbb{E}[D_{\widehat{i}} \mid X = x]}{B} \right| + \left| \widehat{v}_{i^\star} - \frac{\mathbb{E}[D_{i^\star} \mid X = x]}{B} \right| \right) \\
&\leq B \left(\sqrt{\frac{k_0 + k_{\widehat{i}}}{2}} + \sqrt{\frac{k_0 + k_{i^\star}}{2}} \right).
\end{aligned}$$

By Cauchy–Schwarz,

$$\mathrm{reg}(x)^2 \leq B^2 (2k_0 + k_{\widehat{i}} + k_{i^\star}).$$

Since $\widehat{i} \neq i^\star$, their two terms satisfy $k_{\widehat{i}} + k_{i^\star} \leq \sum_{i \neq i_0} k_i$, with $k_{i_0} = 0$. Equation (26) therefore gives

$$2k_0 + k_{\widehat{i}} + k_{i^\star} \leq \max\{N, 2/\alpha\}\, e_\alpha(x).$$

Integrating and applying Jensen's inequality to the square root gives

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\widehat{\pi}) - \mathcal{E}^{\star}_{\ell_{\text{def-adv}}} \leq B \sqrt{\max\{N, 2/\alpha\}\, (\mathcal{E}_{\Phi_{\mathrm{RC},\alpha}} - \mathcal{E}^{\star}_{\Phi_{\mathrm{RC},\alpha}})}.$$

This proves (14). Every step remains valid when the excess risk is infinite, in which case the inequality is immediate. □

D.3 Fractional targets for bounded losses

The consistency proof uses the table's marginal probabilities to recover the expected task loss. Among tables with those same marginals, the shared uniform draw gives the smallest probability that the reference and action targets differ. The next lemma proves this property.

Lemma D.1 (Disagreement in the fractional target)**.** *Fix $d_{i_0}, d_i \in [0, 1]$. Among all joint binary laws with $\Pr(\widetilde{Z}_{i_0} = 1) = d_{i_0}$ and $\Pr(\widetilde{Z}_i = 1) = d_i$, the table in equation (12) minimizes $\Pr(\widetilde{Z}_{i_0} \neq \widetilde{Z}_i)$, with minimum $|d_i - d_{i_0}|$. Every joint target with these marginals recovers the same marginal action mean through the conditional construction, although its conditional targets may differ.*

Proof. For any admissible binary pair,

$$|d_i - d_{i_0}| = |\mathbb{E}(\widetilde{Z}_i - \widetilde{Z}_{i_0})| \leq \mathbb{E}|\widetilde{Z}_i - \widetilde{Z}_{i_0}| = \Pr(\widetilde{Z}_i \neq \widetilde{Z}_{i_0}).$$

The displayed table has off-diagonal mass $(d_i - d_{i_0})_+ + (d_{i_0} - d_i)_+ = |d_i - d_{i_0}|$ and the required marginals, so it attains this lower bound. For any other target with those marginals, the law of total probability gives $\Pr(\widetilde{Z}_i = 1) = (1 - d_{i_0}) \Pr(\widetilde{Z}_i = 1 \mid \widetilde{Z}_{i_0} = 0) + d_{i_0} \Pr(\widetilde{Z}_i = 1 \mid \widetilde{Z}_{i_0} = 1) = d_i$, with arbitrary values on null strata. Applying this identity conditionally to the realized losses, then averaging given X, preserves $\mathbb{E}[D_i \mid X]/B$. The consistency proof above uses this marginal identity and the relative entropy decomposition, so it also applies to these alternative targets when they use only the reference and the action loss. □

D.4 Learning from observed source blocks

Each acquired source reveals a block of action losses, one for each expert. Weighting this entire block by its inverse inclusion probability recovers its full-observation contribution in expectation. The reference and other no-advice terms are observed on every record and require no weighting.

Proposition 6 (Limited advice targets the full classification risk)**.** *Suppose each source has known positive inclusion probability $\rho_k(X, H)$ and, conditional on (X, H), its acquisition indicator is independent of the current label and source outcomes. For every fixed finite logit model,*

$$\mathbb{E}[\widetilde{\Phi}_{\mathrm{RC},\alpha} \mid X, A, Y, H] = \Phi_{\mathrm{RC},\alpha}. \tag{16}$$

Whenever differentiation and conditional expectation may be interchanged, the same identity holds for the gradient. Consequently the population risk of (15) *obeys the consistency bound in Theorem 5.*

Proof. Condition on (X, A, Y, H). The no-advice terms in equation (15) are then fixed. For each paid source, the acquisition assumptions give

$$\mathbb{E}\left[\frac{I_k}{\rho_k(X, H)}\middle| X, A, Y, H\right] = 1.$$

The term $\phi_{j,k}$ is measurable with respect to (X, Y, A^k) and the fixed logits. Linearity of conditional expectation therefore gives

$$\mathbb{E}\left[\frac{I_k}{\rho_k(X, H)}\sum_{j=1}^{J}\phi_{j,k}\middle| X, A, Y, H\right] = \sum_{j=1}^{J}\phi_{j,k}.$$

Summing over sources and adding the no-advice terms proves equation (16). The losses are nonnegative, so Tonelli's theorem permits integration even for an infinite population risk.

