Consistent Learning-to-Defer with Expert-Conditional Advice

Yannis Montreuil [1] **Leina Montreuil** [2] **Axel Carlier** [3] **Lai Xing Ng** [4] **Wei Tsang Ooi** [1]

Abstract

Learning-to-Defer routes each input to the expert that minimizes expected cost, but it assumes the information available to every expert is fixed at decision time. Modern systems violate this assumption: after selecting an expert, one may also choose what information that expert receives—retrieved documents, tool outputs, or escalation context—and the value of such advice is itself expert-dependent. We introduce *Learning-to-Defer with advice*, a formulation in which the system jointly selects an expert and an expert-conditional advice action. The natural sequential surrogate—one router score and one query head per expert—turns out to be inconsistent: we prove that a broad family of such separated router/query surrogates fails to recover the Bayes rule even in the smallest non-trivial case. We resolve this by an augmented surrogate that operates on the composite expert–advice action space, for which we establish an $\mathcal{H}$-consistency bound. Across language, tabular, multi-modal, and synthetic benchmarks, the resulting method improves over Learning-to-Defer while adapting its advice-acquisition behavior to the cost regime, and a theorem-aligned synthetic benchmark instantiates the predicted failure mode of separated surrogates.

1. Introduction

Modern decision systems face two intertwined choices: *who* should handle a given instance, and *what information* that expert should receive. A retrieval-augmented model may answer only after receiving retrieved documents (Guu et al., 2020; Lewis et al., 2020). A tool-augmented agent may be granted access to a calculator, a search engine, or a code interpreter (Schick et al., 2023). These examples concern the second choice; standard Learning-to-Defer (Madras et al., 2018; Mozannar & Sontag, 2020; Mao et al., 2023a; Montreuil et al., 2026b), by contrast, addresses the first. Our setting combines both: the learner must commit to an executed expert–advice pair, where the value of advice depends on the expert who will consume it and the best expert in turn depends on the advice received.

Learning-to-Defer (L2D) treats the information available to every expert as fixed at decision time. The entire question of whether and what to acquire is outside the model. This misses an important class of systems in which the value of extra information depends on the expert who will consume it, and can even change which expert is optimal. While recent work studies the binary ask/not-ask case for a single enriched predictor (Pugnana et al., 2025), our focus is the more general routed setting in which advice is expert-conditioned and may come from multiple sources. We call this setting *Learning-to-Defer with advice*: the system jointly selects an expert and decides what information that expert should receive. The central statistical question is then Bayes consistency (Zhang, 2002; Steinwart, 2007; Bartlett et al., 2006): can one learn a policy that recovers the Bayes-optimal expert-advice decision, that is, the expert whose best advice yields the smallest conditional expected cost?

Because routing and advice are coupled — the best advice depends on the expert, and the best expert depends on the advice — one might keep the sequential structure and learn them with separate heads. We show that this natural decomposition is, in general, inconsistent: even with two experts and binary advice, the separated surrogate fails to recover the Bayes rule. The key insight is that the protocol evaluates expert–advice *pairs*, so the surrogate must operate on the same composite action space. Building on this observation, we introduce an augmented surrogate over the joint expert–advice space and prove an $\mathcal{H}$-consistency guarantee: its minimization is provably aligned with the Bayes-optimal expert–advice decision.

Contributions.

– We introduce *Learning-to-Defer with advice*, a formula-

[1]School of Computing, National University of Singapore, Singapore [2]Departement de Mathematiques, Sorbonne University, Paris, France [3]Federation ENAC ISAE-SUPAERO ONERA, Universite de Toulouse, Toulouse, France [4]Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. Correspondence to: Yannis Montreuil <yannis.montreuil@u.nus.edu>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

tion in which the system jointly selects an expert and an expert-conditional advice action, and derive the Bayes-optimal policy, which compares experts at their best-advised cost (Section 4.3, Lemma 3).

– We prove that a broad family of natural separated router-query surrogates is not Bayes consistent, even in the simplest binary-advice setting (Section 4.4, Theorem 6).

– We introduce an augmented surrogate and prove an $\mathcal{H}$-consistency guarantee together with an excess-risk transfer bound, establishing that surrogate minimization recovers the Bayes-optimal deferral-advice risk in the limit (Section 4.5, Theorem 8, Corollary 9).

– We validate the framework on tabular, language, multimodal, and synthetic experiments, where it improves over standard Learning-to-Defer, adapts its advice-acquisition rate to the cost regime, and empirically exhibits the failure mode predicted for separated surrogates (Section 5, Appendix C.2).

2. Related Work

L2D extends selective prediction and learning with abstention (Chow, 1970; Bartlett & Wegkamp, 2008; Cortes et al., 2016; Geifman & El-Yaniv, 2017; Cao et al., 2022; Ghosh et al., 2017; Mao et al., 2024a;i; Cortes et al., 2024b) by allowing a learner to defer uncertain inputs to external experts (Madras et al., 2018; Mozannar & Sontag, 2020; Verma et al., 2022; Keswani et al., 2021; Kerrigan et al., 2021; Hemmer et al., 2022; Benz & Rodriguez, 2022; Joshi et al., 2021; Raman & Yee, 2021). A substantial line of work develops surrogate losses and statistical guarantees for this problem (Mozannar & Sontag, 2020; Verma et al., 2022; Cao et al., 2024; Mozannar et al., 2023; Liu et al., 2024; Mao et al., 2024f; 2025d; Charusaie et al., 2022; Mao et al., 2024e; Wei et al., 2024; Montreuil et al., 2026b; Mao, 2025; Montreuil et al., 2026a). The framework has also been extended to regression and multi-task settings and deployed in real systems (Mao et al., 2024g; Strong et al., 2024; Palomba et al., 2025; Montreuil et al., 2025b; Strong et al., 2025; Montreuil et al., 2026d), as well as to robust (Montreuil et al., 2025a; 2026e), online (Montreuil et al., 2026f;c), budgeted or missing experts (DeSalvo et al., 2025; Nguyen et al., 2025), and populations of experts unseen at training time (Tailor et al., 2024; Strong et al., 2026) settings. Our consistency analysis builds on a broader line of work on hypothesis-set-specific consistency, initiated by Long & Servedio (2013); Zhang & Agarwal (2020) and later made quantitative through the $\mathcal{H}$-consistency bounds of Awasthi et al. (2022), with subsequent refinements and extensions to a wide range of losses (Mao et al., 2024b; 2023b; 2024h; 2025b; 2024c;d; Cortes et al., 2025a;b; Mao et al., 2025c;a; Zhong, 2025).

Recent work has also considered querying human feedback and feeding it to an enriched predictor rather than treating the human only as a deferred decision maker (Pugnana et al., 2025), in a binary ask/not-ask formulation for a single enriched predictor with realizable-consistency guarantees. We study a strictly more general setting in which the learner selects an executed *expert–advice pair* from multiple experts and advice sources, and establish both $\mathcal{H}$-consistency and Bayes consistency for the proposed augmented surrogate.

3. Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting a pair of random variables (X, Y) with joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space and $\mathcal{Y}$ is an arbitrary measurable output space. We denote by (x, y) a realization of (X, Y). Throughout, conditional expectations given $X = x$ denote fixed measurable versions, and all finite argmax and argmin operations use smallest-index tie-breaking.

Learning-to-Defer. We consider a fixed collection of J experts, indexed by $[J] \coloneqq \{1, \ldots, J\}$. Each expert $j \in [J]$ is associated with a measurable cost function $c_j : \mathcal{X} \times \mathcal{Y} \to [0, C]$, that quantifies the expense of routing (x, y) to expert j. This formulation is agnostic to the nature of the expert: an expert may be a predictive model, a language model system, a human annotator, or any other decision mechanism that induces measurable costs. We model expert prediction as a measurable map $e_j : \mathcal{X} \to \mathcal{Y}$ and decompose the cost as $c_j(x, y) \coloneqq \psi\big(e_j(x), y\big) + \beta_j$, where $\psi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a measurable task-specific loss (e.g., the 0–1 loss for classification) and $\beta_j \geq 0$ is an expert consultation cost (Madras et al., 2018; Mozannar & Sontag, 2020; Verma et al., 2022). We assume the decomposed costs are uniformly bounded by the same finite constant C above. We denote the tuple of expert predictors by $\mathbf{e} \coloneqq (e_1, \ldots, e_J)$.

Given a collection of fixed, pre-trained experts, L2D seeks a routing policy that assigns each instance to the most cost-effective expert (Madras et al., 2018; Mozannar & Sontag, 2020; Narasimhan et al., 2022; Mao et al., 2023a; 2024g; Montreuil et al., 2025b; 2026b). Let $\mathcal{H}_r \subseteq \{s_r : \mathcal{X} \times [J] \to \mathbb{R}\}$ denote a hypothesis class of measurable scoring functions for routing. Each $s_r \in \mathcal{H}_r$ induces a measurable router $r : \mathcal{X} \to [J]$ via $r(x) = \arg\max_{j \in [J]} s_r(x, j)$; we write $\mathbf{s}_r(x) = (s_r(x, j))_{j \in [J]}$ for the full score vector. The deferral loss incurred by routing a realization (x, y) through r is

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

and the population risk is its expectation $\mathcal{E}_{\ell_{\text{def}}}(r) \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[\ell_{\text{def}}(r;\, X, Y, \mathbf{e})\big]$.

$\mathcal{H}_r$-consistency. The indicator functions in ℓ_{def} make it discontinuous and non-differentiable, precluding direct gradient-based optimization (Zhang, 2002; Stein-

wart, 2007; Mozannar & Sontag, 2020). Standard practice replaces it with a convex surrogate Φ_{def} that operates on the scoring function s_r and is amenable to gradient-based methods. The surrogate deferral risk is $\mathcal{E}_{\Phi_{\text{def}}}(s_r) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\Phi_{\text{def}}(s_r;\, X, Y, \mathbf{e})]$, with best-in-class value $\mathcal{E}^*_{\Phi_{\text{def}}}(\mathcal{H}_r) := \inf_{s_r\in\mathcal{H}_r} \mathcal{E}_{\Phi_{\text{def}}}(s_r)$, and let $\mathcal{E}^B_{\ell_{\text{def}}}(\mathcal{H}_r) := \inf_{s_r\in\mathcal{H}_r} \mathcal{E}_{\ell_{\text{def}}}(r_{s_r})$ denote the best achievable true risk within $\mathcal{H}_r$, where r_{s_r} is the router induced by s_r. To obtain population-level excess-risk control, Awasthi et al. (2022) introduced $\mathcal{H}$-consistency bounds, subsequently refined and extended in (Mao et al., 2024b;h; 2025b;a).

Definition 1 ($\mathcal{H}_r$-consistency bound)**.** *The surrogate Φ_{def} satisfies an $\mathcal{H}_r$-consistency bound with respect to ℓ_{def} if there exists a non-decreasing function $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+$ with $\Gamma(0) = 0$ such that for all $s_r \in \mathcal{H}_r$,*

$$\begin{aligned}&\mathcal{E}_{\ell_{\text{def}}}(r) - \mathcal{E}^B_{\ell_{\text{def}}}(\mathcal{H}_r) + \mathcal{U}_{\ell_{\text{def}}}(\mathcal{H}_r)\\ &\quad\le \Gamma\Big(\mathcal{E}_{\Phi_{\text{def}}}(s_r) - \mathcal{E}^*_{\Phi_{\text{def}}}(\mathcal{H}_r) + \mathcal{U}_{\Phi_{\text{def}}}(\mathcal{H}_r)\Big),\end{aligned} \tag{2}$$

where r is the router induced by s_r. For a surrogate loss L depending on s_r, or a true loss evaluated at the router induced by s_r, the associated minimizability gap is $\mathcal{U}_L(\mathcal{H}_r) := \inf_{s_r\in\mathcal{H}_r} \mathcal{E}_L(s_r) - \mathbb{E}_X\left[\inf_{s_r\in\mathcal{H}_r} \mathbb{E}_{Y|X}[L(s_r;\, X, Y, \mathbf{e})]\right]$, applied here to $L \in \{\ell_{\text{def}}, \Phi_{\text{def}}\}$.

When $\mathcal{H}_r$ is rich enough for these gaps to vanish, the bound recovers Bayes consistency (Steinwart, 2007; Awasthi et al., 2022). In the standard L2D setting, such bounds have been established for the comp-sum family of cross-entropy surrogates (Mao et al., 2024e;g; 2023a; Montreuil et al., 2025b; 2026b), guaranteeing that surrogate minimization over a rich enough class recovers the Bayes-optimal router.

4. Problem Formulation and Surrogate Design

The deferral framework of Section 3 assumes that every expert acts on the input x alone. We now extend this setting by allowing the system to acquire additional information—advice—and provide it to the chosen expert before that expert acts. This introduces a genuine statistical complication: the policy must choose both the expert and the advice source from x alone, while the realized advice is revealed only after that commitment. We therefore first formalize the protocol and its Bayes rule, and then study which surrogate constructions remain valid in this extended setting.

4.1. Advice Sources

Advice acquisition introduces two coupled difficulties. First, the value of a source is expert-dependent: the same retrieved passage or tool call may help one expert and be irrelevant, or even harmful, for another (Guu et al., 2020; Lewis et al., 2020; Schick et al., 2023). The choice of *what* to acquire therefore cannot be assessed independently of the choice of *who* receives it. Second, both decisions must be made from x alone. The realized advice value is revealed only after the expert and advice source have already been selected. The difficulty is therefore not merely that the protocol is sequential, but that learning must target a hidden executed expert-advice pair rather than two independently observable decisions.

Formal setup. We formalize this uncertainty by augmenting the data-generating process with K advice variables. The triple (X, A, Y) is jointly distributed, where $A = (A^1, \ldots, A^K)$ and each A^k takes values in a measurable space $\mathcal{A}_k$. Let $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$, and write (x, a, y) for a realization. This model accommodates both deterministic advice, such as a database lookup, and stochastic advice, such as a language model output (Touvron et al., 2023; OpenAI et al., 2024). The action $k = 0$ denotes the decision to acquire no advice, and $[K]_0 := \{0, 1, \ldots, K\}$ is the resulting advice-action set.

Masking. The protocol enforces that at most one source is revealed per instance. Fix a missing-advice symbol $\perp \notin \bigcup_{k=1}^K \mathcal{A}_k$ and define $\widetilde{\mathcal{A}} := (\mathcal{A}_1 \cup \{\perp\}) \times \cdots \times (\mathcal{A}_K \cup \{\perp\})$. For each $k \in [K]_0$, the masking operator $m_k : \mathcal{A} \to \widetilde{\mathcal{A}}$ retains the k-th coordinate and replaces all others with $\perp$; when $k = 0$ every coordinate is masked. We write $\widetilde{a}^{(k)} := m_k(a)$ and $\widetilde{A}^{(k)} := m_k(A)$ for the masked advice.

4.2. The Deferral-Advice Loss

With advice sources and masking in place, we now define how the system acts and what it pays. The goal is to formalize the true loss of the protocol and identify the Bayes-optimal policy.

Costs. In the standard setting, routing to expert j incurs a cost $c_j(x, y)$ that captures both prediction error and consultation fee. With advice, the cost acquires a new dimension: the system may spend an acquisition fee γ_k to provide advice to the expert, hoping that the additional information reduces the expert's prediction error by more than it costs to obtain. For each pair $(j, k) \in [J] \times [K]_0$, let $c_{j,k} : \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \to [0, C]$ be a measurable cost function, uniformly bounded by $C < \infty$.

$$c_{j,k}(x, a, y) := \underbrace{\psi\big(e_j(x, \widetilde{a}^{(k)}),\, y\big)}_{\text{task loss}} + \underbrace{\beta_j}_{\text{expert fee}} + \underbrace{\gamma_k}_{\text{advice fee}}, \tag{3}$$

where $e_j : \mathcal{X} \times \widetilde{\mathcal{A}} \to \mathcal{Y}$ is measurable and extends the expert prediction of Section 3 to accept masked advice $\widetilde{a}^{(k)}$ (reducing to $e_j(x)$ when $k = 0$), and $\beta_j \ge 0$ is the expert consultation cost, while $\gamma_k \ge 0$ is the advice acquisition cost with $\gamma_0 = 0$. The no-advice compatibility condition $c_{j,0}(x, a, y) = c_j(x, y)$ holds by construction. The trade-off is explicit: acquiring advice k incurs the fee γ_k but may

reduce the task loss ψ if the advice is informative for expert j. Whether the trade-off is favorable depends jointly on the expert, the advice source, and the input.

Policy and true loss. A deferral-advice policy is a pair of measurable maps $r : \mathcal{X} \to [J]$ and $q : \mathcal{X} \times [J] \to [K]_0$. Let $\mathcal{H}_q$ be a hypothesis class of measurable query scoring functions $s_q : \mathcal{X} \times [J] \times [K]_0 \to \mathbb{R}$; each $s_q \in \mathcal{H}_q$ induces a query function via $q(x, j) = \arg\max_{k\in[K]_0} s_q(x, j, k)$. We write $\mathbf{S}_q(x) \in \mathbb{R}^{J\times(K+1)}$ for the matrix with entries $s_q(x, j, k)$, so that each row corresponds to one expert and the row-wise argmax gives the advice index for that expert. Upon observing x, the system routes to expert $j = r(x)$ and selects advice index $k = q(x, j)$; the masked advice $\widetilde{A}^{(k)}$ is then revealed to expert j. Although $q(x, \cdot)$ is defined for every expert, only the executed value $q(x, r(x))$ enters the realized loss (see Algorithm 2, Figure 2).

Definition 2 (True deferral-advice loss)**.** *Given a policy (r, q), the true deferral-advice loss on a realization (x, a, y) is*

$$\ell_{\text{def-adv}}(r, q;\, x, a, y, \mathbf{e}) \coloneqq \sum_{j\in[J]} \sum_{k\in[K]_0} c_{j,k}(x, a, y)\, \mathbf{1}\{r(x) = j\}\, \mathbf{1}\{q(x, j) = k\}.$$

The double indicator selects exactly one expert-advice pair per instance. The loss generalizes (1) from J to $J(K + 1)$ actions and depends on the policy only through the executed pair $(r(x),\, q(x, r(x)))$. Since $c_{j,0}(x, a, y) = c_j(x, y)$ by construction, setting $K = 0$ recovers the standard deferral loss ℓ_{def} (Mozannar & Sontag, 2020; Mao et al., 2023a; Montreuil et al., 2026b). A concrete numerical example illustrating the loss is given in Example 1.

4.3. Deferral with Advice Dominates Standard Deferral

We now derive the Bayes-optimal policy. Its structure reveals two properties: the choice of advice is inherently coupled to the choice of expert, and deferral with advice is more powerful than standard Learning-to-Defer.

The Bayes-optimal policy. In Section 4.1 we argued that the value of advice is expert-dependent, and the policy of Definition 2 reflects this by conditioning the query $q(x, j)$ on the expert. We now show that minimizing the expected loss of Definition 2 recovers the desired policy. To determine which expert is best, the system must first evaluate each expert at its best-advised cost—this requires solving the advice selection problem for every expert independently. The optimal query *necessarily* depends on the expert, since the best advice for expert j is generally different from the best advice for expert j'. The router then compares experts on their optimally-advised costs rather than their unaided ones. Write

$$m_{j,k}(x) \coloneqq \mathbb{E}[c_{j,k}(X, A, Y) \mid X = x]\,.$$

Lemma 3 (Bayes-optimal deferral-advice policy)**.** *There exists a measurable Bayes-optimal policy $(r^\star, q^\star)$ that minimizes $\mathbb{E}[\ell_{\text{def-adv}}(r, q;\, X, A, Y, \mathbf{e})]$ over all measurable policies and, for $\mathbb{P}_X$-a.e. x, can be chosen by smallest-index tie-breaking as follows:*

1. The Bayes query selects the best advice for each expert:

$$q^\star(x, j) = \operatorname{minarg\,min}_{k\in[K]_0} m_{j,k}(x). \tag{4}$$

2. The Bayes router selects the best optimally-advised expert:

$$r^\star(x) = \operatorname{minarg\,min}_{j\in[J]} \min_{k\in[K]_0} m_{j,k}(x). \tag{5}$$

The proof is given in Appendix B.1. The result reveals a fundamental coupling: the router does not compare experts on their unaided costs $\mathbb{E}[c_j(X, Y) \mid X = x]$ as in the standard setting, but on their optimally-advised ones. An expert that would never be selected under standard deferral may become optimal once paired with the right advice. The two decisions—who acts and what they see—cannot be separated.

When is advice worth acquiring? In any practical system, acquiring information has a cost, and a sensible policy should query only when the expected benefit justifies the expense. The smallest-index Bayes query of Lemma 3 satisfies exactly this principle. Under the cost decomposition (3), the system acquires advice $k \geq 1$ for expert j if and only if some queried source strictly beats the no-advice option in expected total cost.

Lemma 4 (Advice acquisition condition)**.** *Under the cost decomposition (3) and the smallest-index tie-breaking of Lemma 3, the Bayes-optimal query satisfies $q^\star(x, j) \neq 0$ if and only if there exists $k \in [K]$ such that*

$$\begin{aligned}&\mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(0)}), Y) \mid X = x\big] \\ &\quad - \mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(k)}), Y) \mid X = x\big] > \gamma_k.\end{aligned} \tag{6}$$

When no source satisfies this condition, $q^\star(x, j) = 0$ and the expert acts on x alone.

The proof is given in Appendix B.2. The left-hand side of (6) measures how much advice k improves expert j's prediction; the right-hand side is the price of obtaining it. The system acquires advice only when it is expected to pay for itself.

Deferral with advice dominates L2D. Since the no-advice action k=0 is always available, the learner can recover the standard L2D decision whenever advice is not worth its cost. The point is stronger: for each expert the best advised

cost is at most the unaided cost; propagating this through the comparison over experts yields a strict population-level improvement whenever advice lowers the best achievable conditional cost on a set of positive measure.

Lemma 5 (Deferral with advice dominates standard deferral)**.** *If $\mathcal{E}^\star_{\ell_{\text{def}}}$ denotes the Bayes risk of the standard deferral setting of Section 3, then*

$$\mathcal{E}^\star_{\ell_{\text{def-adv}}} \;\le\; \mathcal{E}^\star_{\ell_{\text{def}}}. \tag{7}$$

The proof, including a necessary and sufficient condition for strict inequality, is given in Appendix B.3.

4.4. The Separated Surrogate Fails

Because the protocol is sequential, the natural first attempt is to learn it sequentially as well: one router score, one query head per expert, and smooth logistic replacements for the hard indicators (Bartlett et al., 2006). This is precisely the construction a learning-theory practitioner would derive by applying the standard recipe—replacing each indicator in the true loss by a smooth convex surrogate term—to the product of routing and query decisions. But does mirroring the control flow suffice for consistency? We show that a broad family of separated surrogates already fails in the smallest non-trivial case: two experts ($J = 2$) and one binary advice ($K = 1$).

Let $s^b_r : \mathcal{X} \to \mathbb{R}$ be a binary router score and, for each expert $j \in \{1, 2\}$, let $s_{q^b_j} : \mathcal{X} \to \mathbb{R}$ be a binary query score. Writing $\mathbf{s}_{q^b} = (s_{q^b_1}, s_{q^b_2})$, the induced decisions are $r^b(x) = 1 + \mathbf{1}\{s^b_r(x) \ge 0\}$ and $q^b_j(x) = \mathbf{1}\{s_{q^b_j}(x) \ge 0\}$.

The natural separated construction. In standard learning theory, one often replaces hard indicators by smooth convex surrogates (Zhang, 2002; Bartlett et al., 2006; Steinwart, 2007). Applying this recipe to the binary executed-pair loss yields logistic potentials for the decoded binary outcomes, $\Phi_0(u) = \log(1 + e^{-u})$ and $\Phi_1(u) = \log(1 + e^u)$. Rather than analyzing only the basic surrogate obtained by directly substituting these terms, we study a strictly larger family that encompasses every natural modification a practitioner might attempt: monotone cost transforms, reparameterized query scores, and flexible positive router weights. Let $\nu : [0, C] \to [0, \infty)$ be a monotone cost transform, $G : \mathbb{R} \to (0, \infty)$ a query amplitude, and $U : \mathbb{R} \to \mathbb{R}$ a strictly increasing query reparameterization. The resulting *router-query separated surrogate* family is

$$\Phi^{r/q}_{\text{def-adv}}(s^b_r, \mathbf{s}_{q^b}) = \sum_{j=1}^{2}\sum_{k=0}^{1} \Psi_j(s^b_r(x))\nu(c_{j,k}(x, a, y)) \cdot G(s_{q^b_j}(x))\, \Phi_k(U(s_{q^b_j}(x))). \tag{8}$$

Setting $\nu = \text{id}$, $G \equiv 1$, $U = \text{id}$, and $\Psi_j = \Phi_{j-1}$ recovers the basic construction obtained by assigning these logistic potentials to the corresponding decoded binary outcomes—the standard recipe in cost-sensitive surrogate design (Cortes et al., 2016; 2024a); we give the full derivation in Appendix B.4.1. The family (8) subsumes every natural modification a practitioner might attempt to rescue the separated architecture—monotone cost transforms, reparameterized query scores, flexible positive router weights—so the negative result below is not an artifact of a specific parameterization but of the separated structure itself.

Where the mismatch enters. This separation has one decisive consequence. After profiling over the query heads, each expert is summarized by a scalar quantity

$$F(u, v) := \inf_{t\in\mathbb{R}} \big[\nu(u)\, G(t)\, \Phi_0(U(t)) + \nu(v)\, G(t)\, \Phi_1(U(t))\big] \tag{9}$$

that depends on the *entire* row, not only on its best entry. The Bayes rule, by contrast, compares row minima: $\min\{c_{1,0}, c_{1,1}\}$ versus $\min\{c_{2,0}, c_{2,1}\}$ (cf. (5)). The surrogate and the Bayes rule compress each row in fundamentally different ways. That difference is the entire obstruction.

We formalize this failure at the level of Bayes consistency: every population surrogate minimizer should decode to a Bayes-optimal action (Zhang, 2002; Bartlett et al., 2006; Steinwart, 2007; Tewari & Bartlett, 2007).

Theorem 6 (Broad Bayes inconsistency of separated router/query surrogates)**.** *Under mild regularity conditions on ν, G, U, Ψ_1, Ψ_2 (Appendix B.4), for every $b \in (0, C)$ and $\varepsilon \in (0, C - b)$, there exists $\delta \in (0, b)$ such that the pointwise conditional cost table*

$$\bar{\mathbf{C}}(x) \;=\; \begin{pmatrix} b-\delta & C \\ b & b+\varepsilon \end{pmatrix} \tag{10}$$

has Bayes-optimal expert $j^\star = 1$, while the unique minimizer of the pointwise surrogate (8) decodes to expert 2. In particular, the entire family (8) is not Bayes consistent for the true loss.

Expert 1 has one excellent action and one catastrophic one, whereas expert 2 has two mediocre actions. Bayes compares row minima, ignores the catastrophic entry, and therefore prefers expert 1. The separated surrogate behaves differently. After profiling out the query head, row 1 is summarized by $F(b - \delta, C)$, which still depends on the catastrophic entry. Appendix B.4.4 shows that $F(u, v)$ is strictly increasing in v for fixed $u \in (0, C)$, so this distortion is unavoidable. The router therefore compares profiled row summaries rather than row minima and selects expert 2.

The contradiction already appears in the smallest non-trivial setting (J=2, K=1) and persists across the entire parameterization family (8). Because this family encompasses every modification the standard surrogate-design recipe suggests, the result rules out the entire class of constructions

a practitioner would naturally derive from the product-of-indicators loss. Appendix C.2 complements the theorem with a synthetic experiment where the separated surrogate's excess risk stays bounded away from zero.

4.5. The Augmented Surrogate

The separated surrogate of Section 4.4 fails because profiling the query heads row by row distorts the router's comparison. This raises the central question of whether the deferral-advice problem admits a calibrated surrogate at all, given that the separated construction is already inconsistent. We answer this question in the affirmative.

Composite actions. Recall that the loss depends only on the executed pair $(r(x),\, q(x, r(x)))$. Define the composite action set $\Pi := [J] \times [K]_0$, of cardinality $|\Pi| = J(K+1)$. Each action $(j,k) \in \Pi$ represents the joint decision to route to expert j and acquire advice k. A composite predictor is a measurable map $\pi : \mathcal{X} \to \Pi$. Let $\mathcal{H}_\pi \subseteq \{s_\pi : \mathcal{X} \times \Pi \to \mathbb{R}\}$ be a hypothesis class of measurable scoring functions; each $s_\pi \in \mathcal{H}_\pi$ induces a composite predictor via $\pi(x) = \arg\max_{(j,k)\in\Pi} s_\pi(x,(j,k))$. We write $\mathbf{s}_\pi(x) = (s_\pi(x,(j,k)))_{(j,k)\in\Pi}$ for the full score vector.

A sequential policy (r,q) induces the composite action $\pi(x) = (r(x),\, q(x,r(x)))$, and every composite action (j,k) can in turn be implemented by a sequential policy with the same realized loss. The composite view simply makes explicit what the loss has depended on all along: the executed expert-advice pair. Proposition 17, proved in Appendix B.5, shows that the sequential and composite formulations therefore induce the same population risks and share the same Bayes-optimal decisions as those identified in Lemma 3.

The augmented surrogate. The composite view aligns the learning problem with the Bayes comparison itself, but the true loss remains non-differentiable. The deferral-advice loss is not a uniform 0–1 loss over Π: it is cost-sensitive, with instance- and action-dependent weights. The key step (Appendix B.6) decomposes the realized loss into an action-independent offset and a weighted mismatch term:

$$\ell_{\text{def-adv}}(\pi;\, x,a,y,\mathbf{e}) = D(x,a,y) + \sum_{i\in\Pi} w_i(x,a,y)\, \mathbf{1}\{\pi(x) \neq i\}, \tag{11}$$

where $D(x,a,y)$ is independent of π and $w_i(x,a,y) := \max_{i'\in\Pi} c_{i'}(x,a,y) - c_i(x,a,y)$. Let Φ_{01}^τ denote the comp-sum multiclass surrogate of Mao et al. (2023b), parameterized by $\tau \geq 0$; at $\tau = 1$ it recovers the standard logistic (cross-entropy) loss. Applying Φ_{01}^τ to the weighted mismatch term yields the augmented deferral-advice surrogate, which we derive in Appendix B.7:

Lemma 7 (Augmented surrogate)**.**

$$\Phi_{\text{def-adv}}^{\text{aug},\tau}(s_\pi;\, x,a,y,\mathbf{e}) := \sum_{i\in\Pi} w_i(x,a,y)\; \Phi_{01}^\tau(\mathbf{s}_\pi(x),\, i). \tag{12}$$

Consistency: the main positive result. Lemma 7 gives the right starting point, but it is not the whole story. Simply replacing the mismatch indicators by a smooth loss is not enough. The separated construction of Section 4.4 is precisely a counterexample: despite its natural form, Theorem 6 shows that its minimizers can still decode to the wrong executed decision. What matters is whether the surrogate preserves the comparison that the protocol actually makes.

The augmented surrogate succeeds where the separated one fails because it never profiles out a subset of scores. The separated construction optimizes query scores row by row and hands the router a scalar summary per expert; that summary depends on the non-minimal entry and distorts the comparison (Theorem 6). The augmented surrogate, by contrast, assigns one score to each composite action (j,k) and compares them all at once. No row-by-row reduction ever occurs, so the surrogate comparison operates on the same objects as the Bayes comparison.

We now establish this intuition formally: the augmented surrogate admits a quantitative excess-risk transfer bound for the entire comp-sum family and hence for every $\tau \geq 0$.

Theorem 8 ($\mathcal{H}_\pi$-consistency of the augmented surrogate)**.** *Assume that $\mathcal{H}_\pi$ satisfies the standard symmetry and completeness conditions required by the comp-sum multiclass $\mathcal{H}_\pi$-consistency theorem for Φ_{01}^τ on the label set Π. For every $\tau \geq 0$, the augmented surrogate $\Phi_{\text{def-adv}}^{\text{aug},\tau}$ satisfies an $\mathcal{H}_\pi$-consistency bound with respect to $\ell_{\text{def-adv}}$. Specifically, for every $s_\pi \in \mathcal{H}_\pi$ and its induced composite predictor π, write $\Phi_{\text{aug}}^\tau := \Phi_{\text{def-adv}}^{\text{aug},\tau}$. Define*

$$\Delta_\ell(\pi) := \mathcal{E}_{\ell_{\text{def-adv}}}(\pi) - \mathcal{E}^B_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) + \mathcal{U}_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi),$$
$$\Delta_\Phi(s_\pi) := \mathcal{E}_{\Phi_{\text{aug}}^\tau}(s_\pi) - \mathcal{E}^*_{\Phi_{\text{aug}}^\tau}(\mathcal{H}_\pi) + \mathcal{U}_{\Phi_{\text{aug}}^\tau}(\mathcal{H}_\pi).$$

Then

$$\Delta_\ell(\pi) \leq \widetilde{\Gamma}_\tau(\Delta_\Phi(s_\pi))\,. \tag{13}$$

In the non-degenerate regime $\mathbb{E}[\|\mathbf{w}(X)\|_1] > 0$, $\widetilde{\Gamma}_\tau(u) := \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \Gamma_\tau\left(\dfrac{u}{\mathbb{E}[\|\mathbf{w}(X)\|_1]}\right)$, where Γ_τ is the non-negative, non-decreasing, concave transfer function for Φ_{01}^τ on Π, and $\mathbf{w}(x) := (\mathbb{E}[w_i(X,A,Y) \mid X = x])_{i\in\Pi}$.

We prove this theorem in Appendix B.8. For the standard log-softmax (cross-entropy) surrogate ($\tau = 1$), $\Gamma_1(u) = \sqrt{2u}$, so the true excess risk decays as $O\big(\sqrt{\mathbb{E}[\|\mathbf{w}(X)\|_1]\,\varepsilon}\,\big)$ in the surrogate excess risk ε, at the standard calibration

rate (Mao et al., 2023b; 2024h). The proof is not a direct application of standard multiclass consistency. The weights $w_i(x, a, y)$ are both instance- and action-dependent, so the surrogate defines a *weighted* cost-sensitive problem rather than a uniform one. The technical core is a reweighting argument that absorbs these weights into an auxiliary distribution $\widetilde{\mathcal{D}}$ on $\mathcal{X} \times \Pi$ (Lemma 19 in the appendix). The transfer function $\widetilde{\Gamma}_\tau$ that emerges is problem-adaptive: it depends on the expected weight norm $\mathbb{E}[\|\mathbf{w}(X)\|_1]$, which encodes the cost structure of the particular expert-advice environment. When no advice is available ($K = 0$, so $\Pi = [J]$), the bound (13) recovers the $\mathcal{H}_r$-consistency bounds established for standard Learning-to-Defer by Mao et al. (2023a; 2024g); Montreuil et al. (2025b; 2026b).

Corollary 9 (Bayes consistency)**.** *Assume the hypotheses of Theorem 8. If the minimizability gaps in Theorem 8 vanish, the class is Bayes-rich in the sense that* $\mathcal{E}^B_{\ell_{\text{def-ady}}}(\mathcal{H}_\pi) = \inf_\pi \mathcal{E}_{\ell_{\text{def-adv}}}(\pi)$*, and* $\mathcal{E}_{\Phi^{\text{aug},\tau}_{\text{def-adv}}}(\widehat{s}_{\pi,n}) - \mathcal{E}^*_{\Phi^{\text{aug},\tau}_{\text{def-adv}}}(\mathcal{H}_\pi) \to 0$ *in probability, then the induced predictors* $\widehat{\pi}_n$ *satisfy* $\mathcal{E}_{\ell_{\text{def-adv}}}(\widehat{\pi}_n) - \inf_\pi \mathcal{E}_{\ell_{\text{def-adv}}}(\pi) \to 0$ *in probability.*

Thus, whenever the hypothesis class is rich enough and the surrogate is well-optimized, the induced policy converges in risk to the Bayes-optimal deferral-advice risk of Lemma 3. When $\mathcal{H}_\pi$ contains all measurable functions, both minimizability gaps vanish (Zhang & Agarwal, 2020; Bartlett et al., 2006; Long & Servedio, 2013) and the bound reduces to classical Bayes-consistency. The separated surrogate offers no such guarantee.

5. Experiments

We evaluate the augmented surrogate on four benchmarks spanning language, tabular, multi-modal, and synthetic domains. **(i) FEVER** (Thorne et al., 2018): fact verification with LLM and NLI experts receiving retrieved evidence (detailed below). **(ii) Sensitive escalation** (Howard et al., 2019): fraud detection where advice reveals progressively more sensitive feature groups (Appendix C.4). **(iii) CLIP prompt escalation** (Radford et al., 2021): zero-shot image classification where advice enlarges the prompt family (Appendix C.5). **(iv) Synthetic**: a theorem-aligned benchmark that isolates the failure mode of separated surrogates predicted by Theorem 6 (Appendix C.2). The separated surrogate of Section 4.4 is defined only for binary routing and binary advice (J=2, K=1); it is therefore evaluated only in the synthetic setting.

5.1. FEVER setup

Task, experts, and advice. We use the FEVER fact-verification benchmark (Thorne et al., 2018). The expert pool contains Qwen3-4B-Instruct, Qwen3-8B (Yang et al., 2025), and DeBERTa-v3-large (He et al., 2021), with consultation costs $\beta = (0.03, 0.05, 0.04)$. Advice consists of retrieved evidence. The five actions are no retrieval, BM25 top-1, top-3, top-5, and a query-reformulation pipeline produced by Qwen2.5-1.5B. The base advice costs are $\gamma^{\text{base}} = (0, 0.015, 0.02, 0.03, 0.01)$, and the deployed cost is $\gamma_k = \lambda\gamma_k^{\text{base}}$ for $\lambda \in \{0, 0.5, 1, 2, 4, 6, 8, 10\}$.

Training protocol and baselines. The composite scorer is an MLP trained with AdamW; full architecture and hyperparameter details are in Appendix C.3. We evaluate the empirical validation average of $\mathcal{E}_{\ell_{\text{def-adv}}}(\pi)$. We compare against four baselines, each isolating a different aspect of the decision: (i) standard L2D (Mao et al., 2023a), restricted to k=0, which can route but never acquire advice; (ii) the best fixed expert–advice pair, an oracle over non-adaptive policies; (iii) two partial-randomization ablations — learned expert with random advice, and learned advice with random expert — that separate the contribution of each decision component; and (iv) a fully random (j, k) baseline as an uninformed floor.

5.2. Results

The value of advice is strongly expert-dependent: on this three-class task, without retrieval all three experts hover around 33% accuracy, but their preferred retrieval actions diverge—DeBERTa is best with top-1, whereas both LLMs prefer top-5 (full per-expert table in Appendix C.3).

Table 1 gives the main result. Our method achieves the lowest mean true loss in seven of the eight cost regimes, with the largest gain at $\lambda = 0$ (0.064 over L2D); as advice cost grows the policy shifts toward no-query decisions, and at $\lambda = 10$ it gracefully recovers the L2D baseline. It also stays below the best fixed expert–advice pair throughout, confirming that the gain stems from input-dependent selection rather than from mere access to a larger action set. The partial-randomization ablations further show that both decision halves matter, and that *Random j, learned k* beats fully random — evidence that the surrogate learns a genuinely expert-conditional advice strategy rather than a global retrieval preference.

5.3. Additional benchmarks

Sensitive escalation (tabular). On the fraud-detection benchmark (Howard et al., 2019), advice reveals progressively more sensitive feature groups to one of three experts (logistic regression, GBDT, MLP). Our method achieves the lowest true loss at every cost multiplier (Table 11 in Appendix C.4), improving over L2D by 0.022 at λ=0 and converging to it at λ=5. The best fixed pair itself changes with the cost regime, confirming that the optimal policy is inherently cost-dependent.

CLIP prompt escalation (multi-modal). On ImageNet-1k (Russakovsky et al., 2015), advice enlarges the prompt family used by frozen CLIP experts (B/32, B/16, L/14) from

Table 1. Validation true deferral-advice loss across advice-cost multipliers λ. Entries are mean $\pm$ std over 4 seeds; lower is better. Bold marks the best mean per row.

λ	Ours	L2D	Best fixed	Learned j, rand. k	Rand. j, learned k	Rand. (j, k)
0	$\mathbf{0.555 \pm 0.002}$	0.619 ± 0.001	0.611	0.633 ± 0.005	0.640 ± 0.007	0.675 ± 0.006
0.5	$\mathbf{0.558 \pm 0.002}$	0.619 ± 0.001	0.619	0.640 ± 0.005	0.657 ± 0.007	0.682 ± 0.006
1	$\mathbf{0.566 \pm 0.001}$	0.619 ± 0.001	0.626	0.647 ± 0.007	0.664 ± 0.007	0.690 ± 0.006
2	$\mathbf{0.580 \pm 0.001}$	0.619 ± 0.001	0.641	0.662 ± 0.006	0.674 ± 0.009	0.705 ± 0.006
4	$\mathbf{0.596 \pm 0.003}$	0.619 ± 0.001	0.671	0.695 ± 0.006	0.695 ± 0.007	0.735 ± 0.007
6	$\mathbf{0.607 \pm 0.002}$	0.619 ± 0.001	0.695	0.723 ± 0.006	0.704 ± 0.007	0.765 ± 0.007
8	$\mathbf{0.613 \pm 0.001}$	0.619 ± 0.001	0.695	0.760 ± 0.006	0.711 ± 0.006	0.795 ± 0.007
10	0.619 ± 0.001	0.619 ± 0.001	0.695	0.788 ± 0.006	0.715 ± 0.008	0.825 ± 0.007

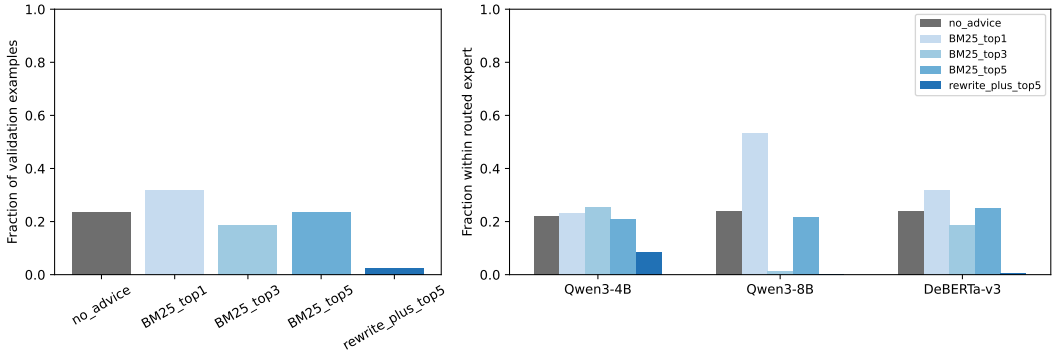

Figure 1. Advice usage learned by the augmented policy on FEVER when advice is free ($\lambda = 0$). The policy queries frequently, but the right panel shows that the preferred advice action remains expert-dependent rather than collapsing to one global retrieval depth.

a single template to ensembles of up to 32. Our method attains the lowest true loss across the entire low- and moderate-cost regime (Table 16 in Appendix C.5), reducing L2D's 0.352 to 0.337 at λ=0 and matching L2D at λ=10. The three experts prefer different prompt families at different cost levels, precisely the regime where a joint expert-advice policy is most natural.