For finite logits, each logit derivative is a bounded Bernoulli residual multiplied by a target weight in $[0, 1]$. Applying the same conditional calculation coordinatewise proves the logit-gradient identity. For a fixed differentiable model, the chain rule gives the corresponding identity for parameter gradients whenever the required moments are integrable. The expected limited loss therefore equals $\mathcal{E}_{\Phi_{\mathrm{RC},\alpha}}$ for every fixed model and has the same unrestricted infimum. Theorem 5 then applies to this common population risk. □

Positive inclusion probabilities are enough for unbiasedness. Concentration and stochastic optimization require additional control of inverse inclusion probabilities; Proposition 6 does not supply such a guarantee.

E Experiments

The experiments evaluate two aspects of the method. With every source observed during training, we compare the decisions learned by reference conditional (RC) full, augmented training, and Learning-to-Defer on FEVER (Appendix E.2.1) and Sensitive (Appendix E.2.2). With a fixed acquisition budget, we compare observing all sources on fewer records against spreading those observations across more records (Appendices E.3.1 and E.3.2). The protocol below specifies the training observations, model selection, and resources used in each comparison.

E.1 Common protocol

Actions and metric. Both tasks have $J = 3$ experts and four paid sources, plus no advice, giving $N = 15$ actions. The task loss is 0–1 classification loss ($B = 1$). For the selected pair (j, k), the reported decision loss is

$$\mathbf{1}\{e_j(X, \widetilde{A}^{(k)}) \neq Y\} + \beta_j + \lambda\gamma_k^{\mathrm{base}}.$$

This loss adds the prediction error, expert fee, and source fee; lower is better. The policy sees only the original input before selecting a pair. Every dev example has a complete source table for evaluating the selected pair; its acquisitions are outside the policy-training budget.

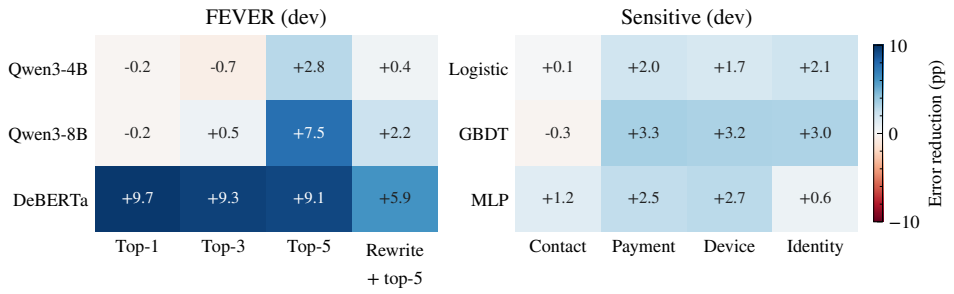

Figure 4: **Advice effects depend on the expert.** Dev task-error reduction relative to no advice, in percentage points, for each fixed expert–source pair. Fees are excluded. Positive values indicate that the source reduces the expert's error.

Reference conditional model. The network outputs two logits for each of the 15 actions. One channel for the chosen no-advice reference supplies the reference logit; both channels for every other action supply its conditional logits. Section 4.3 defines the decoder. RC full trains on every source block for each selected record. RC limited always has the no-advice block and trains on the paid-source blocks in its acquisition mask, weighting each observed paid block by the inverse of its inclusion probability.

On Sensitive, the output head shares parameters across experts and sources. In each score channel, the score for expert j without advice provides a baseline. A paid source adds a source effect and an interaction specific to that expert and source. RC uses this decomposition for both conditional logits. The augmented baseline uses the same trunk and decomposition with one score channel. The two models have 9,166 and 8,711 trainable parameters, respectively.

Configuration and checkpoint selection. All methods evaluate the same number of hyperparameter configurations per task. They use the same backbone architecture, with output heads implementing their respective objectives. For RC, the hyperparameter grid contains 36 configurations:

$$\text{learning rate} \in \{3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}, \quad \text{weight decay} \in \{0, 10^{-4}\},$$
$$\alpha \in \{1/4, 1\}, \qquad i_0 \in \{(1,0),(2,0),(3,0)\}.$$

We fit each configuration with development seeds 7 and 17 and select one configuration per task and RC method. The selected configuration is reused across all reported fees and acquisition budgets. The dev-set criterion and checkpoint selection depend on the task.

On Sensitive, configuration selection uses a separate development dataset. For each fee and development seed, we retain the lowest dev decision loss observed during training. We average these minima over all nine fees and both seeds to select the configuration. The final data splits exclude every record in this development dataset. In the reported runs, each fee uses the checkpoint with the lowest dev decision loss at that fee.

On FEVER, each checkpoint is evaluated by its mean dev loss at $\lambda \in \{0, 2\}$. We select the checkpoint with the lowest mean and use it at every fee. Configuration selection averages this criterion over the two development seeds. Table 2 lists the retained configurations.

Table 2: Reference conditional configurations selected during development and retained for every reported fee and acquisition budget.