Synthetic (theorem-aligned). A binary-setting benchmark (J=2, K=1) instantiates the cost tables from Theorem 6. At n=5000, our method reaches true risk 0.281, essentially matching the Bayes optimum 0.280, while the separated surrogate stalls at 0.334 and L2D at 0.342 (Table 4 in Appendix C.2). Regionwise, the separated surrogate misroutes on the theorem region (18% Bayes-action match), L2D cannot realize the queried Bayes action on the advice-helpful region (0%), and our method achieves 99.3% on both.

6. Conclusion

We introduced *Learning-to-Defer with advice*, a setting in which the system must decide not only which expert should act, but also which additional information that expert should receive. A broad family of natural separated router-query surrogates is not Bayes consistent even in the smallest non-trivial case; we gave an augmented surrogate that learns directly over the composite expert-advice action space and proved an $\mathcal{H}$-consistency guarantee. Empirically, the resulting method improves over Learning-to-Defer across language, tabular, and multi-modal benchmarks while adapting its advice acquisition to the cost regime. Once information acquisition is expert dependent, Learning-to-Defer must reason not only about *who* acts, but also about *what they get to see*.

References

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Multi-class h-consistency bounds. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

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

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

Benz, N. L. C. and Rodriguez, M. G. Counterfactual in-

ference of second opinions. In *Uncertainty in Artificial Intelligence*, pp. 453–463. PMLR, 2022.

Cao, Y., Cai, T., Feng, L., Gu, L., Gu, J., An, B., Niu, G., and Sugiyama, M. Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in neural information processing systems*, 35:521–534, 2022.

Cao, Y., Mozannar, H., Feng, L., Wei, H., and An, B. 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.

Charusaie, M.-A., Mozannar, H., Sontag, D., and Samadi, S. Sample efficient learning of predictors that complement humans, 2022.

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

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In Ortner, R., Simon, H. U., and Zilles, S. (eds.), *Algorithmic Learning Theory*, pp. 67–82, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

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

Cortes, C., Mao, A., Mohri, C., Mohri, M., and Zhong, Y. Cardinality-aware set prediction and top-k classification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=WAT3qu737X.

Cortes, C., Mao, A., Mohri, M., and Zhong, Y. Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025a.

Cortes, C., Mao, A., Mohri, M., and Zhong, Y. Improved balanced classification with theoretically grounded loss functions. In *Advances in Neural Information Processing Systems 39 (NeurIPS 2025)*, 2025b.

DeSalvo, G., Mohri, C., Mohri, M., and Zhong, Y. Budgeted multiple-expert deferral, 2025. URL https://arxiv.org/abs/2510.26706.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (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.

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 1919–1925. AAAI Press, 2017.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

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

Hemmer, P., Schellhammer, S., Vössing, M., Jakubik, J., and Satzger, G. Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. In Raedt, L. D. (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.

Howard, A., Bouchon-Meunier, B., CIS, I., inversion, Lei, J., Lynn@Vesta, Marcus2010, and Abbass, P. H. Ieee-cis fraud detection. https://kaggle.com/competitions/ieee-fraud-detection, 2019. Kaggle.

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

Kerrigan, G., Smyth, P., and Steyvers, M. Combining human predictions with model probabilities via confusion matrices and calibration. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (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.

Keswani, V., Lease, M., and Kenthapadi, K. 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.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Liu, S., Cao, Y., Zhang, Q., Feng, L., and An, B. Mitigating underfitting in learning to defer with consistent losses. In *International Conference on Artificial Intelligence and Statistics*, pp. 4816–4824. PMLR, 2024.

Long, P. and Servedio, R. Consistency versus realizable h-consistency for multiclass classification. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, number 3 in Proceedings of Machine Learning Research, pp. 801–809, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/long13.html.

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

Mao, A. *Theory and Algorithms for Learning with Multi-Class Abstention and Multi-Expert Deferral*. PhD thesis, New York University, 2025. Courant Institute of Mathematical Sciences; arXiv:2512.22886.

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

Mao, A., Mohri, M., and Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pp. 23803–23828. PMLR, 2023b.

Mao, A., Mohri, M., and Zhong, Y. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, pp. 822–867. PMLR, 2024a.

Mao, A., Mohri, M., and Zhong, Y. h-consistency bounds: Characterization and extensions. *Advances in Neural Information Processing Systems*, 36, 2024b.

Mao, A., Mohri, M., and Zhong, Y. $\mathcal{H}$-consistency guarantees for regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34712–34737. PMLR, 2024c.

Mao, A., Mohri, M., and Zhong, Y. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024d.

Mao, A., Mohri, M., and Zhong, Y. Principled approaches for learning to defer with multiple experts. In *ISAIM*, 2024e.

Mao, A., Mohri, M., and Zhong, Y. Realizable h-consistent and bayes-consistent loss functions for learning to defer. *Advances in neural information processing systems*, 37: 73638–73671, 2024f.

Mao, A., Mohri, M., and Zhong, Y. Regression with multi-expert deferral. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024g.

Mao, A., Mohri, M., and Zhong, Y. A universal growth rate for learning with smooth surrogate losses. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024h.

Mao, A., Mohri, M., and Zhong, Y. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In Dasgupta, S., Mandt, S., and Li, Y. (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 2024i. URL https://proceedings.mlr.press/v238/mao24a.html.

Mao, A., Mohri, M., and Zhong, Y. Beyond tsybakov: Model margin noise and $\mathcal{H}$-consistency bounds, 2025a. URL https://arxiv.org/abs/2511.15816.

Mao, A., Mohri, M., and Zhong, Y. Enhanced $\mathcal{H}$-consistency bounds. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory (ALT 2025)*, Proceedings of Machine Learning Research. PMLR, 2025b.

Mao, A., Mohri, M., and Zhong, Y. Principled algorithms for optimizing generalized metrics in binary classification. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025c.

Mao, A., Mohri, M., and Zhong, Y. Mastering multiple-expert routing: Realizable h-consistency and strong guarantees for learning to defer. In *Forty-second International Conference on Machine Learning*, 2025d.

Montreuil, Y., Carlier, A., Ng, L. X., and Ooi, W. T. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *Forty-second International Conference on Machine Learning*, 2025a.

Montreuil, Y., Heng, Y. S., Carlier, A., Ng, L. X., and Ooi, W. T. A two-stage learning-to-defer approach for multi-task learning. In *Forty-second International Conference on Machine Learning*, 2025b.

Montreuil, Y., Carlier, A., Ng, L. X., and Ooi, W. T. Beyond augmented-action surrogates for multi-expert learning-to-defer, 2026a. URL https://arxiv.org/abs/2604.09414.

Montreuil, Y., Carlier, A., Ng, L. X., and Ooi, W. T. 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.

Montreuil, Y., Dang, H. D., Meyer, M., Ng, L. X., Carlier, A., and Ooi, W. T. 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.

Montreuil, Y., Heng, Y. S., Ng, L. X., Carlier, A., and Ooi, W. T. 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.

Montreuil, Y., Ng, L. X., Carlier, A., Ooi, W. T., and Letian, Y. 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.

Montreuil, Y., Yu, L., Carlier, A., Ng, L. X., and Ooi, W. T. Learning to defer in non-stationary time series via switching state-space models. *arXiv preprint arXiv:2601.22538*, 2026f.

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

Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., and Sontag, D. A. 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.

Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A., and Kumar, S. Post-hoc estimators for learning to defer to an expert. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 29292–29304. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bc8f76d9caadd48f77025b1c889d2e2d-Paper-Conference.pdf.

Nguyen, C. C., Do, T.-T., and Carneiro, G. Probabilistic learning to defer: Handling missing expert annotations and controlling workload distribution. In *The Thirteenth International Conference on Learning Representations*, 2025.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O.,

Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Palomba, F., Pugnana, A., Alvarez, J. M., and Ruggieri, S. 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.

Pugnana, A., De Toni, G., Barbera, C., Pellungrini, R., Lepri, B., and Passerini, A. To ask or not to ask: Learning to require human feedback. *arXiv preprint arXiv:2510.08314*, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Robertson, S. and Zaragoza, H. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc, 2009.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

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

Strong, J., Men, Q., and Noble, A. 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.

Strong, J., Men, Q., and Noble, J. A. 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.

Strong, J., Saha, P., Ibrahim, Y., Ouyang, C., and Noble, A. Identity-free deferral for unseen experts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4YG9ufFg58.

Tailor, D., Patra, A., Verma, R., Manggala, P., and Nalisnick, E. Learning to defer to a population: A meta-learning approach. In Dasgupta, S., Mandt, S., and Li, Y. (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.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007. URL http://jmlr.org/papers/v8/tewari07a.html.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. 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.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Verma, R., Barrejon, D., and Nalisnick, E. 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.

Wei, Z., Cao, Y., and Feng, L. Exploiting human-ai dependence for learning to defer. In *Forty-first International Conference on Machine Learning*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhang, M. and Agarwal, S. Bayes consistency vs. h-consistency: The interplay between surrogate loss functions and the scoring function class. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16927–16936. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c4c28b367e14df88993ad475dedf6b77-Paper.pdf.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32, 12 2002. doi: 10.1214/aos/1079120130.

Zhong, Y. *Fundamental Novel Consistency Theory: $\mathcal{H}$-Consistency Bounds*. PhD thesis, New York University, 2025. Courant Institute of Mathematical Sciences.

Appendix

Contents

A. Algorithms, Visualization, Example

A.1. Notation Summary

Table 2 collects the main symbols used throughout the paper. We keep only the objects that are central to the formulation and theory.

A.2. Visualization

A.3. Worked Example of the True Loss

The following example traces the full evaluation of the deferral-advice loss on a single realization, starting from the scoring functions and ending with the scalar loss value.

Example 1 (Evaluating the true loss on one instance)**.** *Consider $J = 3$ experts and $K = 2$ advice sources, so the advice index set is $[K]_0 = \{0, 1, 2\}$.*

Scores. *Suppose the router and query scoring functions produce, for a given input x,*

$$\mathbf{s}_r(x) = (0.8,\ 1.2,\ 0.5), \qquad \mathbf{S}_q(x) = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.9 & 0.4 \\ 0.5 & 0.1 & 0.7 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

Each row of $\mathbf{S}_q(x)$ corresponds to one expert; each column to one advice action ($k = 0, 1, 2$).

Policy decisions. *The router selects the expert with the highest routing score: $r(x) = \arg\max_{j\in[3]} s_r(x, j) = 2$. The query function computes one advice index per expert by taking the row-wise argmax of $\mathbf{S}_q(x)$:*

$$q(x, 1) = 0, \qquad q(x, 2) = 1, \qquad q(x, 3) = 2.$$

Only the executed query matters: $k^\star = q(x, r(x)) = q(x, 2) = 1$.

Realized costs. *For this realization (x, a, y), suppose the full cost table is*

$$\big(c_{j,k}(x, a, y)\big)_{j\in[3],\, k\in\{0,1,2\}} = \begin{pmatrix} 0.35 & 0.42 & 0.38 \\ 0.40 & 0.20 & 0.45 \\ 0.50 & 0.48 & 0.25 \end{pmatrix}.$$

Each entry includes task loss, expert fee, and advice fee.

Loss evaluation. *The double indicator in Definition 2 selects the executed pair $(j, k) = (2, 1)$:*

$$\ell_{\text{def-adv}}(r, q;\, x, a, y, \mathbf{e}) \;=\; c_{2,1}(x, a, y) \;=\; 0.20.$$

The counterfactual query decisions $q(x, 1) = 0$ and $q(x, 3) = 2$ play no role: the protocol only pays for the executed pair. The chosen pair is also the realized minimum of the cost table: comparing the row-wise minima $(0.35, 0.20, 0.25)$, expert 2 with advice $k = 1$ achieves the smallest cost. This is exactly the comparison performed by the Bayes policy in Lemma 3, applied at the level of one realization.

A.4. Algorithms

Algorithms 1, 3, and 2 summarize the two decision rules discussed in the paper. The first two are the proposed augmented approach: training from fully observed executed-pair costs, and deployment-time execution of the learned composite policy. Algorithm 2 then records the inference rule of the separated router/query parameterization, which is the natural sequential baseline analyzed in Section 4.4.

Algorithm 1 Training with the augmented surrogate

Require: Training set $\{(x_n, a_n, y_n)\}_{n=1}^{N}$, composite action set $\Pi = [J] \times [K]_0$, score model $s_\pi(\cdot, \cdot; \theta)$, comp-sum parameter $\tau \geq 0$, learning rate η

Initialize parameters θ
repeat
  Sample a minibatch $B \subseteq \{1, \ldots, N\}$
  for all $n \in B$ **do**
    for all $i = (j, k) \in \Pi$ **do**
      Compute the realized cost $c_i(x_n, a_n, y_n) \equiv c_{j,k}(x_n, a_n, y_n)$
    end for
    Set $w_i(x_n, a_n, y_n) \leftarrow \max_{i' \in \Pi} c_{i'}(x_n, a_n, y_n) - c_i(x_n, a_n, y_n)$ for all $i \in \Pi$
    Form the sample loss

$$L_n(\theta) \leftarrow \sum_{i \in \Pi} w_i(x_n, a_n, y_n)\, \Phi_{01}^{\tau}(\mathbf{s}_\pi(x_n; \theta), i)$$

  end for
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \left(|B|^{-1} \sum_{n \in B} L_n(\theta) \right)$
until convergence
return learned parameters $\widehat{\theta}$

Algorithm 2 Inference with a separated router/query policy

Require: Input x, router score model $s_r(\cdot, \cdot)$, query score model $s_q(\cdot, \cdot, \cdot)$

1: Form the query-score matrix

$$\mathbf{S}_q(x) \leftarrow \big(s_q(x, j, k)\big)_{j \in [J],\, k \in [K]_0} \in \mathbb{R}^{J \times (K+1)}$$

2: For each expert j, take the argmax along row j of $\mathbf{S}_q(x)$:

$$\widehat{k}_j(x) \in \operatorname*{arg\,max}_{k \in [K]_0} s_q(x, j, k)$$

3: Compute the routed expert

$$\widehat{j}(x) \in \operatorname*{arg\,max}_{j \in [J]} s_r(x, j)$$

4: Select the advice actually executed by the routed expert

$$\widehat{k}(x) \leftarrow \widehat{k}_{\widehat{j}(x)}(x)$$

5: **if** $\widehat{k}(x) = 0$ **then**
6: Set $\widetilde{a}^{(\widehat{k}(x))} \leftarrow \widetilde{a}^{(0)}$
7: **else**
8: Acquire advice source $\widehat{k}(x)$ and reveal $\widetilde{a}^{(\widehat{k}(x))}$
9: **end if**
10: Route x to expert $\widehat{j}(x)$ and output $e_{\widehat{j}(x)}(x, \widetilde{a}^{(\widehat{k}(x))})$
11: **return** routed expert $\widehat{j}(x)$, executed advice $\widehat{k}(x)$, and prediction $e_{\widehat{j}(x)}(x, \widetilde{a}^{(\widehat{k}(x))})$

Table 2. Notation summary.

Symbol	Meaning
Data, experts, and advice	
$(X, Y) \sim \mathcal{D}$	Input-output random pair.
(x, y)	One realization of (X, Y).
$A = (A^1, \ldots, A^K)$	Advice random vector.
(x, a, y)	One realization of the full deferral-advice instance.
$[J] = \{1, \ldots, J\}$	Expert index set.
$[K]_0 = \{0, 1, \ldots, K\}$	Advice-action set, where $k = 0$ denotes no advice.
m_k	Masking operator that reveals at most advice source k.
$\widetilde{a}^{(k)}$	Masked realized advice under action k.
$\widetilde{A}^{(k)}$	Masked advice random variable under action k.
e_j	Expert predictor for expert j.
$\mathbf{e} = (e_1, \ldots, e_J)$	Tuple of expert predictors.
ψ	Task loss.
β_j	Expert consultation cost for expert j.
γ_k	Advice acquisition cost for advice action k.
$c_j(x, y)$	Standard L2D cost of routing (x, y) to expert j.
$c_{j,k}(x, a, y)$	Deferral-advice cost of executing expert-advice pair (j, k).
Policies, actions, and risks	
$r : \mathcal{X} \to [J]$	Router in standard Learning-to-Defer.
$q : \mathcal{X} \times [J] \to [K]_0$	Expert-conditional query function.
ℓ_{def}	Standard deferral loss.
$\ell_{\text{def-adv}}$	True deferral-advice loss.
$\mathcal{E}_{\ell_{\text{def}}}$	Population risk associated with ℓ_{def}.
$\mathcal{E}_{\ell_{\text{def-adv}}}$	Population risk associated with $\ell_{\text{def-adv}}$.
$r^\star$	Bayes-optimal router.
$q^\star$	Bayes-optimal query function.
$\Pi = [J] \times [K]_0$	Composite action space of executed expert-advice pairs.
$i \in \Pi$	Composite action index, typically identified with a pair (j, k).
$\pi : \mathcal{X} \to \Pi$	Composite predictor.
Surrogates and consistency quantities	
$\mathcal{H}_r$	Hypothesis class for router scores.
$\mathcal{H}_q$	Hypothesis class for query scores.
$\mathcal{H}_\pi$	Hypothesis class for composite scores.
s_r	Router score function.
$\mathbf{s}_r(x)$	Router score vector over experts.
s_q	Query score function over expert-advice pairs (j, k).
$\mathbf{S}_q(x)$	Query-score matrix with entries $s_q(x, j, k)$.
s_π	Composite score function.
$\mathbf{s}_\pi(x)$	Composite score vector over Π.
$\Phi^{r/q}_{\text{def-adv}}$	Separated router/query surrogate family.
$F(u, v)$	Profiled row summary in the separated-surrogate analysis.
$D(x, a, y)$	Action-independent offset in the composite-loss decomposition.
$w_i(x, a, y)$	Mismatch weight associated with composite action i.
Φ^τ_{01}	Comp-sum multiclass loss indexed by $\tau \geq 0$.
$\Phi^{\text{aug},\tau}_{\text{def-adv}}$	Augmented deferral-advice surrogate.
Γ_τ	Transfer function for the base comp-sum loss on Π.
$\widetilde{\Gamma}_\tau$	Transfer function in the augmented-surrogate consistency bound.

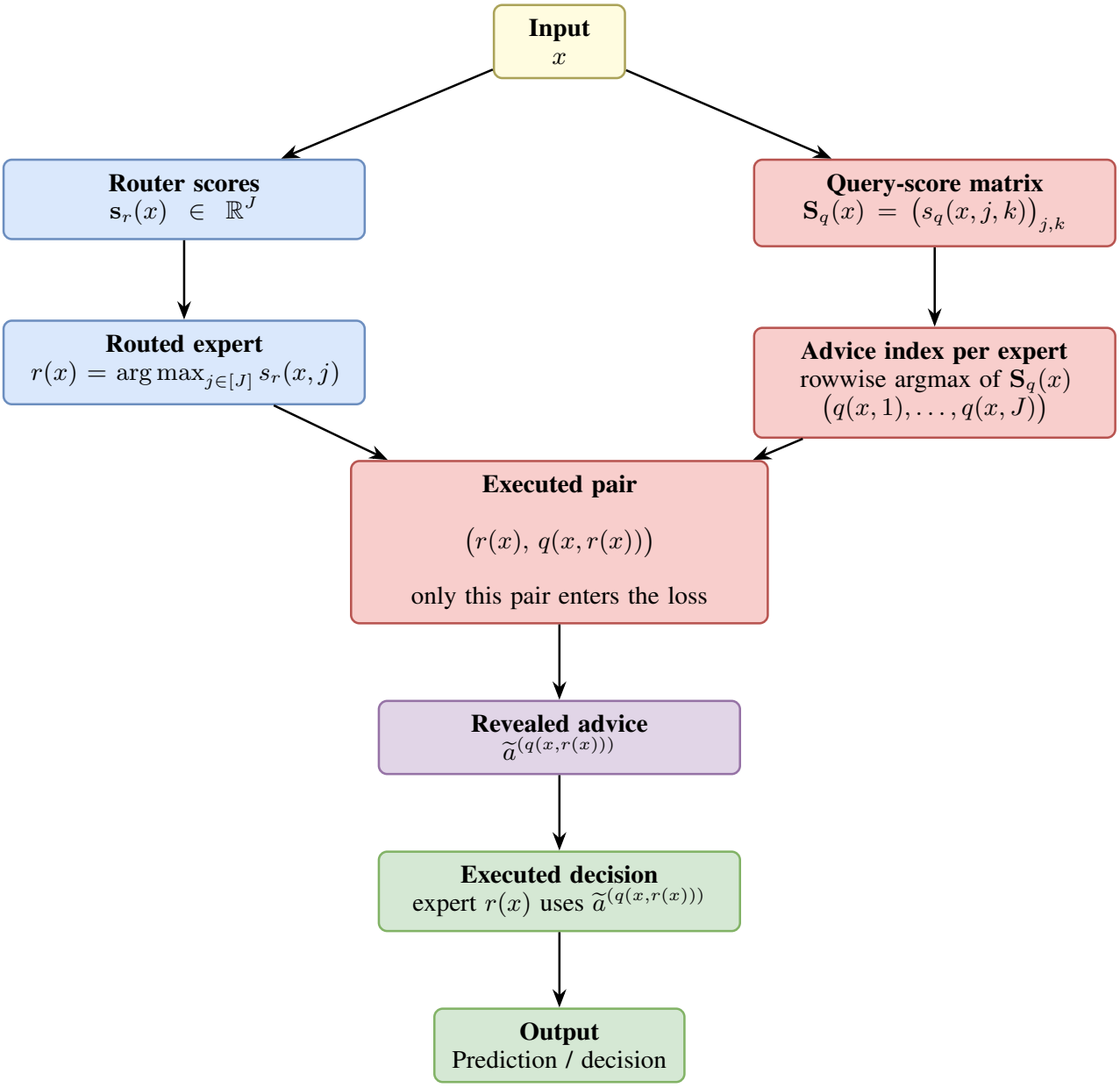

Figure 2. The deferral-advice protocol. From the input x, the router produces expert scores and the query model produces a $J \times (K+1)$ query-score matrix. Taking the argmax along each row yields one advice index per expert, but only the routed row is ever executed: the protocol uses the pair $\big(r(x),\ q(x,r(x))\big)$, reveals the corresponding masked advice, and executes expert $r(x)$ under that advice. When $q(x,r(x)) = 0$, the revealed advice is the null masked advice $\widetilde{a}^{(0)}$.

Algorithm 3 Inference with the learned composite policy

Require: Input x, learned score model $s_\pi(\cdot, \cdot; \widehat{\theta})$

Select the composite action

$$(\widehat{j}, \widehat{k}) \leftarrow \arg\max_{(j,k)\in\Pi} s_\pi(x, (j, k); \widehat{\theta})$$

if $\widehat{k} = 0$ **then**
  Set $\widetilde{a}^{(\widehat{k})} \leftarrow \widetilde{a}^{(0)}$
else
  Acquire advice source $\widehat{k}$ and reveal the masked advice $\widetilde{a}^{(\widehat{k})}$
end if
Route x to expert $\widehat{j}$ and output $e_{\widehat{j}}(x, \widetilde{a}^{(\widehat{k})})$
return executed pair $\widehat{\pi}(x) = (\widehat{j}, \widehat{k})$ and prediction $e_{\widehat{j}}(x, \widetilde{a}^{(\widehat{k})})$

B. Proofs

Example 2 (Running conditional-cost table)**.** *Before turning to the formal derivations, it is useful to keep one concrete conditional cost table in mind. Fix an input x and consider*

$$\bar{\mathbf{C}}(x) \;:=\; \big(\mathbb{E}[c_{j,k}(X,A,Y) \mid X=x]\big)_{j\in[3],\,k\in\{0,1,2\}} = \begin{pmatrix} 0.32 & 0.41 & 0.36 \\ 0.35 & 0.27 & 0.40 \\ 0.38 & 0.33 & 0.24 \end{pmatrix}.$$

Each row corresponds to an expert, each column to an advice action, and $k=0$ denotes the no-advice option. The entries already include the entire executed cost: prediction error, expert fee, and advice fee. The query decision therefore acts row by row, because once the expert is fixed the learner still has to decide which advice source, if any, should be revealed to that expert. In this table, expert 1 is best with no advice, expert 2 is best with advice 1, and expert 3 is best with advice 2. The expert-wise Bayes query therefore keeps only the row minima

$$(0.32,\ 0.27,\ 0.24),$$

and the Bayes router then selects expert 3, because 0.24 is the smallest of these three values.

Standard Learning-to-Defer would reason differently. It would only compare the unaided column

$$(0.32,\ 0.35,\ 0.38),$$

and would therefore select expert 1. The richer protocol changes the decision because it allows each expert to be evaluated under its own best advice.

This example will be useful in the next three proofs. It illustrates the three main ideas: the best advice is expert-dependent, the router should compare experts only after this expert-wise query choice has been made, and the additional advice menu can strictly improve the achievable Bayes risk.

B.1. Proof of Lemma 3

Lemma 3 (Bayes-optimal deferral-advice policy)**.** *There exists a measurable Bayes-optimal policy $(r^\star, q^\star)$ that minimizes $\mathbb{E}[\ell_{\text{def-adv}}(r,q;\,X,A,Y,\mathbf{e})]$ over all measurable policies and, for $\mathbb{P}_X$-a.e. x, can be chosen by smallest-index tie-breaking as follows:*

1. The Bayes query selects the best advice for each expert:

$$q^\star(x,j) = \operatorname{minarg\,min}_{k\in[K]_0}\ m_{j,k}(x). \tag{4}$$

2. The Bayes router selects the best optimally-advised expert:

$$r^\star(x) = \operatorname{minarg\,min}_{j\in[J]}\ \min_{k\in[K]_0}\ m_{j,k}(x). \tag{5}$$

Proof The key point is that the protocol pays only for the executed pair $(r(x), q(x, r(x)))$. Once this is made explicit, the Bayes problem reduces to a pointwise comparison of the conditional expected costs of the finitely many expert-advice pairs.

Indeed, by Definition 2, the population risk of any policy (r,q) can be written as

$$\mathbb{E}\big[\ell_{\text{def-adv}}(r,q;\,X,A,Y,\mathbf{e})\big] = \mathbb{E}\Bigg[\sum_{j\in[J]}\sum_{k\in[K]_0} c_{j,k}(X,A,Y)\,\mathbf{1}\{r(X)=j\}\,\mathbf{1}\{q(X,j)=k\}\Bigg].$$

Since exactly one pair (j,k) is selected for each X, this simplifies to

$$\mathbb{E}\big[c_{r(X),\,q(X,r(X))}(X,A,Y)\big].$$

Thus the loss depends on the policy only through the executed pair. Applying the tower property of conditional expectation gives

$$\mathbb{E}\big[c_{r(X),\,q(X,r(X))}(X,A,Y)\big] = \mathbb{E}\Big[\mathbb{E}\big[c_{r(X),\,q(X,r(X))}(X,A,Y) \mid X\big]\Big].$$

Since $r(X)$ and $q(X, r(X))$ are deterministic functions of X, conditioning on X fixes the pair $(j, k) = (r(X),\ q(X, r(X)))$, and the inner expectation becomes $\mathbb{E}[c_{j,k}(X, A, Y) \mid X]$. It is therefore enough to minimize this conditional expectation pointwise in x.

For $\mathbb{P}_X$-a.e. x, the optimization problem is

$$\min_{(j,k)\in[J]\times[K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big].$$

Since $[J] \times [K]_0$ is a finite set and each $c_{j,k}$ is bounded (hence integrable), the conditional expectations $\mathbb{E}[c_{j,k}(X, A, Y) \mid X = x]$ are well-defined measurable functions of x. The minimum of finitely many measurable functions is measurable, and choosing the smallest index among the finitely many minimizers produces a measurable selector. This minimum can be decomposed as

$$\min_{(j,k)\in[J]\times[K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big] = \min_{j\in[J]} \min_{k\in[K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big],$$

so the optimal decision decomposes in the same way. For each expert j, the smallest-index Bayes query is

$$q^\star(x, j) = \operatorname*{minarg\,min}_{k\in[K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big].$$

Once these expert-wise best query values are fixed, the router compares the resulting optimally-advised costs and therefore satisfies

$$r^\star(x) = \operatorname*{minarg\,min}_{j\in[J]} \mathbb{E}\big[c_{j,q^\star(x,j)}(X, A, Y) \mid X = x\big] = \operatorname*{minarg\,min}_{j\in[J]} \min_{k\in[K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big].$$

This is exactly the pair of Bayes decisions stated in (4) and (5). Example 3 shows this decomposition concretely: the query first selects one minimizing column per row, and the router then compares only these row minima. Consequently, the Bayes risk is

$$\mathcal{E}^\star_{\ell_{\text{def-adv}}} = \mathbb{E}\Big[\min_{j\in[J]} \min_{k\in[K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X\big]\Big].$$

■

Example 3 (Bayes query and Bayes router on a conditional cost table)**.** *Fix an input x and suppose the conditional executed-pair costs are*

$$\bar{\mathbf{C}}(x) \;:=\; \big(\mathbb{E}[c_{j,k}(X, A, Y) \mid X = x]\big)_{j\in[3],\, k\in\{0,1,2\}} = \begin{pmatrix} 0.32 & 0.41 & 0.36 \\ 0.35 & 0.27 & 0.40 \\ 0.38 & 0.33 & 0.24 \end{pmatrix}.$$

Each row corresponds to one expert, and each column to one advice action, with $k = 0$ denoting no advice. The Bayes query acts row by row: for expert 1, the smallest entry is 0.32, so $q^\star(x, 1) = 0$; for expert 2, the smallest entry is 0.27, so $q^\star(x, 2) = 1$; and for expert 3, the smallest entry is 0.24, so $q^\star(x, 3) = 2$. Once these query decisions are fixed, the router does not compare the full rows anymore. It compares the best achievable cost of each expert,

$$\big(\mathbb{E}[c_{1,q^\star(x,1)} \mid X = x],\ \mathbb{E}[c_{2,q^\star(x,2)} \mid X = x],\ \mathbb{E}[c_{3,q^\star(x,3)} \mid X = x]\big) = (0.32,\ 0.27,\ 0.24),$$

and therefore selects

$$r^\star(x) = 3.$$

First determine, for each expert, which advice makes that expert cheapest; then route to the expert whose best-advised conditional cost is smallest.

B.2. Proof of Lemma 4

Lemma 4 (Advice acquisition condition)**.** *Under the cost decomposition (3) and the smallest-index tie-breaking of Lemma 3, the Bayes-optimal query satisfies $q^\star(x, j) \neq 0$ if and only if there exists $k \in [K]$ such that*

$$\begin{aligned} &\mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(0)}), Y) \mid X = x\big] \\ &\quad - \mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(k)}), Y) \mid X = x\big] > \gamma_k. \end{aligned} \tag{6}$$

When no source satisfies this condition, $q^\star(x, j) = 0$ and the expert acts on x alone.

Proof Once the expert is fixed, the only remaining decision is whether the gain from revealing advice justifies its acquisition cost. The smallest-index tie-breaking in Lemma 3 is what converts the resulting strict comparison with the no-advice option into an "if and only if" statement.

Fix an expert $j \in [J]$ and an input x. For every $k \in [K]_0$,

$$\mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big] = \mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(k)}), Y) \mid X = x\big] + \beta_j + \gamma_k.$$

The expert fee β_j does not depend on the advice index and therefore plays no role in the comparison across k. For a fixed expert, the smallest-index Bayes query of Lemma 3 thus solves

$$q^\star(x, j) \;\in\; \operatorname*{arg\,min}_{k \in [K]_0} \Big\{\mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(k)}), Y) \mid X = x\big] + \gamma_k\Big\}.$$

The no-advice action corresponds to $k = 0$ and has cost

$$\mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(0)}), Y) \mid X = x\big],$$

because $\gamma_0 = 0$. The policy therefore strictly prefers a queried advice source $k \geq 1$ to no advice exactly when

$$\mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(k)}), Y) \mid X = x\big] + \gamma_k \;<\; \mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(0)}), Y) \mid X = x\big],$$

which rearranges to

$$\mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(0)}), Y) \mid X = x\big] - \mathbb{E}\big[\psi(e_j(X, \widetilde{A}^{(k)}), Y) \mid X = x\big] \;>\; \gamma_k.$$

This is exactly condition (6). Hence, if some $k \in [K]$ satisfies (6), then that queried action has strictly smaller conditional total cost than $k = 0$, so the Bayes-optimal query cannot be 0. Conversely, if no queried action satisfies (6), then every $k \in [K]$ has conditional total cost at least as large as that of $k = 0$. Because ties are broken in favor of the smallest advice index and 0 is the smallest index in $[K]_0$, the selected Bayes query is then $q^\star(x, j) = 0$. This proves the lemma. ■

B.3. Proof of Lemma 5

Lemma 5 (Deferral with advice dominates standard deferral)**.** *If $\mathcal{E}^\star_{\ell_{\text{def}}}$ denotes the Bayes risk of the standard deferral setting of Section 3, then*

$$\mathcal{E}^\star_{\ell_{\text{def-adv}}} \;\leq\; \mathcal{E}^\star_{\ell_{\text{def}}}. \tag{7}$$

Proof The intuition is simple. Deferral with advice enlarges the menu of admissible actions, but it never removes the standard L2D action. For any fixed expert, the best cost achievable with advice can therefore only improve upon, or match, the unaided one. The content of the lemma is that this row-wise comparison is exactly the one the Bayes router uses, so it lifts directly to the population Bayes risk.

Fix an expert $j \in [J]$ and an input x in a full-measure set on which the relevant conditional expectations are defined. Since the null action $k = 0$ belongs to $[K]_0$, minimizing over the larger advice set immediately gives

$$\min_{k \in [K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big] \;\leq\; \mathbb{E}\big[c_{j,0}(X, A, Y) \mid X = x\big]. \tag{14}$$

Expert j can always recover its unaided behavior by choosing $k = 0$.

Because (14) holds for every expert, it remains true after taking the outer minimum over $j \in [J]$:

$$\min_{j \in [J]} \min_{k \in [K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X = x\big] \;\leq\; \min_{j \in [J]} \mathbb{E}\big[c_{j,0}(X, A, Y) \mid X = x\big].$$

The left-hand side is precisely the conditional quantity minimized by the Bayes router of Lemma 3. On the right-hand side, the action $k = 0$ coincides with standard Learning-to-Defer, so $c_{j,0}(x, a, y) = c_j(x, y)$ and therefore

$$\mathbb{E}\big[c_{j,0}(X, A, Y) \mid X = x\big] = \mathbb{E}\big[c_j(X, Y) \mid X = x\big].$$

Integrating over X therefore yields

$$\mathcal{E}^\star_{\ell_{\text{def-adv}}} = \mathbb{E}\Big[\min_{j \in [J]} \min_{k \in [K]_0} \mathbb{E}\big[c_{j,k}(X, A, Y) \mid X\big]\Big] \;\leq\; \mathbb{E}\Big[\min_{j \in [J]} \mathbb{E}\big[c_j(X, Y) \mid X\big]\Big] = \mathcal{E}^\star_{\ell_{\text{def}}},$$

which proves (7). ■

Remark 10 (Equality condition in Lemma 5)**.** *The equality case must be stated at the level of the* best available *expert, not expert by expert. Define*

$$g_{\mathrm{adv}}(x) := \min_{j\in[J]} \min_{k\in[K]_0} \mathbb{E}\big[c_{j,k}(X,A,Y) \mid X=x\big], \qquad g_{\mathrm{def}}(x) := \min_{j\in[J]} \mathbb{E}\big[c_{j,0}(X,A,Y) \mid X=x\big].$$

The first part of the proof shows that $g_{\mathrm{adv}}(x) \le g_{\mathrm{def}}(x)$ for $\mathbb{P}_X$-a.e. x. Equivalently, the gap

$$h(x) := g_{\mathrm{def}}(x) - g_{\mathrm{adv}}(x)$$

is nonnegative for $\mathbb{P}_X$-a.e. x. Equality in (7),

$$\mathbb{E}[g_{\mathrm{adv}}(X)] = \mathbb{E}[g_{\mathrm{def}}(X)],$$

is therefore equivalent to $\mathbb{E}[h(X)] = 0$. Since $h(X) \ge 0$ almost surely, this happens if and only if

$$g_{\mathrm{adv}}(x) = g_{\mathrm{def}}(x) \qquad \text{for } \mathbb{P}_X\text{-a.e. } x.$$

Advice may improve some non-selected experts without changing the Bayes risk. What matters is whether it lowers the smallest *conditional cost available at x on a set of positive $\mathbb{P}_X$-measure.*

Example 4 (Lemma 5 on the running table)**.** *Return to Example 2. The conditional cost table is*

$$\bar{\mathbf{C}}(x) = \begin{pmatrix} 0.32 & 0.41 & 0.36 \\ 0.35 & 0.27 & 0.40 \\ 0.38 & 0.33 & 0.24 \end{pmatrix}.$$

The Bayes query function is obtained row by row:

$$q^\star(x,1) = 0, \qquad q^\star(x,2) = 1, \qquad q^\star(x,3) = 2.$$

This means that expert 1 should act without advice, expert 2 should receive advice source 1, and expert 3 should receive advice source 2. The query decision therefore acts within *each row: once the expert is fixed, it keeps only the advice action that makes that expert cheapest. After this expert-wise selection, the three rows reduce to the best-achievable cost vector*

$$\big(\mathbb{E}[c_{1,q^\star(x,1)} \mid X=x],\, \mathbb{E}[c_{2,q^\star(x,2)} \mid X=x],\, \mathbb{E}[c_{3,q^\star(x,3)} \mid X=x]\big) = (0.32,\ 0.27,\ 0.24).$$

The Bayes router therefore chooses

$$r^\star(x) = 3,$$

because expert 3 now has the smallest achievable conditional cost.

For comparison, standard Learning-to-Defer only sees the no-advice column,

$$\big(\mathbb{E}[c_{1,0} \mid X=x],\, \mathbb{E}[c_{2,0} \mid X=x],\, \mathbb{E}[c_{3,0} \mid X=x]\big) = (0.32,\ 0.35,\ 0.38),$$

so its Bayes router would select expert 1. Denoting this standard router by $r^\star_{\mathrm{def}}$, we therefore have

$$r^\star_{\mathrm{def}}(x) = 1 \qquad \text{whereas} \qquad r^\star(x) = 3.$$

The reason is exactly the one formalized in Lemma 5: once the advice menu is available, each row can only stay the same or improve, because $k = 0$ remains feasible. In this example, the no-advice comparison vector

$$(0.32,\ 0.35,\ 0.38)$$

is replaced by the best-advised vector

$$(0.32,\ 0.27,\ 0.24).$$

This new vector is coordinatewise no larger, and it is strictly smaller in the second and third coordinates. What matters for the lemma is the smallest entry of these two vectors, because that entry determines the Bayes decision at x. Here the best achievable conditional cost drops from

$$g_{\mathrm{def}}(x) = 0.32 \qquad \text{to} \qquad g_{\mathrm{adv}}(x) = 0.24.$$

This is why the richer protocol changes the selected expert and strictly lowers the Bayes-risk contribution at this input. Had advice improved only a non-selected row while leaving the minimum at 0.32, the overall Bayes risk would have remained unchanged. This is precisely the distinction captured by the equality condition in Lemma 5.

B.4. Proof of Theorem 6

We begin from the exact binary deferral-advice loss and show how the usual logistic replacement recipe leads naturally to a separated router/query surrogate. We then identify the quantity that this surrogate actually compares after the query heads are optimized out: each expert row is compressed into a profiled summary $F(c_{j,0}, c_{j,1})$, not into its row minimum. The final step is to construct a bounded cost table on which these two comparisons disagree. This shows that the failure is not tied to one peculiar parameterization, but to the separated architecture itself.

We work in the binary setting $J = 2$, $K = 1$ introduced in Section 4.4. The surrogate (8) is parameterized by a cost transform ν, query functions G and U, the logistic pair Φ_0, Φ_1, and router weights Ψ_1, Ψ_2.

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

$$r^b(x) = 1 + \mathbf{1}\{s_r^b(x) \geq 0\}, \qquad q_j^b(x) = \mathbf{1}\{s_{q_j^b}(x) \geq 0\},$$

with $q^b(x, j) = q_j^b(x)$.