Task	Method	Learning rate	Weight decay	α	Reference
Sensitive	RC full	10^{-3}	0	1	$(1,0)$
Sensitive	RC limited	10^{-3}	0	1	$(1,0)$
FEVER	RC full	10^{-3}	10^{-4}	$1/4$	$(2,0)$
FEVER	RC limited	10^{-3}	0	1	$(1,0)$

All-source runs use policy seeds 101–104. Acquisition curves use seeds 101–108; none of these seeds were used for configuration selection. Data splits and cached expert outcomes are fixed across runs. Seeds vary policy initialization and, for limited training, source assignment.

Compared methods. The augmented baseline scores all 15 expert–source actions with the fixed-bound surrogate Φ^{C}_{aug} in Definition 3. For each fee, C is fixed and bounds every pair cost. Augmented training uses complete source tables on its selected training records, including in the acquisition comparison. Standard Learning-to-Defer selects only among the three no-advice experts. On Sensitive, RC full and augmented training use the same trunk and structured head design. On both tasks, RC learns input-dependent conditional error probabilities.

Baseline results are reused only when the split, expert cache, fees, and training observations match. All-source augmented and Learning-to-Defer runs use the same prepared data, fee grid, and seeds 101–104 as RC full. Augmented training retains its fee-specific checkpoint selection, Learning-to-Defer uses one no-advice checkpoint, and RC follows the task-specific rules above.

The acquisition comparison matches paid-source counts and optimizer updates across RC full, RC limited, and augmented training. The augmented runs use the same records and complete source tables as RC full. This comparison tests how well the methods use a fixed number of paid observations.

E.2 All-source experiments

All-source training observes every paid source on each policy-training record. This comparison evaluates the learned decisions when every action outcome is available during training. RC full and augmented training can select an expert–source pair; standard Learning-to-Defer selects an expert without advice. We vary the advice fee while keeping the experts, sources, and data fixed. Tables 3 and 4 give the losses; Figure 1 plots the same fee sweeps.

E.2.1 FEVER retrieval

FEVER (Thorne et al., 2018) classifies claims as `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFO`. A stratified sample of 15,000 examples from its labeled development split provides 11,250 policy-training and 3,750 dev examples. The fixed experts are Qwen3-4B-Instruct, Qwen3-8B-Instruct (Yang et al., 2025), and DeBERTa-v3-large (He et al., 2021), with consultation fees $\beta = (0.03, 0.05, 0.04)$.

The paid sources are BM25 top-1, top-3, top-5, and a reformulated query followed by BM25 top-5. Including the no-advice option, the base fees are $\gamma^{\mathrm{base}} = (0, 0.015, 0.02, 0.03, 0.01)$, with $\lambda \in \{0, 0.5, 1, 2, 4, 6, 8, 10\}$. Policy features are frozen, normalized, claim-only `all-MiniLM-L6-v2` embeddings. The RC network maps these embeddings through ReLU layers of sizes 128 and 64, followed by a rank-64 head for expert–source predictions. AdamW uses batch size 128, gradient clipping at 10, a constant learning rate, and 50 epochs. Dev loss is evaluated after epoch 1, every two epochs, and at the final epoch.

Table 3: All-source dev results on FEVER (mean ± sample standard deviation over seeds 101–104). Each method uses its stated dev checkpoint rule. Lower is better.

Fee	Reference conditional full	Augmented surrogate	Learning-to-Defer
0	**0.555 ± 0.004**	0.563 ± 0.002	0.627 ± 0.002
0.5	**0.560 ± 0.003**	0.565 ± 0.006	0.627 ± 0.002
1	**0.566 ± 0.006**	0.572 ± 0.002	0.627 ± 0.002
2	**0.576 ± 0.005**	0.584 ± 0.004	0.627 ± 0.002
4	**0.592 ± 0.003**	0.600 ± 0.003	0.627 ± 0.002
6	**0.603 ± 0.004**	0.614 ± 0.006	0.627 ± 0.002
8	**0.612 ± 0.003**	0.621 ± 0.004	0.627 ± 0.002
10	**0.619 ± 0.006**	0.627 ± 0.004	0.627 ± 0.002

Table 3 shows that RC full has the lowest mean dev loss at every tested fee. At $\lambda = 2$, its mean is 0.576, versus 0.584 for augmented training and 0.627 for Learning-to-Defer. Its gap to the no-advice baseline narrows as advice becomes more expensive. The fee sweep changes decision costs while the cached expert outcomes remain fixed.

E.2.2 SENSITIVE FEATURE ACQUISITION

Sensitive uses the IEEE-CIS fraud data (IEEE Computational Intelligence Society, 2019). A restricted transaction view is always available. Four nested paid sources successively reveal contact metadata, payment-profile attributes, device-profile attributes, and identity attributes. Including no advice, the base fees are $\gamma^{\text{base}} = (0, 0.08, 0.12, 0.16, 0.20)$, with $\lambda \in \{0, 0.04, 0.12, 0.20, 1, 2, 3, 4, 5\}$. The fixed experts are logistic regression, a histogram gradient-boosted tree, and an MLP, with fees $\beta = (0.02, 0.06, 0.09)$.

Disjoint balanced splits contain 8,000 examples for expert fitting, 8,000 for policy training, and 4,000 for the dev set.