B.4.1. Proof of Proposition 11

Proposition 11 (Natural separated construction)**.** *For a realization (x, a, y), the binary executed-pair loss admits the exact expansion*

$$\ell^b_{\text{def-adv}}(r^b, q^b;\, x, a, y, \mathbf{e}) := \sum_{j=1}^{2} \sum_{k=0}^{1} c_{j,k}(x, a, y)\, \mathbf{1}\{r^b(x) = j\}\, \mathbf{1}\{q_j^b(x) = k\}. \tag{15}$$

Assigning logistic surrogate potentials to the decoded binary outcomes in the four terms of (15) yields the basic separated surrogate $\Phi^{b,r/q}_{\text{def-adv}}(s_r^b, \mathbf{s}_{q^b})$.

Proof This proposition records the most direct surrogate construction suggested by the protocol itself. In the binary setting there are only four executed outcomes: route to expert 1 or expert 2, and for the selected expert either do not query or query the single advice source. Writing these four cases out explicitly makes it clear how the usual logistic replacement enters.

Fix (x, a, y). Since exactly one expert and one binary advice decision are executed, expanding the two sums in (15) gives

$$\begin{aligned}
\ell^b_{\text{def-adv}}(r^b, q^b;\, x, a, y, \mathbf{e}) = {} & c_{1,0}(x, a, y)\, \mathbf{1}\{s_r^b(x) < 0\}\, \mathbf{1}\{s_{q_1^b}(x) < 0\} \\
& + c_{1,1}(x, a, y)\, \mathbf{1}\{s_r^b(x) < 0\}\, \mathbf{1}\{s_{q_1^b}(x) \geq 0\} \\
& + c_{2,0}(x, a, y)\, \mathbf{1}\{s_r^b(x) \geq 0\}\, \mathbf{1}\{s_{q_2^b}(x) < 0\} \\
& + c_{2,1}(x, a, y)\, \mathbf{1}\{s_r^b(x) \geq 0\}\, \mathbf{1}\{s_{q_2^b}(x) \geq 0\}.
\end{aligned} \tag{16}$$

This is the exact executed-pair loss written in terms of the decoded binary scores.

Assigning logistic potentials to the decoded binary outcomes yields

$$\Phi^{b,r/q}_{\text{def-adv}}(s_r^b, \mathbf{s}_{q^b}) := \sum_{j=1}^{2} \Phi_{j-1}(s_r^b(x)) \Big[c_{j,0}(x, a, y)\, \Phi_0(s_{q_j^b}(x)) + c_{j,1}(x, a, y)\, \Phi_1(s_{q_j^b}(x)) \Big].$$

This is exactly the basic logistic construction described in Section 4.4. This proves the natural separated construction. ■

Regularity conditions for the larger separated family. Equation (8) extends the basic construction by allowing monotone cost transforms, query reparameterizations, and flexible router weights. We impose the following conditions. They are not designed to make the counterexample artificially strong. They simply capture the regularity one would naturally ask of a well-behaved separated surrogate.

(N1) *Cost transform:* $\nu : [0, C] \to [0, \infty)$ is continuous, differentiable on $(0, C)$, and strictly increasing there, so $\nu'(c) > 0$ for $c \in (0, C)$.

(Q1) *Amplitude and reparameterization:* $G : \mathbb{R} \to (0, \infty)$ is continuous and strictly positive; $U : \mathbb{R} \to \mathbb{R}$ is continuous, strictly increasing, and surjective.

(Q2) *Well-posedness:* for every $(u, v) \in (0, C]^2$, $t \mapsto \nu(u)\, G(t)\, \Phi_0(U(t)) + \nu(v)\, G(t)\, \Phi_1(U(t))$ is coercive and strictly convex with a unique minimizer $t^\star(u, v)$, and the map $(u, v) \mapsto t^\star(u, v)$ is continuous on $(0, C]^2$.

(P1) *Router weights:* $\Psi_1, \Psi_2 \in C^2(\mathbb{R})$ with $\Psi_j > 0$, $\Psi_1' < 0$, $\Psi_2' > 0$, $\Psi_j'' > 0$.

(P2) *Routing ratio:* $\rho(r) := -\Psi_1'(r)/\Psi_2'(r)$ is a strictly decreasing bijection $\mathbb{R} \to (0, +\infty)$ with $\rho(0) = 1$.

These conditions are satisfied, for instance, by the standard logistic (cross-entropy) surrogate with softmax router weights on the strictly positive cost regime used in the theorem below.

Containment of the traditional surrogate. The familiar construction $\Phi^{b,r/q}_{\text{def-adv}}$ is recovered by taking $\nu(c) = c$, $G(q) \equiv 1$, and $U(q) = q$, together with the usual logistic router weights. The theorem proved below therefore applies to a broad family of positive-cost separated models that still preserves the same separated router/query architecture. The next lemmas isolate the precise object this architecture optimizes.

The expert summary. The next three lemmas formalize the quantity introduced in the main text. They show that after profiling over the query score of expert j, the entire row $(c_{j,0}(x, a, y), c_{j,1}(x, a, y))$ is compressed into the scalar summary $F(c_{j,0}(x, a, y), c_{j,1}(x, a, y))$. The key point is that this summary remains strictly sensitive to the non-minimal entry of the row. This is precisely where the surrogate departs from the Bayes comparison.

Definition 12 (Expert summary)**.** *For $(u, v) \in [0, C]^2$, define*

$$F(u, v) \;:=\; \inf_{t \in \mathbb{R}} \big[\nu(u)\, G(t)\, \Phi_0(U(t)) + \nu(v)\, G(t)\, \Phi_1(U(t))\big]. \tag{17}$$

B.4.2. Proof of Lemma 13

Lemma 13 (Well-defined expert summary)**.** *Under (N1)–(Q2), for every $(u, v) \in (0, C]^2$, the infimum in (17) is attained at a unique point $t^\star(u, v) \in \mathbb{R}$.*

Proof Fix $(u, v) \in (0, C]^2$ and define

$$f(t; u, v) := \nu(u)\, G(t)\, \Phi_0(U(t)) + \nu(v)\, G(t)\, \Phi_1(U(t)).$$

By (Q2), $f(\cdot; u, v)$ is coercive and strictly convex on $\mathbb{R}$. Coercivity implies that every sublevel set is compact, hence a global minimizer exists. Strict convexity then implies that this minimizer is unique. ■

B.4.3. Proof of Lemma 14

Lemma 14 (Expert summary after profiling)**.** *Under (P1) and (Q2), for every fixed $s_r^b : \mathcal{X} \to \mathbb{R}$ and every realization (x, a, y) whose four binary costs lie in $(0, C]$,*

$$\begin{aligned}\inf_{s_{q_1^b}, s_{q_2^b}} \Phi^{\nu,r/q}_{\text{def-adv}}(s_r^b, \mathbf{s}_{q^b}) = \Psi_1(s_r^b(x))\, F(c_{1,0}(x, a, y), c_{1,1}(x, a, y)) \\ + \Psi_2(s_r^b(x))\, F(c_{2,0}(x, a, y), c_{2,1}(x, a, y)).\end{aligned} \tag{18}$$

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 s_r^b.

Proof Fix (x, a, y) and write $t_r = s_r^b(x)$, $t_1 = s_{q_1^b}(x)$, and $t_2 = s_{q_2^b}(x)$. By (8),

$$\begin{aligned}\Phi^{\nu,r/q}_{\text{def-adv}}(s_r^b, \mathbf{s}_{q^b}) = \Psi_1(t_r)\Big[&\nu(c_{1,0}(x, a, y))\, G(t_1)\, \Phi_0(U(t_1)) \\ &+ \nu(c_{1,1}(x, a, y))\, G(t_1)\, \Phi_1(U(t_1))\Big] \\ + \Psi_2(t_r)\Big[&\nu(c_{2,0}(x, a, y))\, G(t_2)\, \Phi_0(U(t_2)) \\ &+ \nu(c_{2,1}(x, a, y))\, G(t_2)\, \Phi_1(U(t_2))\Big].\end{aligned}$$

The score t_1 appears only in the first bracket and t_2 appears only in the second. Because $\Psi_1(t_r)$ and $\Psi_2(t_r)$ are strictly positive by (P1), they factor through the infima over t_1 and t_2. Therefore,

$$\begin{aligned}
&\inf_{s_{q_1^b}, s_{q_2^b}} \Phi^{\nu,r/q}_{\text{def-adv}}(s_r^b, \mathbf{s}_{q^b}) \\
&= \Psi_1(t_r) \inf_{t_1\in\mathbb{R}} \Big[\nu(c_{1,0}(x,a,y))\, G(t_1)\, \Phi_0(U(t_1)) + \nu(c_{1,1}(x,a,y))\, G(t_1)\, \Phi_1(U(t_1))\Big] \\
&\quad + \Psi_2(t_r) \inf_{t_2\in\mathbb{R}} \Big[\nu(c_{2,0}(x,a,y))\, G(t_2)\, \Phi_0(U(t_2)) + \nu(c_{2,1}(x,a,y))\, G(t_2)\, \Phi_1(U(t_2))\Big].
\end{aligned}$$

Each infimum is exactly an instance of the expert summary (17). This gives (18). The characterization of the minimizers follows directly from Lemma 13. ■

B.4.4. Proof of Lemma 15

Lemma 15 (Continuity and strict monotonicity of the expert summary)**.** *Under (N1)–(Q2), the map $F : (0, C]^2 \to (0, \infty)$ is continuous. Moreover, for every fixed $u \in (0, C)$, the map $v \mapsto F(u, v)$ is differentiable on $(0, C)$ and satisfies*

$$\frac{\partial F}{\partial v}(u, v) = \nu'(v)\, G(t^\star(u, v))\, \Phi_1(U(t^\star(u, v))) > 0. \tag{19}$$

In particular, for every fixed $u \in (0, C)$, the function $v \mapsto F(u, v)$ is strictly increasing on $(0, C)$.

Proof Define

$$f(t; u, v) := \nu(u)\, G(t)\, \Phi_0(U(t)) + \nu(v)\, G(t)\, \Phi_1(U(t)).$$

By (N1) and (Q1), the map $(t, u, v) \mapsto f(t; u, v)$ is continuous. By Lemma 13, each pair $(u, v) \in (0, C]^2$ has a unique minimizer $t^\star(u, v)$, and the continuity of $(u, v) \mapsto t^\star(u, v)$ is part of (Q2). Hence

$$F(u, v) = f(t^\star(u, v); u, v)$$

is continuous. Since $\nu(u) > 0$ and $\nu(v) > 0$ on $(0, C]$, and since G, Φ_0, Φ_1 are strictly positive, the attained value is also strictly positive.

Now fix $u \in (0, C)$ and $v \in (0, C)$. Since the minimizer is unique, Danskin's theorem applies to the function $v \mapsto \inf_{t\in\mathbb{R}} f(t; u, v) = F(u, v)$ and yields

$$\frac{\partial F}{\partial v}(u, v) = \frac{\partial}{\partial v} f(t; u, v)\Big|_{t=t^\star(u,v)}.$$

Differentiating f with respect to v gives

$$\frac{\partial}{\partial v} f(t; u, v) = \nu'(v)\, G(t)\, \Phi_1(U(t)).$$

Substituting $t = t^\star(u, v)$ proves (19). Every factor on the right-hand side is strictly positive: $\nu'(v) > 0$ by (N1), $G(t^\star(u, v)) > 0$ by (Q1), and $\Phi_1(z) = \log(1 + e^z) > 0$ for every $z \in \mathbb{R}$. Therefore $\partial F/\partial v(u, v) > 0$, so $F(u, \cdot)$ is strictly increasing. ■

Lemmas 14 and 15 are the structural core of the impossibility result. Once the query scores are profiled out, expert j no longer appears through its row minimum. It appears through the scalar summary $F(c_{j,0}(x, a, y), c_{j,1}(x, a, y))$, and this summary depends strictly on the non-minimal entry. Bayes and the surrogate are therefore comparing different objects.

B.4.5. Proof of Lemma 16

Lemma 16 (Router minimizer)**.** *Under (P1)–(P2), for any $A, B > 0$ the function $t \mapsto A\,\Psi_1(t) + B\,\Psi_2(t)$ has a unique minimizer $t_r^\star(A, B)$ satisfying*

$$A > B \implies t_r^\star > 0, \qquad A = B \implies t_r^\star = 0, \qquad A < B \implies t_r^\star < 0.$$

That is, the router selects the expert with the smaller *profiled summary.*

Proof Strict convexity follows from (P1), since

$$\frac{d^2}{dt^2}\big[A\,\Psi_1(t) + B\,\Psi_2(t)\big] = A\,\Psi_1''(t) + B\,\Psi_2''(t) > 0.$$

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

$$A\,\Psi_1'(t_r^\star) + B\,\Psi_2'(t_r^\star) = 0,$$

or equivalently

$$-\frac{\Psi_1'(t_r^\star)}{\Psi_2'(t_r^\star)} = \frac{B}{A}.$$

The left-hand side is $\rho(t_r^\star)$, which is a strictly decreasing bijection from $\mathbb{R}$ to $(0,\infty)$ by (P2); since $B/A \in (0,\infty)$, this FOC admits a unique solution, and by strict convexity any critical point of a C^2 strictly convex function is its global minimizer. Since $\rho(0) = 1$, we obtain

$$\frac{B}{A} < 1 \iff t_r^\star(A,B) > 0, \qquad \frac{B}{A} = 1 \iff t_r^\star(A,B) = 0, \qquad \frac{B}{A} > 1 \iff t_r^\star(A,B) < 0.$$

This is exactly the stated sign characterization. ■

B.4.6. Completion of the proof of Theorem 6

Theorem 6 (Broad Bayes inconsistency of separated router/query surrogates)**.** *Under mild regularity conditions on ν, G, U, Ψ_1, Ψ_2 (Appendix B.4), for every $b \in (0,C)$ and $\varepsilon \in (0, C-b)$, there exists $\delta \in (0,b)$ such that the pointwise conditional cost table*

$$\bar{\mathbf{C}}(x) \;=\; \begin{pmatrix} b-\delta & C \\ b & b+\varepsilon \end{pmatrix} \tag{10}$$

has Bayes-optimal expert $j^\star = 1$, while the unique minimizer of the pointwise surrogate (8) decodes to expert 2. In particular, the entire family (8) is not Bayes consistent for the true loss.

Proof Fix $b \in (0,C)$ and $\varepsilon \in (0, C-b)$. We will construct a bounded cost table on which Bayes and the separated surrogate rank the two experts in opposite orders. The reason this can happen is the one identified in the main text: Bayes compares row minima, whereas the separated surrogate compares profiled row summaries.

Because $C > b+\varepsilon$ and $v \mapsto F(b,v)$ is strictly increasing by Lemma 15,

$$F(b,C) > F(b, b+\varepsilon). \tag{20}$$

The profile of the first row is therefore already larger than that of the second row when both rows have the same minimum value b. Since $(u,v) \mapsto F(u,v)$ is continuous, we can perturb the first row minimum slightly downward without destroying this strict inequality. More precisely, there exists $\delta \in (0,b)$ such that

$$F(b-\delta, C) > F(b, b+\varepsilon). \tag{21}$$

Now consider the cost table (10). Its row minima are

$$\min\{b-\delta, C\} = b-\delta, \qquad \min\{b, b+\varepsilon\} = b.$$

Since $\delta > 0$, Bayes compares these two values and selects expert 1.

The separated surrogate behaves differently. By Lemma 14, profiling out the two query scores reduces the pointwise objective to

$$\widetilde{\Phi}^{\nu,r/q}_{\text{def-adv}}(t) = F(b-\delta, C)\,\Psi_1(t) + F(b, b+\varepsilon)\,\Psi_2(t).$$

Define

$$A := F(b-\delta, C), \qquad B := F(b, b+\varepsilon).$$

By (21), we have $A > B$. Lemma 16 then implies that the profiled router objective has a unique minimizer $t_r^\star(A, B) > 0$. Under the decoding rule

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

this router minimizer decodes to expert 2.

To conclude, note that the query minimizers are also unique: for each row, Lemma 13 gives a unique minimizer $t^\star(c_{j,0}, c_{j,1})$, and Lemma 14 shows that these minimizers are independent of the router score. Hence the full pointwise surrogate has a unique minimizer

$$\big(t_r^\star(A, B),\ t^\star(b - \delta, C),\ t^\star(b, b + \varepsilon)\big),$$

and this minimizer decodes to expert 2 even though Bayes selects expert 1. Thus the separated surrogate fails the defining pointwise criterion for Bayes consistency. ■

The counterexample already appears within a bounded cost table, so the failure does not rely on any singular limiting construction. Nor does it disappear once the surrogate family is made more flexible: the contradiction persists under monotone cost transforms, positive query amplitudes, smooth reparameterizations, and well-behaved router weights. The obstruction is more basic. The separated architecture compresses each row before routing, whereas the Bayes rule compares rows only through their minima.

B.5. Proof of the Sequential–Composite Equivalence

Proposition 17 (Sequential–composite equivalence)**.** *For every sequential policy (r, q), the composite predictor $\pi(x) = (r(x), q(x, r(x)))$ incurs the same realized loss as (r, q) under $\ell_{\text{def-adv}}$. Conversely, every composite predictor $\pi : \mathcal{X} \to \Pi$ 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 proof is simply an identification of the object that the loss depends on. Although the protocol is written sequentially, the realized loss only sees the executed pair consisting of the routed expert and the advice ultimately shown to that expert.

Let (r, q) be any sequential policy and define $\pi(x) = (r(x), q(x, r(x)))$. By Definition 2,

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

where $\pi(x) = (\pi_1(x), \pi_2(x))$. Thus every sequential policy induces a composite predictor with exactly the same realized loss at every input.

Conversely, let $\pi : \mathcal{X} \to \Pi$ and write $\pi(x) = (\pi_1(x), \pi_2(x))$. To implement the same executed pair sequentially, define

$$r(x) = \pi_1(x).$$

The query function only matters at the routed expert, so it is enough to require

$$q(x, r(x)) = \pi_2(x).$$

One convenient choice is to set $q(x, j) = \pi_2(x)$ for every $j \in [J]$. Then $q(x, r(x)) = \pi_2(x)$ and therefore

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

So every composite predictor can be realized by a sequential policy with the same pointwise loss.

Since the two constructions agree pointwise in both directions, the sequential and composite formulations induce exactly the same set of achievable realized losses. Their population risk functionals therefore have the same infima and the same Bayes-optimal decisions. ■

B.6. Proof of Proposition 18

Proposition 18 (Mismatch decomposition)**.** *For every composite predictor* $\pi : \mathcal{X} \to \Pi$ *and every realization* (x, a, y)*,*

$$\ell_{\text{def-adv}}(\pi;\, x, a, y, \mathbf{e}) \;=\; D(x, a, y) \;+\; \sum_{i \in \Pi} w_i(x, a, y)\, \mathbf{1}\{\pi(x) \neq i\},$$

where $M(x, a, y) = \max_{t \in \Pi} c_t(x, a, y)$*,* $w_i(x, a, y) = M(x, a, y) - c_i(x, a, y) \geq 0$*, and* $D(x, a, y) = \sum_{i \in \Pi} c_i(x, a, y) - (|\Pi| - 1)\, M(x, a, y)$*.*

Proof The decomposition is easiest to see by fixing the composite action that is actually executed. Let

$$i_\pi \coloneqq \pi(x) \in \Pi.$$

Since the predictor selects exactly one composite action, the true loss is just the realized cost of that action:

$$\ell_{\text{def-adv}}(\pi;\, x, a, y, \mathbf{e}) \;=\; c_{i_\pi}(x, a, y).$$

Now expand the mismatch term. The indicator $\mathbf{1}\{\pi(x) \neq i\}$ vanishes only at the executed action $i = i_\pi$, so the sum keeps every weight except the one attached to the chosen pair:

$$\begin{aligned}\sum_{i \in \Pi} w_i\, \mathbf{1}\{\pi(x) \neq i\} = \sum_{i \neq i_\pi} w_i &= \sum_{i \neq i_\pi} \big(M - c_i\big) \\ &= (|\Pi| - 1)\, M - \sum_{i \neq i_\pi} c_i.\end{aligned}$$

Adding the offset

$$D = \sum_{i \in \Pi} c_i - (|\Pi| - 1)\, M$$

therefore gives

$$D + \sum_{i \in \Pi} w_i\, \mathbf{1}\{\pi(x) \neq i\} = \sum_{i \in \Pi} c_i - (|\Pi| - 1)\, M + (|\Pi| - 1)\, M - \sum_{i \neq i_\pi} c_i = c_{i_\pi},$$

which is exactly $\ell_{\text{def-adv}}(\pi;\, x, a, y, \mathbf{e})$. This proves the decomposition. ■

B.7. Proof of Lemma 7

Proof This lemma is a derivation rather than a separate statistical claim. Once the true loss is written over the composite action space, the only action-dependent part is the weighted mismatch term. The augmented surrogate is obtained by replacing each mismatch indicator with a multiclass surrogate on Π.

Indeed, Proposition 18 shows that

$$\ell_{\text{def-adv}}(\pi;\, x, a, y, \mathbf{e}) \;=\; D(x, a, y) + \sum_{i \in \Pi} w_i(x, a, y)\, \mathbf{1}\{\pi(x) \neq i\},$$

where $D(x, a, y)$ does not depend on the chosen action. Thus, for surrogate design, the relevant object is

$$\sum_{i \in \Pi} w_i(x, a, y)\, \mathbf{1}\{\pi(x) \neq i\}.$$

The comp-sum family (Mao et al., 2023b) provides a multiclass surrogate for prediction on the composite label space Π. Applying it label by label to the weighted mismatch term leads directly to

$$\Phi^{\text{aug},\tau}_{\text{def-adv}}(s_\pi;\, x, a, y, \mathbf{e}) \;\coloneqq\; \sum_{i \in \Pi} w_i(x, a, y)\, \Phi^{\tau}_{01}(\mathbf{s}_\pi(x), i),$$

which is exactly (12). In other words, the augmented surrogate is the comp-sum analogue of the weighted mismatch representation of the true loss on the composite action space. ■

B.8. Proof of Theorem 8

The proof proceeds by reducing the augmented objective to an ordinary multiclass problem on the composite action space Π, with example-dependent weights absorbed into a new data distribution.

B.8.1. Proof of Lemma 19

Lemma 19 (Weighted multiclass reduction)**.** *Let* $\mathbf{w}(x) := (\mathbb{E}[w_i(X, A, Y) \mid X = x])_{i\in\Pi}$*, and define*

$$d(x) := \mathbb{E}[D(X, A, Y) \mid X = x].$$

For every $i \in \Pi$*, write*

$$\bar{w}_i(x) := \mathbb{E}[w_i(X, A, Y) \mid X = x].$$

Define a probability vector $p(x) = (p_i(x))_{i\in\Pi}$ *on* Π *by*

$$p_i(x) := \begin{cases} \bar{w}_i(x)/\|\mathbf{w}(x)\|_1, & \|\mathbf{w}(x)\|_1 > 0, \\ 1/|\Pi|, & \|\mathbf{w}(x)\|_1 = 0, \end{cases} \qquad i \in \Pi,$$

and, for any composite predictor π *and any score function* s_π*, define the conditional multiclass risks*

$$C^p_{01}(\pi, x) := \sum_{i\in\Pi} p_i(x)\, \mathbf{1}\{\pi(x) \neq i\}, \qquad C^p_{\Phi^\tau_{01}}(s_\pi, x) := \sum_{i\in\Pi} p_i(x)\, \Phi^\tau_{01}(\mathbf{s}_\pi(x), i).$$

Assume $\mathbb{E}[\|\mathbf{w}(X)\|_1] > 0$*, and define a probability distribution* $\widetilde{\mathcal{D}}$ *on* $\mathcal{X} \times \Pi$ *by*

$$\widetilde{\mathbb{P}}(X \in B, I = i) \;:=\; \frac{1}{\mathbb{E}[\|\mathbf{w}(X)\|_1]}\, \mathbb{E}\big[\bar{w}_i(X)\, \mathbf{1}\{X \in B\}\big] \tag{22}$$

for every measurable set $B \subseteq \mathcal{X}$ *and every* $i \in \Pi$*. Let* $\ell_{01}(\pi;\, x, i) := \mathbf{1}\{\pi(x) \neq i\}$ *denote the multiclass* 0–1 *loss on* Π*. Then, for every scoring function* $s_\pi \in \mathcal{H}_\pi$ *and its decoded predictor* $\pi(x) = \arg\max_{i\in\Pi} s_\pi(x, i)$*,*

$$C_{\ell_{\text{def-adv}}}(\pi, x) = d(x) + \|\mathbf{w}(x)\|_1\, C^p_{01}(\pi, x), \tag{23}$$

$$C_{\Phi^{\text{aug},\tau}}(s_\pi, x) = \|\mathbf{w}(x)\|_1\, C^p_{\Phi^\tau_{01}}(s_\pi, x), \tag{24}$$

where

$$C_{\ell_{\text{def-adv}}}(\pi, x) := \mathbb{E}\big[\ell_{\text{def-adv}}(\pi;\, X, A, Y, \mathbf{e}) \mid X = x\big]$$

and

$$C_{\Phi^{\text{aug},\tau}}(s_\pi, x) := \mathbb{E}\big[\Phi^{\text{aug},\tau}_{\text{def-adv}}(s_\pi;\, X, A, Y, \mathbf{e}) \mid X = x\big].$$

Consequently,

$$\mathcal{E}_{\Phi^{\text{aug},\tau}}(s_\pi) = \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{E}_{\Phi^\tau_{01}}(s_\pi; \widetilde{\mathcal{D}}), \tag{25}$$

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\pi) = \mathbb{E}[D(X, A, Y)] + \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{E}_{\ell_{01}}(\pi; \widetilde{\mathcal{D}}). \tag{26}$$

Moreover, the best-in-class risks and minimizability gaps satisfy

$$\mathcal{E}^*_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi) = \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{E}^*_{\Phi^\tau_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}), \tag{27}$$

$$\mathcal{U}_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi) = \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{U}_{\Phi^\tau_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}), \tag{28}$$

$$\mathcal{E}^B_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) = \mathbb{E}[D(X, A, Y)] + \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{E}^B_{\ell_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}), \tag{29}$$

$$\mathcal{U}_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) = \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{U}_{\ell_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}). \tag{30}$$

Proof The point of the lemma is that, once we condition on $X = x$, the deferral-advice problem becomes an ordinary multiclass problem on the composite action space Π, up to an additive offset and a positive scale.

Fix x. By Proposition 18,

$$\ell_{\text{def-adv}}(\pi;\, X, A, Y, \mathbf{e}) = D(X, A, Y) + \sum_{i \in \Pi} w_i(X, A, Y)\, \mathbf{1}\{\pi(X) \neq i\}.$$

Taking the conditional expectation given $X = x$ gives

$$C_{\ell_{\text{def-adv}}}(\pi, x) = d(x) + \sum_{i \in \Pi} \bar{w}_i(x)\, \mathbf{1}\{\pi(x) \neq i\}.$$

If $\|\mathbf{w}(x)\|_1 = 0$, then every $\bar{w}_i(x)$ vanishes, so the action-dependent term disappears and

$$C_{\ell_{\text{def-adv}}}(\pi, x) = d(x).$$

Because the multiplier $\|\mathbf{w}(x)\|_1$ is also zero, both (23) and (24) hold automatically in this degenerate case, regardless of the arbitrary choice of $p(x)$.

Assume now that $\|\mathbf{w}(x)\|_1 > 0$. Then $\bar{w}_i(x) = \|\mathbf{w}(x)\|_1\, p_i(x)$ for every $i \in \Pi$, and therefore

$$C_{\ell_{\text{def-adv}}}(\pi, x) = d(x) + \|\mathbf{w}(x)\|_1 \sum_{i \in \Pi} p_i(x)\, \mathbf{1}\{\pi(x) \neq i\},$$

which is exactly (23).

The same normalization applies to the surrogate:

$$C_{\Phi^{\text{aug},\tau}}(s_\pi, x) = \sum_{i \in \Pi} \bar{w}_i(x)\, \Phi_{01}^{\tau}(\mathbf{s}_\pi(x), i) = \|\mathbf{w}(x)\|_1 \sum_{i \in \Pi} p_i(x)\, \Phi_{01}^{\tau}(\mathbf{s}_\pi(x), i),$$

which proves (24).

Thus, for each fixed input x, the conditional deferral-advice problem is exactly a weighted multiclass problem on Π, up to the action-independent offset $d(x)$.

To pass from this conditional reduction to a population statement, we average these weighted multiclass problems over X. Since $\bar{w}_i(X) \geq 0$ for all $i \in \Pi$, the right-hand side of (22) defines a finite non-negative measure on $\mathcal{X} \times \Pi$. Its total mass is

$$\sum_{i \in \Pi} \widetilde{\mathbb{P}}(X \in \mathcal{X}, I = i) = \frac{1}{\mathbb{E}[\|\mathbf{w}(X)\|_1]} \sum_{i \in \Pi} \mathbb{E}[\bar{w}_i(X)] = \frac{1}{\mathbb{E}[\|\mathbf{w}(X)\|_1]}\, \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big] = 1,$$

so $\widetilde{\mathcal{D}}$ is a probability distribution.

Integrating (24) over X yields

$$\mathcal{E}_{\Phi^{\text{aug},\tau}}(s_\pi) = \mathbb{E}\Big[\sum_{i \in \Pi} \bar{w}_i(X)\, \Phi_{01}^{\tau}(\mathbf{s}_\pi(X), i)\Big] = \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{E}_{\Phi_{01}^{\tau}}(s_\pi; \widetilde{\mathcal{D}}),$$

which is (25). Likewise, (23) gives

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\pi) = \mathbb{E}[D(X, A, Y)] + \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\, \mathcal{E}_{\ell_{01}}(\pi; \widetilde{\mathcal{D}}),$$

which is (26).

The remaining identities follow from the same affine scaling. Taking infima over $s_\pi \in \mathcal{H}_\pi$ or over the induced predictors gives (27) and (29). Since the minimizability gaps are defined through these same best-in-class risks, the same argument also yields (28) and (30). ■

Theorem 8 ($\mathcal{H}_\pi$-consistency of the augmented surrogate)**.** *Assume that $\mathcal{H}_\pi$ satisfies the standard symmetry and completeness conditions required by the comp-sum multiclass $\mathcal{H}_\pi$-consistency theorem for Φ_{01}^{τ} on the label set Π. For every $\tau \geq 0$, the*

augmented surrogate $\Phi^{\mathrm{aug},\tau}_{\text{def-adv}}$ satisfies an $\mathcal{H}_\pi$-consistency bound with respect to $\ell_{\text{def-adv}}$. Specifically, for every $s_\pi \in \mathcal{H}_\pi$ and its induced composite predictor π, write $\Phi^{\tau}_{\mathrm{aug}} := \Phi^{\mathrm{aug},\tau}_{\text{def-adv}}$. Define

$$\Delta_\ell(\pi) := \mathcal{E}_{\ell_{\text{def-adv}}}(\pi) - \mathcal{E}^B_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) + \mathcal{U}_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi),$$
$$\Delta_\Phi(s_\pi) := \mathcal{E}_{\Phi^\tau_{\mathrm{aug}}}(s_\pi) - \mathcal{E}^*_{\Phi^\tau_{\mathrm{aug}}}(\mathcal{H}_\pi) + \mathcal{U}_{\Phi^\tau_{\mathrm{aug}}}(\mathcal{H}_\pi).$$

Then

$$\Delta_\ell(\pi) \le \widetilde{\Gamma}_\tau(\Delta_\Phi(s_\pi))\,. \tag{13}$$

In the non-degenerate regime $\mathbb{E}[\|\mathbf{w}(X)\|_1] > 0$, $\widetilde{\Gamma}_\tau(u) \;:=\; \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\,\Gamma_\tau\left(\frac{u}{\mathbb{E}[\|\mathbf{w}(X)\|_1]}\right)$, where Γ_τ is the non-negative, non-decreasing, concave transfer function for Φ^τ_{01} on Π, and $\mathbf{w}(x) := (\mathbb{E}[w_i(X,A,Y) \mid X = x])_{i\in\Pi}$.

Proof If $\mathbb{E}[\|\mathbf{w}(X)\|_1] = 0$, then $\bar{w}_i(X) = 0$ almost surely for every $i \in \Pi$. By Proposition 18, the true risk then reduces to the expectation of the action-independent term $D(X, A, Y)$, so every predictor is optimal and the theorem is trivial. We therefore assume $\mathbb{E}[\|\mathbf{w}(X)\|_1] > 0$.

The key idea is to compare the surrogate and the true loss at the level where the Bayes policy itself is defined, namely conditionally on the input x. Lemma 19 shows that, once $X = x$ is fixed, both the true deferral-advice risk and the augmented surrogate reduce to weighted multiclass risks on the composite action space Π:

$$C_{\ell_{\text{def-adv}}}(\pi, x) = d(x) + \|\mathbf{w}(x)\|_1\, C^p_{01}(\pi, x), \qquad C_{\Phi^{\mathrm{aug},\tau}}(s_\pi, x) = \|\mathbf{w}(x)\|_1\, C^p_{\Phi^\tau_{01}}(s_\pi, x).$$

The additive term $d(x)$ is independent of the chosen composite action, so it plays no role in the comparison between policies. What matters is that the action-dependent part of the true loss and the surrogate are governed by the same conditional weight vector $p(x)$.

It is therefore natural to look first at the conditional excess risk. For each input x, define

$$\Delta C_{\ell_{\text{def-adv}}}(\pi, x) := C_{\ell_{\text{def-adv}}}(\pi, x) - \inf_{s'_\pi \in \mathcal{H}_\pi} C_{\ell_{\text{def-adv}}}(\pi_{s'_\pi}, x),$$
$$\Delta C_{\Phi^{\mathrm{aug},\tau}}(s_\pi, x) := C_{\Phi^{\mathrm{aug},\tau}}(s_\pi, x) - \inf_{s'_\pi \in \mathcal{H}_\pi} C_{\Phi^{\mathrm{aug},\tau}}(s'_\pi, x).$$

Here $\pi_{s'_\pi}$ denotes the predictor induced by s'_π. When $\|\mathbf{w}(x)\|_1 > 0$, the identities above show that these conditional gaps are exactly the weighted multiclass excess risks induced by $p(x)$, up to the common scale factor $\|\mathbf{w}(x)\|_1$. This is the local reason the augmented surrogate is aligned with the correct comparison. The remainder of the proof turns this pointwise observation into an $\mathcal{H}_\pi$-consistency statement at the population level.

This identifies the right conditional comparison, but the theorem is an $\mathcal{H}_\pi$-consistency statement rather than a purely pointwise calibration claim. To keep the hypothesis-class restriction explicit, we gather these conditional weighted problems into the multiclass distribution $\widetilde{\mathcal{D}}$ defined in Lemma 19. Under that distribution, the comp-sum loss Φ^τ_{01} satisfies the usual multiclass $\mathcal{H}_\pi$-consistency inequality with transfer function Γ_τ. Write

$$\Delta\mathcal{E}_{01}(\pi; \widetilde{\mathcal{D}}) := \mathcal{E}_{\ell_{01}}(\pi; \widetilde{\mathcal{D}}) - \mathcal{E}^B_{\ell_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}) + \mathcal{U}_{\ell_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}),$$
$$\Delta\mathcal{E}_{\Phi^\tau_{01}}(s_\pi; \widetilde{\mathcal{D}}) := \mathcal{E}_{\Phi^\tau_{01}}(s_\pi; \widetilde{\mathcal{D}}) - \mathcal{E}^*_{\Phi^\tau_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}) + \mathcal{U}_{\Phi^\tau_{01}}(\mathcal{H}_\pi; \widetilde{\mathcal{D}}).$$

By the symmetry and completeness assumptions on $\mathcal{H}_\pi$ in Theorem 8, the comp-sum $\mathcal{H}_\pi$-consistency bound of Mao et al. (2023b) applies and gives

$$\Delta\mathcal{E}_{01}(\pi; \widetilde{\mathcal{D}}) \;\le\; \Gamma_\tau\Big(\Delta\mathcal{E}_{\Phi^\tau_{01}}(s_\pi; \widetilde{\mathcal{D}})\Big)\,.$$

Lemma 19 shows that this ordinary multiclass problem is exactly the population counterpart of the conditional reduction above. Define

$$\widetilde{\Gamma}_\tau(u) \;:=\; \mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\,\Gamma_\tau\left(\frac{u}{\mathbb{E}[\|\mathbf{w}(X)\|_1]}\right),$$

so that $\mathbb{E}[\|\mathbf{w}(X)\|_1]\,\Gamma_\tau(u) = \widetilde{\Gamma}_\tau(\mathbb{E}[\|\mathbf{w}(X)\|_1]\,u)$ by construction. Multiplying the previous inequality by $\mathbb{E}[\|\mathbf{w}(X)\|_1]$ gives

$$\mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\,\Delta\mathcal{E}_{01}(\pi; \widetilde{\mathcal{D}}) \;\le\; \widetilde{\Gamma}_\tau\Big(\mathbb{E}\big[\|\mathbf{w}(X)\|_1\big]\,\Delta\mathcal{E}_{\Phi^\tau_{01}}(s_\pi; \widetilde{\mathcal{D}})\Big)\,.$$

We now translate this inequality back to deferral with advice using (25)–(30). The additive term $\mathbb{E}[D(X, A, Y)]$ appears in both the incurred true risk and the best-in-class true risk, so it cancels from the excess-risk difference. We obtain

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\pi) - \mathcal{E}^{B}_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) + \mathcal{U}_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) \leq \widetilde{\Gamma}_\tau(\mathcal{E}_{\Phi^{\text{aug},\tau}}(s_\pi) - \mathcal{E}^{*}_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi) + \mathcal{U}_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi)),$$

which is exactly (13).

This proves the claimed $\mathcal{H}_\pi$-consistency bound. When the minimizability gaps vanish, the right-hand side tends to zero whenever the surrogate excess risk tends to zero, so the augmented surrogate recovers the best-in-class deferral-advice risk. If the class is Bayes-rich, this best-in-class risk is the Bayes-optimal deferral-advice risk.

From (Mao et al., 2023b), we have:

$$\Gamma_\tau(u) = \begin{cases} \sqrt{2^\tau(2-\tau)u} & \tau \in [0,1) \\ \sqrt{2|\Pi|^{\tau-1}u} & \tau \in [1,2) \\ (\tau-1)|\Pi|^{\tau-1}u & \tau \in [2,+\infty) \end{cases} \tag{31}$$

■

B.9. Proof of Corollary 9

Proof The corollary is the statistical form of Theorem 8. Once the two minimizability gaps vanish, the theorem compares the true best-in-class excess risk of the induced policy directly to the surrogate excess risk of the learned score; the Bayes-richness assumption identifies this best-in-class risk with the unrestricted Bayes risk.

Under the stated assumption,

$$\mathcal{U}_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) = \mathcal{U}_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi) = 0.$$

Moreover, by Bayes-richness,

$$\mathcal{E}^{B}_{\ell_{\text{def-adv}}}(\mathcal{H}_\pi) = \inf_\pi \mathcal{E}_{\ell_{\text{def-adv}}}(\pi).$$

Applying Theorem 8 to the learned score $\widehat{s}_{\pi,n}$ and its induced predictor $\widehat{\pi}_n$ therefore gives, for every n,

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\widehat{\pi}_n) - \inf_\pi \mathcal{E}_{\ell_{\text{def-adv}}}(\pi) \;\leq\; \widetilde{\Gamma}_\tau(\mathcal{E}_{\Phi^{\text{aug},\tau}}(\widehat{s}_{\pi,n}) - \mathcal{E}^{*}_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi)).$$

Thus the only remaining quantity to control is the surrogate excess risk.

By assumption,

$$\mathcal{E}_{\Phi^{\text{aug},\tau}}(\widehat{s}_{\pi,n}) - \mathcal{E}^{*}_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi) \to 0$$

in probability. Since Γ_τ is concave, non-negative, non-decreasing, and satisfies $\Gamma_\tau(0) = 0$, the definition of $\widetilde{\Gamma}_\tau$ in Theorem 8 implies that $\widetilde{\Gamma}_\tau$ is continuous at 0. It therefore follows that

$$\widetilde{\Gamma}_\tau(\mathcal{E}_{\Phi^{\text{aug},\tau}}(\widehat{s}_{\pi,n}) - \mathcal{E}^{*}_{\Phi^{\text{aug},\tau}}(\mathcal{H}_\pi)) \to 0$$

in probability as well. Combining this with the preceding excess-risk bound proves that

$$\mathcal{E}_{\ell_{\text{def-adv}}}(\widehat{\pi}_n) - \inf_\pi \mathcal{E}_{\ell_{\text{def-adv}}}(\pi) \to 0$$

in probability. ■

C. Experiments

C.1. Global Experimental Protocol

Before turning to the individual tasks, we summarize the experimental choices that are shared across the appendix. The goal is to evaluate the learned policy in the regime studied by the theory. The learner does not predict an abstract label directly. It chooses an executed expert-advice pair, and the value of that choice is revealed only through the resulting realized cost. The experimental protocol is therefore designed to mirror the formal setup as closely as possible.

Evaluation metric. Our primary metric is always the empirical average of the true deferral-advice loss,

$$\widehat{\mathcal{E}}_{\ell_{\text{def-adv}}}(\pi) \;=\; \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} \ell_{\text{def-adv}}(\pi; x_i, a_i, y_i, \mathbf{e}_i),$$

evaluated on the validation split of the corresponding experiment. This is the only quantity used to compare policies across values of λ. The reason is structural. Once advice costs vary, two policies may have very different consultation patterns even when they achieve similar predictive accuracy. Accuracy alone therefore no longer represents the deployment objective of the paper. The true loss remains the correct comparison criterion because it accounts jointly for task error, expert fees, and advice costs. The appendix therefore keeps the focus on true-loss tables and uses advice-distribution figures only as diagnostics to explain how a method achieves a certain true loss rather than to rank methods.