Policy features are standardized using training statistics. All three advice-aware methods use a 64-unit ReLU trunk, a second 64-unit score layer, and the rank-32 structured head described above. AdamW uses batch size 64 and gradient clipping at 10. RC uses a constant learning rate; augmented training uses 5 percent linear warmup followed by cosine decay. All methods train for 100 epochs. Dev loss is evaluated after the first update, every two updates, and at the final update. For each reported fee, the checkpoint with the lowest dev decision loss is selected, retaining the earliest in a tie.

Table 4: All-source dev results on Sensitive (mean $\pm$ sample standard deviation over seeds 101–104). Each method uses its stated dev checkpoint rule. Lower is better.

Fee	Reference conditional full	Augmented surrogate	Learning-to-Defer
0	**0.244 ± 0.002**	0.248 ± 0.001	0.273 ± 0.001
0.04	**0.247 ± 0.002**	0.252 ± 0.001	0.273 ± 0.001
0.12	**0.253 ± 0.001**	0.258 ± 0.002	0.273 ± 0.001
0.2	**0.256 ± 0.000**	0.262 ± 0.001	0.273 ± 0.001
1	**0.266 ± 0.001**	0.277 ± 0.001	0.273 ± 0.001
2	**0.266 ± 0.001**	0.273 ± 0.002	0.273 ± 0.001
3	**0.266 ± 0.001**	0.272 ± 0.001	0.273 ± 0.001
4	**0.266 ± 0.001**	0.270 ± 0.002	0.273 ± 0.001
5	**0.266 ± 0.001**	0.271 ± 0.001	0.273 ± 0.001

Table 5: Cost decomposition for reference conditional full on Sensitive dev (means over seeds 101–104). Decision loss is the sum of task loss, expert cost, and advice cost. Advice rate is the fraction of inputs using a paid source.

Fee	Decision loss	Task loss	Expert cost	Advice cost	Advice rate
0.2	0.2561	0.2127	0.0329	0.0105	36.17%
1	0.2660	0.2273	0.0299	0.0088	7.81%
2	0.2659	0.2366	0.0292	0.0001	0.03%
5	0.2659	0.2364	0.0295	0.0000	0.00%

Figure 5 shows that RC full requests no paid advice at $\lambda \in \{4, 5\}$ in these runs. At these fees it selects among no-advice experts using RC error estimates. Its selected expert may differ from that of the separately trained Learning-to-Defer router, although both target the same no-advice Bayes decision under their population consistency conditions.

Table 4 shows that RC full has the lowest mean dev loss at all nine fees. Table 5 details its prediction error, fees, and advice use. Its advice use falls from 36.17% at $\lambda = 0.2$ to 7.81% at $\lambda = 1$ and 0.03% at $\lambda = 2$. Its mean loss decreases slightly from 0.2660 to 0.2659 between the latter two fees because the reduction in advice cost slightly exceeds the increase in task loss. The selected checkpoint and decisions can change with the fee, so the measured loss can decrease between two fees. The Bayes risk is nondecreasing in the fee multiplier.

E.3 TRAINING UNDER EQUAL ACQUISITION BUDGETS

This comparison fixes the number of paid source requests. RC full acquires all four sources for each selected record. Below the full acquisition budget, RC limited acquires fewer sources per record to cover more inputs. For example, on Sensitive a budget of $Q = 8{,}000$ requests gives RC full all four

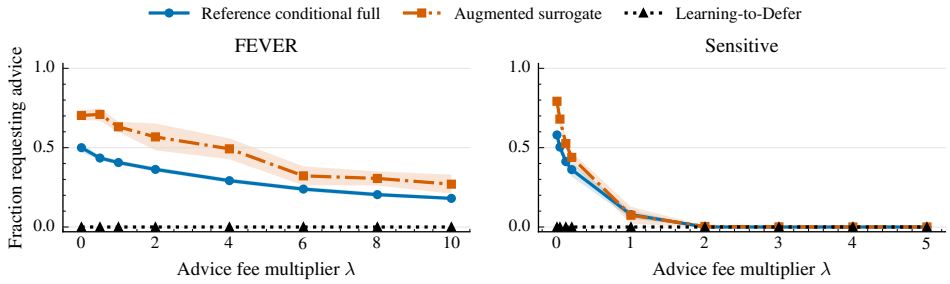

Figure 5: **Advice use across fees.** Mean fraction of evaluated examples on which each method requests a paid source; bands show one sample standard deviation over seeds 101–104. Both tasks use dev data. Learning-to-Defer has no advice action.

sources on 2,000 transactions, or RC limited one source on each of 8,000 transactions. In both cases, an acquired source can be supplied to all three fixed experts.

Let n be the number of available policy-training records. The nine budget fractions are

$$\left\{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{5}{8}, \frac{3}{4}, 1\right\}.$$

For fraction η in this set, the paid-source budget is $Q = 4\lfloor \eta n \rfloor$. RC limited uses $\min\{Q, n\}$ labeled records, each with at least one acquired source. It assigns each used record a random cyclic ordering of the four sources, independently of its outcomes. The design gives as many records as possible a first source before adding second, third, and fourth sources to records in their assigned order. If a record receives s sources, each source has inclusion probability $s/4$; the observed paid terms receive the corresponding inverse weights in equation (15). RC full observes all four sources on a random subset of $Q/4$ records. Augmented training uses the same subset and complete source tables as RC full.