Learned policy. In every experiment, the learned policy acts on the composite action space $\Pi = [J] \times [K]_0$. The deployed decision is therefore a single executed pair (j, k). Still, it is useful to parameterize the scores in a way that mirrors the protocol. We write the score of action (j, k) as

$$s_\theta(x, (j, k)) \;=\; u_\theta(x, j) + v_\theta(x, j, k),$$

where $u_\theta(x, j)$ plays the role of an expert-routing score and $v_\theta(x, j, k)$ is an expert-conditional advice adjustment. Both depend on the input through a shared learned representation, so the same construction can be used for text and tabular inputs.

A convenient instantiation first maps the input to a latent representation $z_\theta(x) \in \mathbb{R}^d$ and then introduces routing embeddings $a_j \in \mathbb{R}^d$, advice-side expert embeddings $m_j \in \mathbb{R}^d$, advice embeddings $g_k \in \mathbb{R}^d$, and score biases $\rho_j, \delta_k \in \mathbb{R}$:

$$\begin{aligned} u_\theta(x, j) &:= \rho_j + \langle a_j,\ z_\theta(x) \rangle, \\ v_\theta(x, j, k) &:= \delta_k + \big\langle m_j,\ g_k \odot z_\theta(x) \big\rangle, \end{aligned}$$

where $\odot$ denotes elementwise multiplication. This factorization preserves three properties that are important in practice. First, routing remains input-dependent through u_θ. Second, the value of advice acquisition can depend jointly on the input and an expert through v_θ. Third, the expert-advice interaction is non-separable, while the number of parameters still scales through a shared low-rank representation rather than independently over all $J(K + 1)$ composite actions.

Decoding remains a single argmax over composite actions:

$$\pi_\theta(x) \;=\; \operatorname*{arg\,max}_{(j,k) \in \Pi}\ s_\theta(x, (j, k)).$$

Resources. All experiments were conducted on an internal cluster using an NVIDIA A100 GPU with 40 GB of VRAM.

Reading the appendix tables. The main paper reports only the true deferral-advice loss, because that is the deployment metric of interest and the one that remains meaningful across cost regimes. The appendix keeps the same true-loss tables, but adds expert-by-advice summaries and advice-distribution figures for selected values of λ. These additional diagnostics help explain how the learned policy adapts as advice becomes more or less expensive, and they make explicit how the deployment-optimal expert-advice pair changes with the cost regime.

C.1.1. Baselines

We use the same baseline family throughout, and each baseline is included for a specific reason. All methods are evaluated on exactly the same precomputed executed-pair table, so differences in performance come from the decision rule rather than from stochastic variation in the experts or advice sources.

Our method. This is our main method. It predicts directly over the composite action space $\Pi = [J] \times [K]_0$ and is the only learned policy that is allowed to coordinate expert selection and advice acquisition jointly. Its role is therefore to measure the benefit of solving the full deferral-advice problem.

L2D. This is the standard no-query baseline obtained by restricting the action space to the J actions $(j, 0)$, as established in (Mao et al., 2023a; 2024g; Montreuil et al., 2025b). It answers the first natural question: is explicit advice acquisition useful at all beyond ordinary learning to defer? If our method does not improve over L2D, then the added advice mechanism is not buying anything under the true deployment objective.

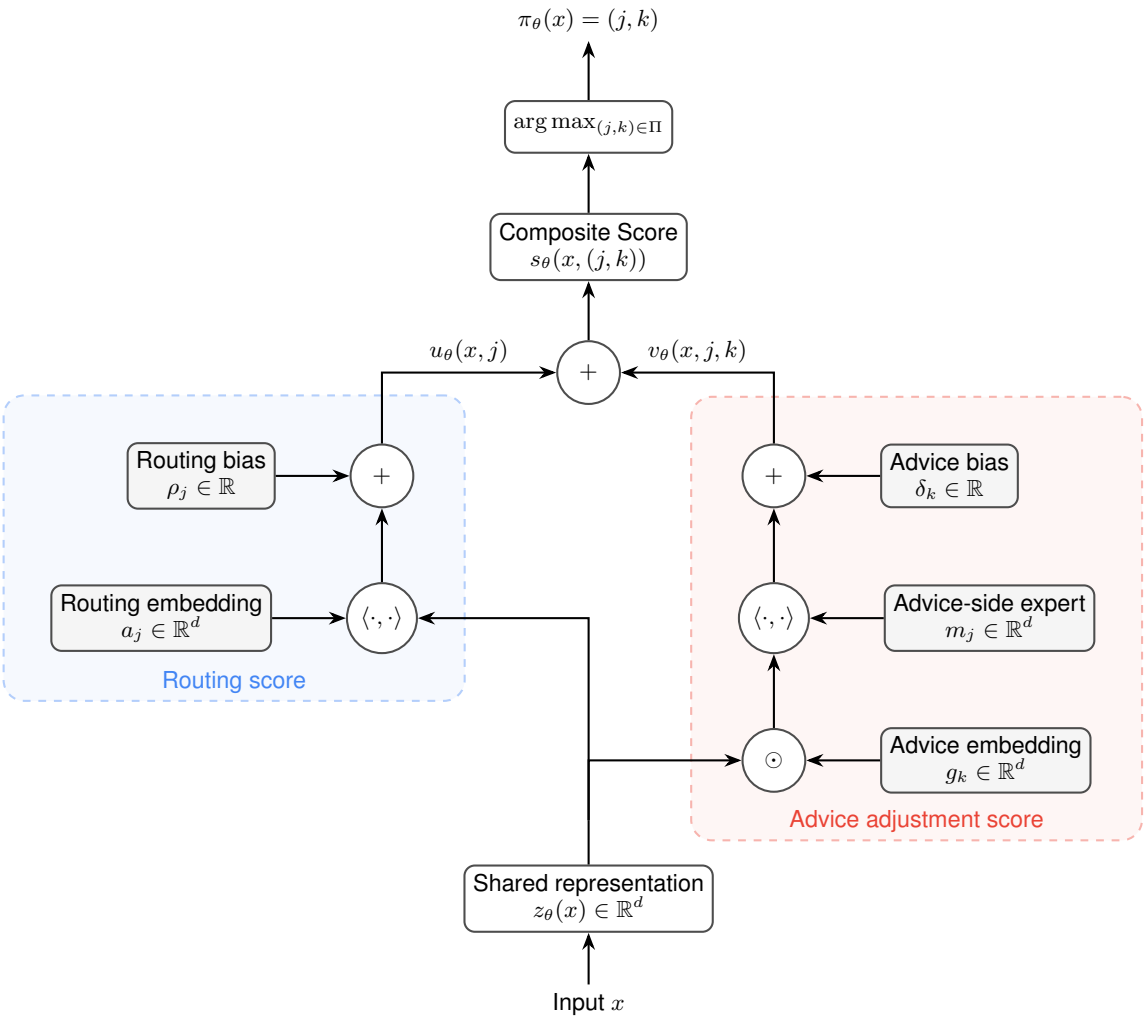

Figure 3. Structured parameterization of the composite policy score. The final decision remains a single argmax over executed expert-advice pairs, while the score decomposes into a routing term and an expert-conditional advice adjustment.

Best fixed pair. This baseline selects a single constant composite action (j, k) and applies it to every input. It is an oracle over non-adaptive policies, and therefore measures how much is gained by input-dependent decision-making rather than by choosing one good expert-advice pair globally.

Partial-randomization ablations. These are the two most informative diagnostic baselines. In "learned expert, random advice," we keep the expert chosen by our learned policy and replace only the advice action by a uniform random choice. In "random expert, learned advice," we do the reverse. Taken together, these two baselines separate the contribution of expert selection from that of advice selection. If the full method improves over both, then the gain cannot be attributed to only one half of the decision.

Random (j, k). The final baseline samples both the expert and the advice action uniformly. Its role is not diagnostic but calibrating: it provides an uninformed floor against which the learned and structured methods can be compared.

C.2. Synthetic Theorem-Aligned Experiment

This appendix includes one synthetic benchmark whose role is purely theoretical. The goal is not to mimic a realistic deployment pipeline, but to instantiate the exact mechanism of Theorem 6 and show that the observed failure of the separated surrogate is the one predicted by the theory. The same benchmark also contains a second region in which advice is genuinely useful, so that standard L2D is structurally suboptimal. The synthetic experiment therefore isolates the two distinct reasons why the augmented formulation should dominate the alternatives.

Goal, data, and metric. We work in the minimal binary setting of the theorem: two experts ($J = 2$) and one advice action ($K = 1$), so the composite action space is $\Pi = \{(1, 0), (1, 1), (2, 0), (2, 1)\}$. Inputs are sampled uniformly from $[-1, 1]^2$, and the first coordinate partitions the space into two regions:

$$R_- := \{x \in [-1, 1]^2 : x_1 < 0\}, \qquad R_+ := \{x \in [-1, 1]^2 : x_1 \geq 0\}.$$

The second coordinate is irrelevant to the Bayes rule; it is included only so that the learned decision maps can be visualized on a two-dimensional grid. We evaluate every policy by its expected true deferral-advice loss on an independent test split drawn from the same synthetic environment. Because the environment is defined directly through executed costs, the evaluation metric matches the object studied in the theory exactly.

Certified cost construction. We fix $\lambda_0 = 0.08$ and assign one conditional executed-cost table to each region. On R_-, we use the certified theorem table

$$T^- = \begin{pmatrix} 0.38 & 1.08 \\ 0.50 & 0.51 \end{pmatrix},$$

while on R_+ we use the advice-helpful table

$$T^+ = \begin{pmatrix} 0.55 & 0.18 \\ 0.30 & 0.90 \end{pmatrix}.$$

Table 3 records these two regions in the same executed-action format as the main paper. Every queried entry is realizable as a Bernoulli task loss plus the fixed advice fee λ_0, so the synthetic environment is not an abstract cost table detached from any prediction model.

Table 3. Certified conditional executed-cost tables used in the synthetic benchmark. Each row is one region of the input space, and each column is one executed expert-advice pair.

Region	$(1, 0)$	$(1, 1)$	$(2, 0)$	$(2, 1)$	Bayes action	Bayes risk
R_-	0.380	1.080	0.500	0.510	$(1, 0)$	0.380
R_+	0.550	0.180	0.300	0.900	$(1, 1)$	0.180

The left region is the theorem region. Bayes compares row minima and therefore prefers expert 1, since $\min_k T^-_{1,k} = 0.38 < 0.50 = \min_k T^-_{2,k}$. The exact binary separated surrogate from the paper behaves differently: its profiled row summaries satisfy

$$F(0.38, 1.08) = 0.8371 \ > \ 0.7000 = F(0.50, 0.51),$$

so the profiled surrogate reverses the comparison and favors expert 2. The right region serves a different purpose. There, Bayes chooses the queried action $(1, 1)$ with cost 0.18, whereas L2D is restricted to the no-advice column and must therefore pick expert 2 with cost 0.30. Since both regions have positive probability, the separated surrogate and L2D each incur a positive irreducible excess risk, but for different reasons.

Policies and exact separated baseline. We compare our method, L2D, the same random baselines used in the real experiments, and one *exact* implementation of the binary separated surrogate analyzed in Section 4.4. This is important. The separated baseline is not an approximate router-query variant. It is precisely the special case obtained by setting $\nu = \mathrm{id}$, $G \equiv 1$, $U = \mathrm{id}$, and $\Psi_j = \Phi_{j-1}$, so that each indicator in the true binary loss is replaced by its logistic envelope. Concretely, the model uses one binary router score s^b_r and one expert-conditional query score $s_{q^b_1}$ or $s_{q^b_2}$ per row, exactly as in the theorem statement.

Training protocol. All learned methods — our augmented surrogate, the separated baseline, and the L2D baseline — share the same backbone: a two-layer MLP over the synthetic input $x \in \mathbb{R}^2$, with hidden widths $(32, 32)$ and ReLU activations. They are trained with AdamW (learning rate 3×10^{-3}, no weight decay, gradient clipping at norm 10) for 120 epochs with batch size 256 and no learning-rate scheduler. Our method uses the augmented comp-sum surrogate at $\tau = 1$; the separated baseline uses the exact binary logistic surrogate described above; and the L2D baseline is restricted to the k=0 slice and trained with the standard cross-entropy deferral surrogate. We sweep the training size over $n \in \{250, 500, 1000, 2500, 5000\}$ and report averages over five seeds. The figures and tables below use the largest train size, $n = 5000$, for the pointwise comparisons.

Table 4 is the central quantitative result. Our method reaches 0.281 ± 0.001 true risk, essentially matching the Bayes optimum of 0.280. The separated surrogate stalls far above this level, with excess risk 0.054 ± 0.004, even though it is

Table 4. Synthetic test performance at the largest train size ($n = 5000$). Our method is the only one whose excess risk is essentially zero.

Method	Test true risk	Test excess risk	Advice rate (%)	Bayes-action match (%)
Our	0.281 ± 0.001	0.001 ± 0.000	49.6	**99.3**
Separated	0.334 ± 0.004	0.054 ± 0.004	62.5	58.4
L2D	0.342 ± 0.001	0.062 ± 0.000	0.0	48.8
Learned route, random advice	0.661 ± 0.002	0.381 ± 0.002	49.6	24.5
Random route, no advice	0.432 ± 0.001	0.152 ± 0.001	0.0	24.9
Random (j, k)	0.550 ± 0.003	0.269 ± 0.004	50.1	24.9

trained with the exact binary surrogate analyzed in the theorem. L2D also remains suboptimal, but for a different reason: it never queries, so it cannot realize the Bayes action on R_+. The random baselines are much worse, which confirms that the gain is not coming from a trivial bias in the environment.

Table 5. Regionwise diagnosis at the largest train size ($n = 5000$). The theorem region R_- isolates the failure of the separated surrogate, while the advice-helpful region R_+ isolates the limitation of L2D.

Method	Risk on R_-	Risk on R_+	Bayes match on R_- (%)	Bayes match on R_+ (%)
Our	**0.381**	**0.181**	**99.3**	**99.3**
Separated	0.484	0.184	18.0	99.0
L2D	0.383	0.300	97.5	0.0
Learned route, random advice	0.727	0.594	48.9	0.1
Random route, no advice	0.440	0.424	49.8	0.0
Random (j, k)	0.617	0.482	24.7	25.1

Table 5 shows that the two failure modes are exactly the ones predicted by the theory. On R_-, the separated surrogate matches the Bayes action only 18.0% of the time, even though it is nearly perfect on R_+; this is the empirical footprint of the profiled-row distortion proved in Theorem 6. L2D exhibits the opposite pattern. It is nearly optimal on R_-, where the Bayes action is no-advice, but it matches the Bayes action 0% of the time on R_+ because the Bayes decision there is the queried action $(1, 1)$. Our method is nearly perfect on both regions, which is exactly the consistency story established by Theorem 8.

Figures 4 and 5 make the same point visually. The excess risk curve shows that our method is the only one that continues to improve toward Bayes optimality as the sample size grows. The action maps then show *why*: our method learns the correct composite decision on both halves of the space, while the exact separated surrogate and L2D fail on the region singled out by the corresponding theory.

C.3. FEVER Experimental Details

This appendix records the implementation details that are omitted from the main paper for space reasons. We focus on the FEVER (Thorne et al., 2018) experiment reported in Section 5, since it is the main empirical study in the paper.

Data, action space, and metric. FEVER (Thorne et al., 2018) is a fact-verification benchmark built from short natural-language claims paired with Wikipedia evidence. Each example must be classified as `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFO` depending on whether the claim is supported by the evidence available in Wikipedia. A typical example is a claim such as "The Eiffel Tower is located in Berlin"; the correct label is `REFUTES` once the relevant Wikipedia passages are retrieved. The experiment uses $15{,}000$ examples sampled from the labelled development split, with a stratified $75/25$ train-validation partition. The composite action space has size 3×5, corresponding to three experts and five advice actions. All policy comparisons use the global protocol of Section C.1; in particular, the reported numbers are validation averages of the true deferral-advice loss.

Experts and advice construction. The expert pool contains Qwen3-4B-Instruct (Yang et al., 2025), Qwen3-8B, and DeBERTa-v3-large (He et al., 2021) fine-tuned on entailment tasks, with respective consultation costs $\beta = (0.03, 0.05, 0.04)$. Advice is derived from Wikipedia evidence retrieved with BM25 (Robertson & Zaragoza, 2009). The five advice actions are: no retrieval, BM25 top-1, BM25 top-3, BM25 top-5, and query reformulation followed by BM25 top-5. The reformulated query is produced by Qwen2.5-1.5B. Query costs take the form $\gamma_k = \lambda \gamma_k^{\text{base}}$ with $\gamma^{\text{base}} = (0, 0.015, 0.02, 0.03, 0.01)$. This construction is chosen to induce heterogeneous expert-advice trade-offs, rather than to mimic exact monetary costs.

Prompt templates. The expert-facing prompts are deliberately simple so that the experiment isolates the interaction between routing and advice rather than prompt engineering. For the two Qwen experts, the system instruction is to return

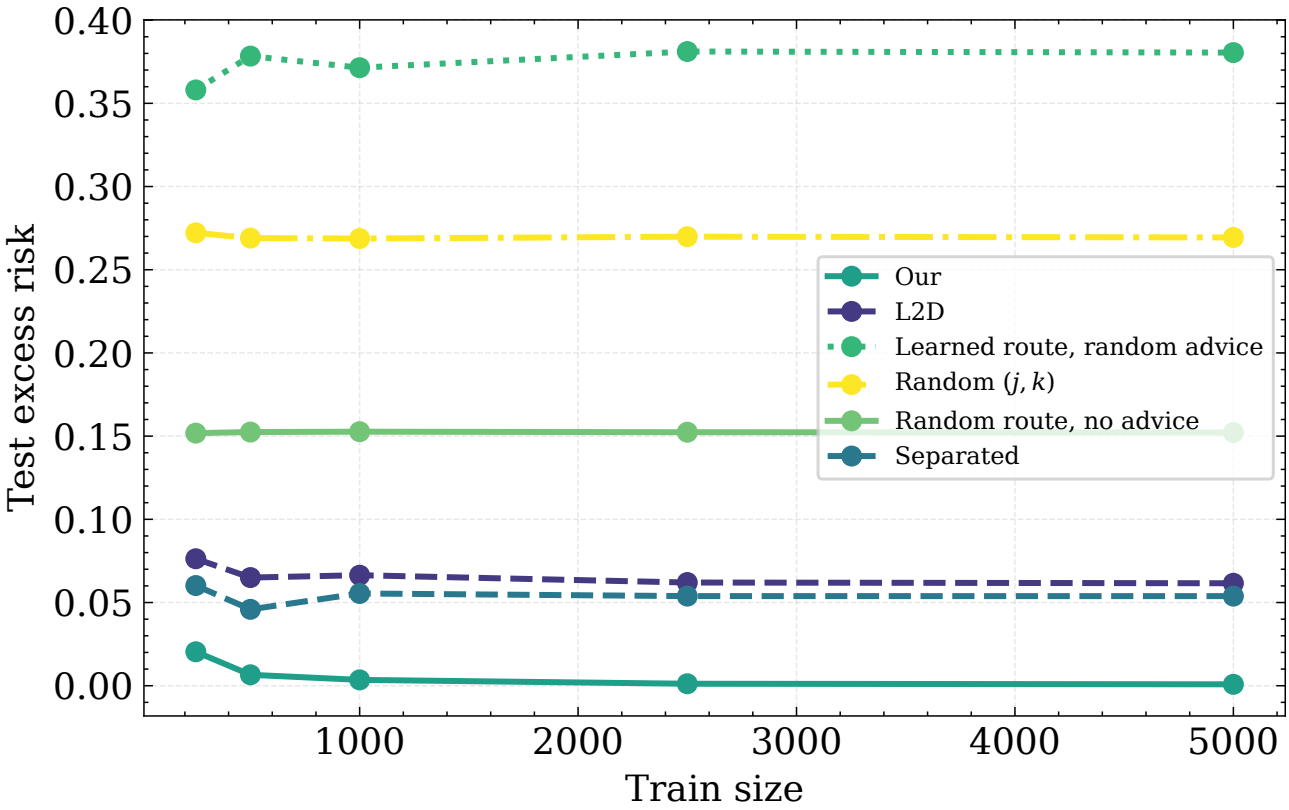

Figure 4. Synthetic test excess risk as a function of train size. Our method approaches zero excess risk, whereas the exact separated surrogate and L2D remain bounded away from the Bayes optimum.

exactly one of `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFO`. The user message then instantiates the template

```
Claim: <claim>.
Evidence: <retrieved passages, or EMPTY when k = 0>.
Answer with exactly one label: SUPPORTS, REFUTES, or NOT_ENOUGH_INFO.
```

For example, if the claim is "The Eiffel Tower is located in Berlin," the top-1 advice may supply a passage stating that the Eiffel Tower is in Paris, in which case the desired output is `REFUTES`. The reformulator is prompted separately to rewrite the claim into short retrieval text, e.g.

```
Rewrite the claim as a concise Wikipedia search query. Return only the query
text.
```

The exact prompt strings used in the runs are the ones stored in the released experiment configuration.