Let T_{ep} be the full-data epoch count and m_{batch} the batch size. Each method receives

$$U(Q) = T_{\text{ep}} \left\lceil \frac{Q/4}{m_{\text{batch}}} \right\rceil$$

optimizer updates, capped at the full-data count. The settings are $(T_{\text{ep}}, m_{\text{batch}}) = (100, 64)$ for Sensitive and $(50, 128)$ for FEVER. Thus the three methods receive equal update counts at each budget even though they cover different numbers of records.

Table 6 reports the labels, source requests, expert outcomes, and updates used at each budget. Augmented training uses the same resources as RC full. Here n_{dev} is the dev-set size. Counts refer to distinct observed records and outcomes: reusing a cached outcome in later updates does not add an acquisition. On Sensitive, a source is a requested cumulative feature group, so Q counts source requests rather than distinct feature values.

Table 6: Resources used at acquisition budget Q with three experts and four paid sources. The dev table has $n_{\text{dev}} = 3{,}750$ records on FEVER and $4{,}000$ on Sensitive and is outside Q.

Resource	Reference conditional full	Reference conditional limited
Training labels used	$Q/4$	$\min\{Q, n\}$
Paid source requests	Q	Q
Expert outcomes under paid sources	$3Q$	$3Q$
No-advice expert outcomes	$3Q/4$	$3\min\{Q, n\}$
Dev source requests	$4n_{\text{dev}}$	$4n_{\text{dev}}$
Optimizer updates	$U(Q)$	$U(Q)$

Figure 2 reports the mean and sample standard deviation over seeds 101–108. At equal Q, RC limited can use up to four times as many labels and no-advice outcomes as RC full. Table 6 records these additional observations alongside the matched source requests and optimizer updates. Both designs have positive source inclusion probabilities, as required by the population consistency results.

E.3.1 FEVER RETRIEVAL

For FEVER, $n = 11{,}250$, giving budgets from $Q = 1{,}404$ to $45{,}000$. At small budgets, RC limited covers more distinct claims and no-advice outcomes; later acquisitions fill in sources for those claims. At $Q = 45{,}000$, both allocation rules observe the complete training table. Figure 2 reports dev loss at the primary fee $\lambda = 2$.

Table 7: Reference conditional full, reference conditional limited and augmented training on FEVER (dev) at $\lambda = 2$ with the same number Q of paid source observations and optimizer updates (mean $\pm$ sample standard deviation over seeds 101–108). Lower is better.

Budget Q	RC full	RC limited	Augmented	Limited − full	Limited lower
1404	0.685 ± 0.013	0.674 ± 0.024	0.651 ± 0.007	-0.0116	6/8
2812	0.668 ± 0.021	0.656 ± 0.011	0.644 ± 0.005	-0.0125	6/8
5624	0.643 ± 0.004	0.623 ± 0.008	0.639 ± 0.006	-0.0201	8/8
11248	0.627 ± 0.006	0.601 ± 0.007	0.624 ± 0.006	-0.0267	8/8
16872	0.607 ± 0.007	0.589 ± 0.006	0.614 ± 0.006	-0.0181	8/8
22500	0.597 ± 0.005	0.579 ± 0.005	0.606 ± 0.004	-0.0180	8/8
28124	0.587 ± 0.006	0.579 ± 0.005	0.601 ± 0.003	-0.0078	6/8
33748	0.582 ± 0.007	0.580 ± 0.005	0.594 ± 0.007	-0.0027	5/8
45000	0.574 ± 0.003	0.571 ± 0.006	0.587 ± 0.004	-0.0031	6/8

Table 7 shows that RC limited has lower mean loss than RC full at all nine FEVER budgets. From $Q = 5{,}624$ through $Q = 22{,}500$, the reduction ranges from 0.018 to 0.027, and every paired seed favors RC limited. Augmented training has the lowest mean at the two smallest budgets, $Q = 1{,}404$ and $2{,}812$; RC limited has the lowest mean from $Q = 5{,}624$ onward. At the two largest budgets, the mean difference between the two RC methods is about 0.003, and individual seeds favor different methods. At the complete-data budget, both methods observe the same training table but retain separately selected hyperparameters, including different reference experts, so their fitted policies can differ.

E.3.2 SENSITIVE FEATURE ACQUISITION

For Sensitive, $n = 8{,}000$, giving budgets from $Q = 1{,}000$ to $32{,}000$. At small budgets, RC limited covers more transactions, whereas RC full observes all four nested feature groups on each selected transaction. At $Q = 32{,}000$, both rules observe every source on every training record. Figure 2 reports dev loss at the primary fee $\lambda = 0.12$. The seed variation includes policy fitting and random source ordering, but not a new data split or refitted experts.

Table 8: Reference conditional full, reference conditional limited and augmented training on Sensitive (dev) at $\lambda = 0.12$ with the same number Q of paid source observations and optimizer updates (mean $\pm$ sample standard deviation over seeds 101–108). Lower is better.

Budget Q	RC full	RC limited	Augmented	Limited − full	Limited lower
1000	0.277 ± 0.007	0.273 ± 0.004	0.285 ± 0.005	-0.0037	6/8
2000	0.272 ± 0.005	0.269 ± 0.003	0.279 ± 0.003	-0.0029	6/8
4000	0.266 ± 0.002	0.264 ± 0.002	0.274 ± 0.002	-0.0022	6/8
8000	0.263 ± 0.004	0.256 ± 0.002	0.270 ± 0.003	-0.0065	8/8
12000	0.259 ± 0.003	0.254 ± 0.002	0.266 ± 0.002	-0.0043	8/8
16000	0.257 ± 0.002	0.254 ± 0.001	0.264 ± 0.002	-0.0031	7/8
20000	0.255 ± 0.002	0.253 ± 0.001	0.263 ± 0.002	-0.0022	6/8
24000	0.254 ± 0.001	0.253 ± 0.001	0.261 ± 0.001	-0.0015	7/8
32000	0.252 ± 0.002	0.252 ± 0.002	0.259 ± 0.001	0.0000	1/8

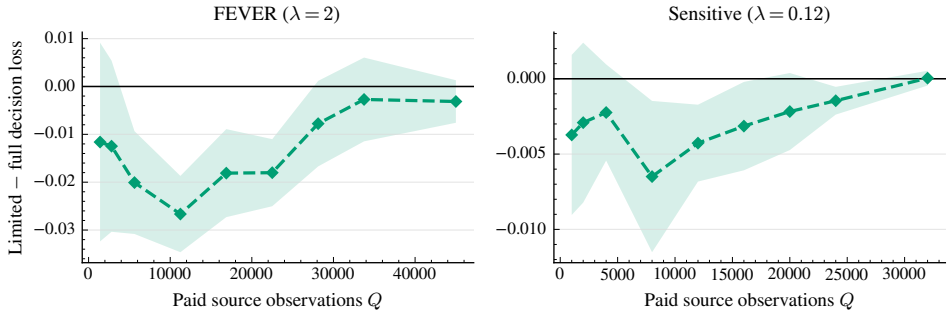

Figure 6: **Paired difference at equal acquisition budgets.** Mean within-seed decision-loss difference between reference conditional limited and reference conditional full; bands show one sample standard deviation over seeds 101–108 on the FEVER and Sensitive dev sets. Values below zero favor the limited method.

Table 8 shows that RC limited has lower mean dev loss at the first eight budgets. At $Q = 1{,}000$, its mean is 0.273, versus 0.277 for RC full. At $Q = 8{,}000$, one source per transaction gives 0.256, versus 0.263 when all four sources are observed on 2,000 transactions. At $Q = 32{,}000$, both methods observe the complete training table and both means round to 0.252; their unrounded difference is negligible relative to the reported variation (Figure 6).

In Table 8, augmented training uses the same records and sources as RC full. Its mean dev loss exceeds both RC methods at every Sensitive budget. At $Q = 8{,}000$, the means are 0.270 for augmented training, 0.263 for RC full, and 0.256 for RC limited. At the complete-table budget $Q = 32{,}000$, the augmented mean is 0.259, versus 0.252 for each RC allocation after rounding.

E.4 Controlled acquisition-variance experiment

We isolate the variance reduction in Theorem 4 and Proposition C.2 using one input and independent training records. The reference error is $E \sim \mathrm{Ber}(1/2)$, and the action error satisfies $Z \mid E = 0 \sim \mathrm{Ber}(a)$ and $Z \mid E = 1 \sim \mathrm{Ber}(b)$. Acquisition $I \sim \mathrm{Ber}(\rho)$ is independent of (E, Z). Every record reveals E; only acquired records reveal Z. The informative-reference case uses $(a, b) = (0.1, 0.9)$; the uninformative-reference control uses $(a, b) = (0.5, 0.5)$. Both cases have reference and action error rates of $1/2$, so only their dependence changes. These error laws are realizable by fixed binary experts: take $Y = E$, reference prediction 0, advice $A = E \oplus Z$, and advised prediction A, where $\oplus$ denotes exclusive OR. All fees are zero, and the fixed augmented bound is $C = 1.1$.

Estimators and predictions. On each dataset, we compare the augmented importance-weighted estimate, an acquired-only average, and the RC estimate:

$$\widehat{c}_{\mathrm{aug}} = C - \frac{1}{n\rho}\sum_{t=1}^{n} I_t(C - Z_t), \qquad \widehat{c}_{\mathrm{acq}} = \frac{\sum_{t=1}^{n} I_t Z_t}{\sum_{t=1}^{n} I_t}, \qquad \widehat{c}_{\mathrm{RC}} = (1-\widehat{p})\widehat{a} + \widehat{p}\widehat{b}.$$

Here $\widehat{p}$ uses all records, while $\widehat{a}, \widehat{b}$ use the acquired records in their respective reference groups. These are the joint empirical binary-cross-entropy minimizers over the closed probability cube (Proposition C.3). The acquired-only average removes acquisition-count noise without using the reference; it therefore separates that effect from the additional gain due to conditioning. All three estimates use the same records and acquisition mask.

Write $\mu = \mathbb{E}Z$, $V_B = p(1-p)(b-a)^2$, and $V_W = (1-p)a(1-a) + pb(1-b)$, with $p = 1/2$. The predicted scaled variances are

$$n\,\mathrm{Var}(\widehat{c}_{\mathrm{aug}}) = \frac{V_B + V_W}{\rho} + (\rho^{-1} - 1)(C - \mu)^2,$$
$$n\,\mathrm{Var}(\widehat{c}_{\mathrm{acq}}) \longrightarrow \frac{V_B + V_W}{\rho}, \qquad n\,\mathrm{Var}(\widehat{c}_{\mathrm{RC}}) \longrightarrow V_B + \frac{V_W}{\rho}.$$

The first identity is exact; the limits take $n \to \infty$ at fixed (p, a, b, ρ). For the acquired-only ratio, the delta method gives influence function $I(Z - \mu)/\rho$, whose variance is $(V_B + V_W)/\rho$. For both

ratio estimators, bounded outcomes and exponentially small probabilities of near-zero denominators also give convergence of the scaled second moments. Thus the limiting variance gap between the acquired-only average and RC is $(\rho^{-1} - 1)V_B$. It is positive in the informative case and zero in the control.

Protocol. We generate 50,000 independent datasets for each $n \in \{512, 2048, 8192\}$ and $\rho \in \{1, 1/2, 1/4, 1/8, 1/16\}$ in each case. The master seed is 20260926, with separate NumPy SeedSequence streams indexed by case, sample size, and acquisition probability. We draw the multinomial counts of the eight (E, Z, I) outcomes, which has exactly the same distribution as generating n independent records. The estimates are computed in float64 without numerical optimization. Empty acquired groups receive mean $1/2$, and no repetitions are discarded; no empty groups occurred in this run. The design and seed were fixed before execution on one cluster CPU job. Figure 7 shows $n = 8192$; Table 9 reports every setting. Error bars quantify Monte Carlo uncertainty across independent datasets.

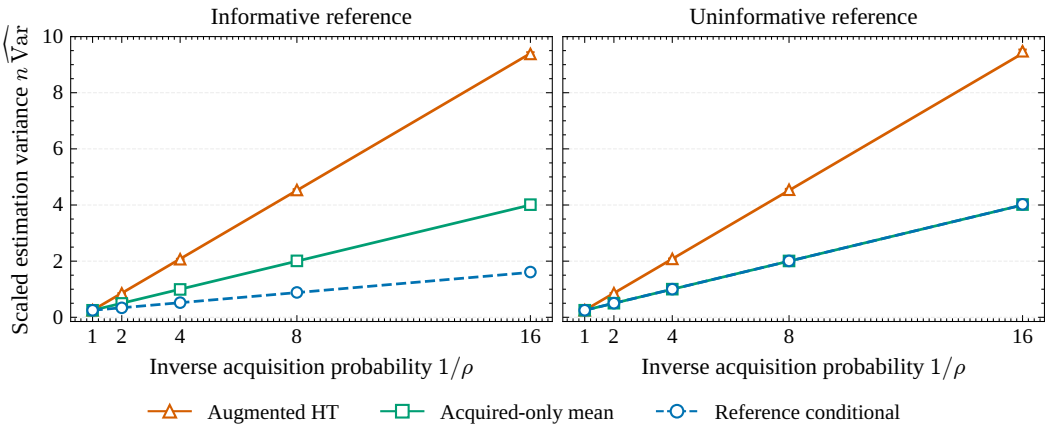

Figure 7: **Reference conditioning reduces acquisition variance.** Markers show measured $n\widehat{\mathrm{Var}}$ at $n = 8192$; error bars show one Monte Carlo standard error over 50,000 datasets. Lines show the exact augmented variance and the asymptotic variances of the acquired-only and RC estimates. Conditioning reduces variance beyond acquisition-count normalization when the reference is informative (left). The acquired-only and RC curves nearly coincide in the control (right). All three estimates coincide at $\rho = 1$.

Results. With an informative reference and $n = 8192$, the scaled variances at $\rho = 1/4$ are 2.069 for augmented importance weighting, 0.993 for the acquired-only average, and 0.521 for RC. The corresponding predictions are 2.080, 1.000, and 0.520. RC reduces variance by 47.5% relative to the acquired-only average, compared with the predicted 48%. At $\rho = 1/16$, this reduction is 59.8%, compared with the predicted 60%. In the uninformative control at $\rho = 1/4$, the acquired-only and RC scaled variances are 1.0022 and 1.0025, respectively, against a common limit of 1.000. Every empirical mean is within 2.51 Monte Carlo standard errors of the true mean $1/2$; the largest absolute bias is 7.0×10^{-4}. The supplementary code includes all estimates, bias and mean-squared-error summaries, and checks against the variance formulas. The comparison directly exhibits the reference-dependent reduction predicted by Proposition C.2.

Table 9: Scaled estimation variance $n\widehat{\mathrm{Var}}$ over 50,000 independent datasets per setting, with one Monte Carlo standard error after $\pm$. All prespecified sample sizes and acquisition probabilities are included.

n	$1/\rho$	Augmented HT	Acquired-only mean	RC
		Informative reference: $(a, b) = (0.1, 0.9)$		
512	1	0.249 ± 0.002	0.249 ± 0.002	0.249 ± 0.002
512	2	0.850 ± 0.005	0.498 ± 0.003	0.341 ± 0.002
512	4	2.087 ± 0.013	1.004 ± 0.006	0.521 ± 0.003
512	8	4.503 ± 0.029	2.010 ± 0.013	0.897 ± 0.006
512	16	9.425 ± 0.061	4.116 ± 0.026	1.688 ± 0.011
2 048	1	0.252 ± 0.002	0.252 ± 0.002	0.252 ± 0.002
2 048	2	0.859 ± 0.005	0.501 ± 0.003	0.339 ± 0.002
2 048	4	2.123 ± 0.013	1.012 ± 0.006	0.523 ± 0.003
2 048	8	4.527 ± 0.029	2.000 ± 0.013	0.877 ± 0.006
2 048	16	9.335 ± 0.059	4.068 ± 0.026	1.620 ± 0.010
8 192	1	0.250 ± 0.002	0.250 ± 0.002	0.250 ± 0.002
8 192	2	0.852 ± 0.005	0.493 ± 0.003	0.337 ± 0.002
8 192	4	2.069 ± 0.013	0.993 ± 0.006	0.521 ± 0.003
8 192	8	4.534 ± 0.028	2.010 ± 0.013	0.885 ± 0.006
8 192	16	9.382 ± 0.060	4.013 ± 0.025	1.611 ± 0.010
		Uninformative reference: $(a, b) = (0.5, 0.5)$		
512	1	0.250 ± 0.002	0.250 ± 0.002	0.250 ± 0.002
512	2	0.860 ± 0.005	0.503 ± 0.003	0.504 ± 0.003
512	4	2.078 ± 0.013	1.012 ± 0.006	1.017 ± 0.006
512	8	4.558 ± 0.029	2.030 ± 0.013	2.059 ± 0.013
512	16	9.484 ± 0.060	4.153 ± 0.026	4.300 ± 0.027
2 048	1	0.248 ± 0.002	0.248 ± 0.002	0.248 ± 0.002
2 048	2	0.853 ± 0.005	0.498 ± 0.003	0.498 ± 0.003
2 048	4	2.078 ± 0.013	1.002 ± 0.006	1.003 ± 0.006
2 048	8	4.501 ± 0.028	2.010 ± 0.013	2.018 ± 0.013
2 048	16	9.255 ± 0.059	3.999 ± 0.025	4.026 ± 0.026
8 192	1	0.251 ± 0.002	0.251 ± 0.002	0.251 ± 0.002
8 192	2	0.859 ± 0.005	0.502 ± 0.003	0.502 ± 0.003
8 192	4	2.080 ± 0.013	1.002 ± 0.006	1.003 ± 0.006
8 192	8	4.544 ± 0.029	2.007 ± 0.013	2.009 ± 0.013
8 192	16	9.477 ± 0.060	4.016 ± 0.026	4.025 ± 0.026

E.5 Reproducibility

The code for the experiments is provided in the anonymous supplementary ZIP archive.

F Additional Related Work

Selective prediction and abstention. Selective prediction allows a system to reject a prediction at a specified cost. Work in this area studies the error–rejection tradeoff (Chow, 1970), convex losses for rejection (Bartlett & Wegkamp, 2008), and selective classification with deep networks (Geifman & El-Yaniv, 2017). Predictor–rejector methods learn prediction and rejection scores jointly (Cortes et al., 2016; 2024; Ni et al., 2019; Mao et al., 2024a). Other formulations accommodate arbitrary task losses (Cao et al., 2022) or represent rejection as an additional scored action (Mao et al., 2024d).

Collaboration with human experts. L2D research studies fairness and accuracy when selecting among multiple experts (Keswani et al., 2021), and trains models to complement their capabilities (Hemmer et al., 2022). Related approaches combine human predictions with model probabilities through confusion matrices and calibration (Kerrigan et al., 2021), infer counterfactual second opinions (Benz & Rodriguez, 2022), and improve deferral through fine-tuning (Raman & Yee, 2021).

Deferral losses and algorithms. Consistent surrogates have been developed for deferral to one expert (Mozannar & Sontag, 2020) and multiple experts (Verma et al., 2022; Mao et al., 2024b). Further work studies sample efficiency (Charusaie et al., 2022), calibrated softmax parameterizations (Cao et al., 2024), and exact optimization algorithms (Mozannar et al., 2023). Deferral can also be learned after fitting the predictor (Mao et al., 2023), or extended to selecting several experts (Montreuil et al., 2026b).

Training observations and changing experts. DeSalvo et al. (2025) reduce training costs by querying a subset of experts on each labeled example. Nguyen et al. (2025) address missing expert annotations and workload allocation. Our limited setting acquires advice sources: one acquired source supplies outcomes for every fixed expert, while every no-advice outcome is available on each record. Separate work studies deferral to experts unseen during training (Tailor et al., 2024; Strong et al., 2026), online routing with varying experts (Montreuil et al., 2026c), and deferral in non-stationary time series (Montreuil et al., 2026f).

Tasks and applications. L2D has been extended to regression (Mao et al., 2024c), multi-task learning (Montreuil et al., 2025b), and adversarially robust decisions (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 for extractive question answering (Montreuil et al., 2026d). Palomba et al. (2025) develop a causal framework for evaluating deferral systems.