Policy training and baselines. Input features are frozen `all-MiniLM-L6-v2` sentence embeddings (dimension 384). All learned methods — our augmented surrogate and the L2D baseline — share the same backbone: a two-layer MLP with hidden dimensions $(128, 64)$, ReLU activations, and no dropout. They are trained with AdamW (weight decay 10^{-4}, gradient clipping at norm 10) for 50 epochs with batch size 128 and no learning-rate scheduler. We train L2D with a learning rate 5×10^{-3} and our augmented surrogate with learning rate 10^{-3}. The augmented comp-sum surrogate uses $\tau = 1$; the L2D baseline uses the same architecture but is restricted to the k=0 slice and trained with the standard cross-entropy deferral surrogate. The best-fixed, partial-randomization, and joint-random baselines are exactly those defined in Section C.1, instantiated on the FEVER expert pool and advice set.

Reading the main-paper FEVER table. Table 6 shows that advice utility is expert-dependent: the best retrieval strategy is not shared across experts, which is precisely the regime in which a joint expert-advice policy is needed. Table 1 then evaluates policies on the true deployment metric, reporting mean and standard deviation over 4 random seeds. Its central

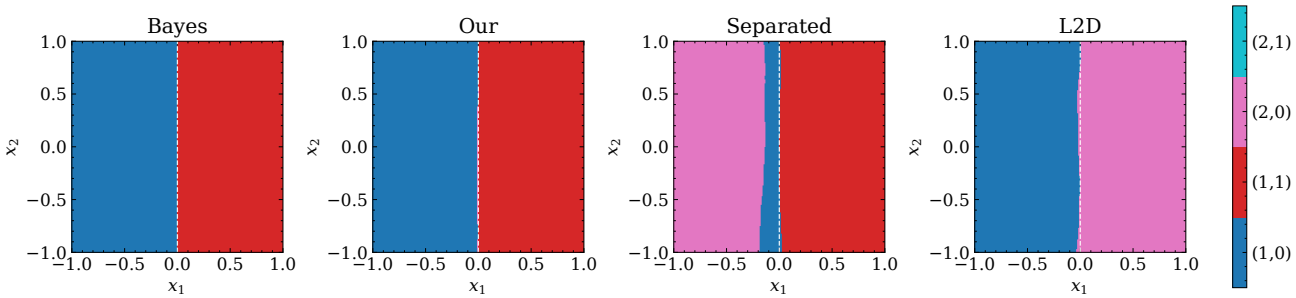

Figure 5. Learned decision maps at the largest train size ($n = 5000$). The vertical dashed line marks the transition between the theorem region R_- and the advice-helpful region R_+. Our method recovers the Bayes split, the separated surrogate selects the wrong expert on the left region, and L2D cannot realize the queried Bayes action on the right region.

Table 6. Validation accuracy (%) of each expert under each advice action at $\lambda = 0$. Bold marks the best advice action for each expert. The preferred action is expert-dependent.

	k=0 (none)	k=1 (top-1)	k=2 (top-3)	k=3 (top-5)	k=4 (reform.)
Qwen3-4B-Instruct	33.5	33.4	32.9	**36.3**	34.0
Qwen3-8B	33.5	33.4	34.1	**41.1**	35.7
DeBERTa-v3-large	33.2	**42.9**	42.5	42.3	39.2

message is that our method achieves the lowest mean loss for seven of the eight retained cost regimes and ties L2D when advice becomes too expensive at $\lambda = 10$, while also dominating the non-adaptive and randomized references. The best fixed composite action changes with λ: it is DeBERTa with top-1 retrieval for $\lambda \in \{0, 0.5, 1, 2, 4\}$, and the cheapest no-advice expert for $\lambda \in \{6, 8, 10\}$.

Table 7. Best advice action for each expert across the retained FEVER cost grid. Each cell reports the best advice index k and the corresponding true loss. Here $k = 0$ denotes no advice, $k = 1$ BM25 top-1, $k = 2$ BM25 top-3, $k = 3$ BM25 top-5, and $k = 4$ reformulation followed by BM25 top-5.

Expert	$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	$\lambda = 6$	$\lambda = 8$	$\lambda = 10$
Qwen3-4B-Instruct	k=3 / 0.667	k=3 / 0.682	k=0 / 0.695	k=0 / 0.695	k=0 / 0.695	k=0 / **0.695**	k=0 / **0.695**	k=0 / **0.695**
Qwen3-8B	k=3 / 0.639	k=3 / 0.654	k=3 / 0.669	k=3 / 0.699	k=0 / 0.715	k=0 / 0.715	k=0 / 0.715	k=0 / 0.715
DeBERTa-v3-large	k=1 / **0.611**	k=1 / **0.619**	k=1 / **0.626**	k=1 / **0.641**	k=1 / **0.671**	k=1 / 0.701	k=0 / 0.708	k=0 / 0.708

Table 7 makes the mechanism explicit over the full retained grid. At low advice cost, different experts prefer different advice actions: the two Qwen experts favor deeper retrieval, whereas DeBERTa prefers the cheapest retrieval action. As λ increases, these choices do not collapse simultaneously. Qwen3-4B switches to no advice already at $\lambda = 1$, Qwen3-8B remains advice-seeking until $\lambda = 4$, and DeBERTa keeps top-1 retrieval until $\lambda = 6$ before reverting to no advice at $\lambda = 8$. This staggered transition is exactly the pattern the joint expert-advice formulation is designed to capture.

Figures 6 and 7 show how our method allocates mass across advice actions. As the advice cost increases, the policy shifts away from expensive advice, but it does not do so uniformly. The reallocation toward cheaper actions remains expert-dependent.

Tables 8 and 9 show directly how the deployment-optimal expert-advice pair changes with the cost regime. At $\lambda = 2$, the lowest loss is achieved by DeBERTa with top-1 retrieval. At $\lambda = 10$, the optimum shifts to the cheapest no-query expert. What matters operationally is therefore the full deferral-advice loss, not retrieval quality in isolation.

C.4. Sensitive-Escalation Experimental Details

This appendix reports the tabular sensitive-escalation experiment that is deferred from the main paper. The objective is the same as in FEVER, but in a different regime: advice corresponds to progressively more sensitive feature groups rather than retrieved text. The question is whether the composite expert-advice formulation still improves the true deployment objective once the system must trade predictive value against both expert fees and escalation costs.

Data, action space, and metric. We use the IEEE-CIS fraud detection benchmark (Howard et al., 2019). Each example is a payment event with tabular transaction fields, optional identity attributes, and a binary fraud label. A useful operational picture is an online card transaction for which the system initially observes only a restricted low-sensitivity view, such as

Algorithm 4 FEVER inference example

Require: Claim x, learned composite scorer $s_\pi(\cdot, \cdot; \widehat{\theta})$
Define the FEVER expert set $\{$Qwen3-4B-Instruct, Qwen3-8B, DeBERTa-v3$\}$
Define the advice actions $\{0$, top-1, top-3, top-5, reformulate+top-5$\}$
Select the executed pair

$$(\widehat{j}, \widehat{k}) \leftarrow \underset{(j,k)\in[3]\times[4]_0}{\arg\max} \; s_\pi(x, (j,k); \widehat{\theta})$$

if $\widehat{k} = 0$ **then**
  Set $\widetilde{a}^{(\widehat{k})} \leftarrow \widetilde{a}^{(0)}$ and issue no retrieval request
else if $\widehat{k} =$ top-1 **then**
  Run BM25 on claim x and keep the top-1 retrieved passage
  Mask all non-selected advice slots to form $\widetilde{a}^{(\widehat{k})}$
else if $\widehat{k} =$ top-3 **then**
  Run BM25 on claim x and keep the top-3 retrieved passages
  Mask all non-selected advice slots to form $\widetilde{a}^{(\widehat{k})}$
else if $\widehat{k} =$ top-5 **then**
  Run BM25 on claim x and keep the top-5 retrieved passages
  Mask all non-selected advice slots to form $\widetilde{a}^{(\widehat{k})}$
else
  Use the reformulation model to rewrite the claim query
  Mask all non-selected advice slots to form $\widetilde{a}^{(\widehat{k})}$
end if
Route the claim and masked advice to expert $\widehat{j}$
Output the FEVER label $e_{\widehat{j}}(x, \widetilde{a}^{(\widehat{k})}) \in \{$SUPPORTS, REFUTES, NOT ENOUGH INFO$\}$
return executed pair $\widehat{\pi}(x) = (\widehat{j}, \widehat{k})$ and prediction $e_{\widehat{j}}(x, \widetilde{a}^{(\widehat{k})})$

Table 8. Expert-by-advice true deferral-advice loss at $\lambda = 2$. Lower is better.

	k=0	k=1	k=2	k=3	k=4
Qwen3-4B-Instruct	0.695	0.726	0.741	0.727	0.710
Qwen3-8B	0.715	0.746	0.749	0.699	0.713
DeBERTa-v3-large	0.708	**0.641**	0.655	0.677	0.668

coarse merchant, amount, and transaction metadata. Before sending the case to an expert, it may decide whether to reveal additional information. For instance, the policy may request contact metadata, a payment-profile summary, a device profile, or stricter identity attributes if these extra signals are worth their escalation cost. We follow the split used in the experiment configuration and train on $120{,}000$ examples with an $8{,}000$-example validation set. With three experts and five advice actions, the composite action space again has size 3×5. As in FEVER, all policy comparisons follow the global protocol of Section C.1; in particular, the reported numbers are validation averages of the true deferral-advice loss.

Experts and advice construction. The expert pool contains a logistic-regression model, a histogram GBDT, and an MLP, with consultation costs $\beta = (0.02, 0.06, 0.09)$. Advice does not take the form of retrieved text here. Instead, it corresponds to revealing additional feature groups beyond the base low-sensitivity transaction view. The advice levels are nested: each action reveals all information from the previous levels and then adds one more group. Concretely, the five advice actions are:

- $k = 0$: no additional information.
- $k = 1$: the base view plus low-security contact metadata.
- $k = 2$: the information in $k = 1$ plus medium-security payment-profile attributes.
- $k = 3$: the information in $k = 2$ plus high-security device-profile attributes.
- $k = 4$: the information in $k = 3$ plus strict identity attributes.

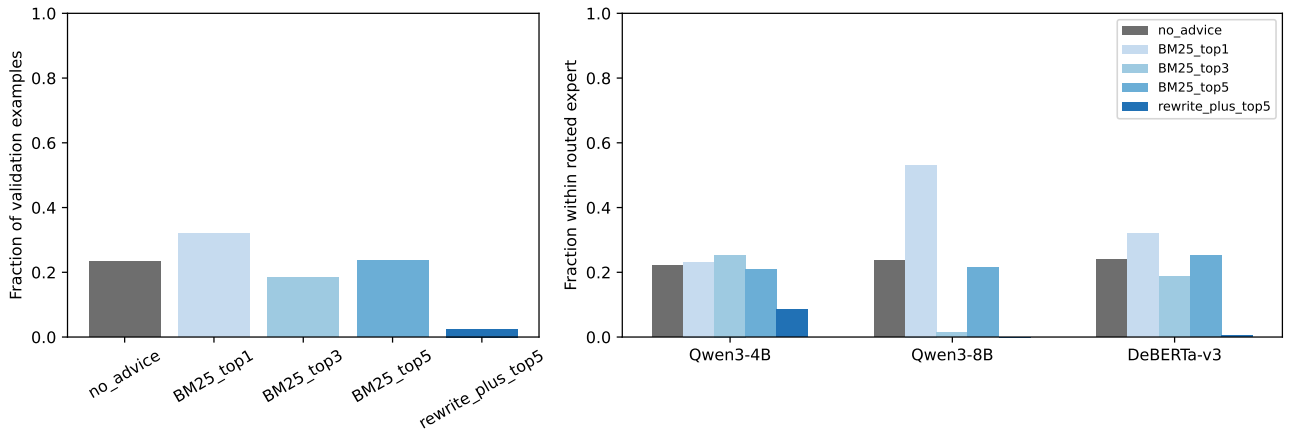

Figure 6. Deployed advice distribution of our method at $\lambda = 0$ in FEVER. The left panel shows the overall fraction of validation examples assigned to each advice action. The right panel conditions on the routed expert. When advice is free, the learned policy uses queried advice on most examples, but the preferred advice level already varies across experts.

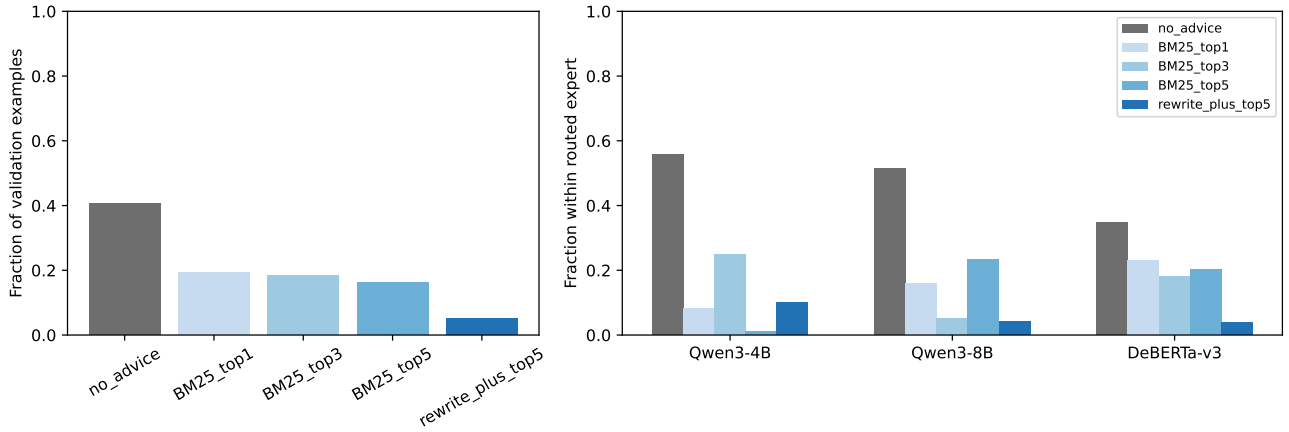

Figure 7. Deployed advice distribution of our method at $\lambda = 2$ in FEVER. Compared with Figure 6, the policy shifts mass away from expensive advice actions, but it does not collapse uniformly to no advice. The expert-conditional panel shows that the transition remains expert-dependent.

Their base costs are $\gamma^{\text{base}} = (0, 0.08, 0.12, 0.16, 0.20)$, multiplied by $\lambda \in \{0, 0.04, 0.12, 0.20, 5.0\}$. Richer advice can improve fraud detection, but each escalation level carries a larger operational and privacy cost.

The experts themselves are trained offline in a way that matches this nested advice protocol. For each training example, we do not duplicate the example across all five advice levels. Instead, we sample one advice level uniformly from $\{0, 1, 2, 3, 4\}$, reveal exactly the corresponding prefix of feature groups, and train the expert on that masked view augmented with binary reveal indicators. This produces one memory-safe pooled training set in which each expert sees a mixture of low- and high-information cases. The logistic model uses balanced class weights, the histogram GBDT uses class-balanced sample weights, and the MLP additionally oversamples the minority class. After this offline training stage, each expert is evaluated under all five advice levels on the train and validation splits to populate the executed expert-advice cost tensor used by the policy learner.

Policy training and baselines. Input features are the standardized tabular representation (identity features with z-score normalization). All learned methods — our augmented surrogate and the L2D baseline — share the same backbone: a single-hidden-layer MLP with hidden dimension 64, ReLU activation, and no dropout. They are trained with AdamW (learning rate 10^{-3}, no weight decay, gradient clipping at norm 10) for 4,000 epochs with batch size 512 and a cosine learning-rate scheduler with 5% linear warmup. The augmented comp-sum surrogate uses $\tau = 1$; the L2D baseline uses the same architecture but is restricted to the k=0 slice and trained with the standard cross-entropy deferral surrogate. For every example and every pair (j, k), we precompute the expert prediction and the resulting realized cost, so all methods are trained and evaluated on exactly the same expert-advice outcome tensor. The best-fixed, partial-randomization, and joint-random

Table 9. Expert-by-advice true deferral-advice loss at $\lambda = 10$. Lower is better.

	k=0	k=1	k=2	k=3	k=4
Qwen3-4B-Instruct	**0.695**	0.846	0.901	0.967	0.790
Qwen3-8B	0.715	0.866	0.909	0.939	0.793
DeBERTa-v3-large	0.708	0.761	0.815	0.917	0.748

baselines are exactly those defined in Section C.1, instantiated on this fraud-detection expert pool and advice set.

Reading the appendix sensitive-escalation tables. The tables below play the same role as their FEVER counterparts, but in a different modality. Table 10 first asks how each expert behaves when advice cost changes: which advice action is best for that expert, and how quickly does it revert to no advice as escalation becomes expensive? Table 11 then evaluates full policies on the true deployment metric. The detailed tables at $\lambda = 0$, $\lambda = 0.12$, and $\lambda = 5$ make the cost-dependent expert-advice optimum explicit.

Table 10. Best advice action for each expert across the retained sensitive-escalation cost grid. Each cell reports the best advice index k and the corresponding true loss. Here $k = 0$ denotes no advice, $k = 1$ low-security metadata, $k = 2$ medium-security profile, $k = 3$ high-security device profile, and $k = 4$ strict identity attributes.

Expert	$\lambda = 0$	$\lambda = 0.04$	$\lambda = 0.12$	$\lambda = 0.20$	$\lambda = 5$
Linear	k=4 / 0.308	k=3 / 0.314	k=0 / 0.325	k=0 / 0.325	k=0 / 0.325
Tree	k=4 **/ 0.271**	k=4 **/ 0.279**	k=2 **/ 0.291**	k=2 **/ 0.301**	k=0 **/ 0.317**
MLP	k=2 / 0.359	k=2 / 0.364	k=0 / 0.369	k=0 / 0.369	k=0 / 0.369

Table 10 highlights the same phenomenon as in FEVER in a tabular setting. The best advice action depends both on the expert and on the cost regime. When escalation is cheap, both the linear model and the tree expert prefer the most revealing strict-security tier (k=4), while the MLP prefers the medium-security tier (k=2). As the escalation cost increases, these preferences shift back toward cheaper advice or no advice, but not at the same rate for all experts. This is precisely the kind of expert-dependent trade-off that motivates a joint expert-advice policy.

Table 11. Validation true deferral-advice loss in the sensitive-escalation experiment. Entries are mean $\pm$ standard deviation over 4 random seeds. Lower is better.

Method	λ=0	λ=0.04	λ=0.12	λ=0.20	λ=5
Our	**0.260 $\pm$ 0.001**	**0.267 $\pm$ 0.001**	**0.274 $\pm$ 0.002**	**0.279 $\pm$ 0.002**	**0.281 $\pm$ 0.001**
L2D	0.282 $\pm$ 0.001	0.282 $\pm$ 0.001	0.282 $\pm$ 0.001	0.282 $\pm$ 0.001	0.282 $\pm$ 0.001
Best fixed pair	0.271	0.279	0.291	0.301	0.317
Learned expert, random advice	0.287 $\pm$ 0.003	0.297 $\pm$ 0.003	0.304 $\pm$ 0.003	0.313 $\pm$ 0.003	0.837 $\pm$ 0.004
Random expert, learned advice	0.310 $\pm$ 0.003	0.314 $\pm$ 0.003	0.322 $\pm$ 0.003	0.325 $\pm$ 0.004	0.338 $\pm$ 0.003
Random (j, k)	0.325 $\pm$ 0.003	0.330 $\pm$ 0.003	0.339 $\pm$ 0.003	0.348 $\pm$ 0.003	0.885 $\pm$ 0.006

Table 11 shows a consistent pattern across the entire cost sweep. Our method attains the lowest mean true loss at every value of λ, with the largest gains over L2D when escalation is cheap and only a small residual gap at $\lambda = 5$. This is the expected pattern as the optimal policy moves gradually toward no-advice actions. Our method also stays below the best fixed pair throughout the sweep. That fixed pair changes with the cost regime, moving from the tree expert with strict-security advice at $\lambda \in \{0, 0.04\}$, to the same expert with medium-security advice at $\lambda \in \{0.12, 0.20\}$, and finally to the no-advice tree expert at $\lambda = 5$. The partial-randomization ablations further show that the gain comes from coordinating expert selection and advice acquisition rather than from either component alone; in particular, randomizing the advice becomes especially damaging at high cost, because unnecessary escalation is then heavily penalized.

Figures 8 and 9 show how the policy allocates mass across advice levels. The transition with λ is not simply "query" versus "no-query." As advice becomes more expensive, the learned policy first moves toward cheaper advice levels and only then collapses toward no advice, which explains why the augmented and L2D risks nearly coincide at $\lambda = 5$.

Tables 12, 13, and 14 show how the deployment-optimal action changes with the escalation cost. At $\lambda = 0$, the tree expert with the strict-security advice tier is best. At $\lambda = 0.12$, the optimum has already moved to the medium-security tier for that same expert. At $\lambda = 5$, the best action is the no-advice tree expert, and all queried actions are much more costly. The relevant choice is therefore the expert-advice pair with smallest true deferral-advice loss for the current value of λ.

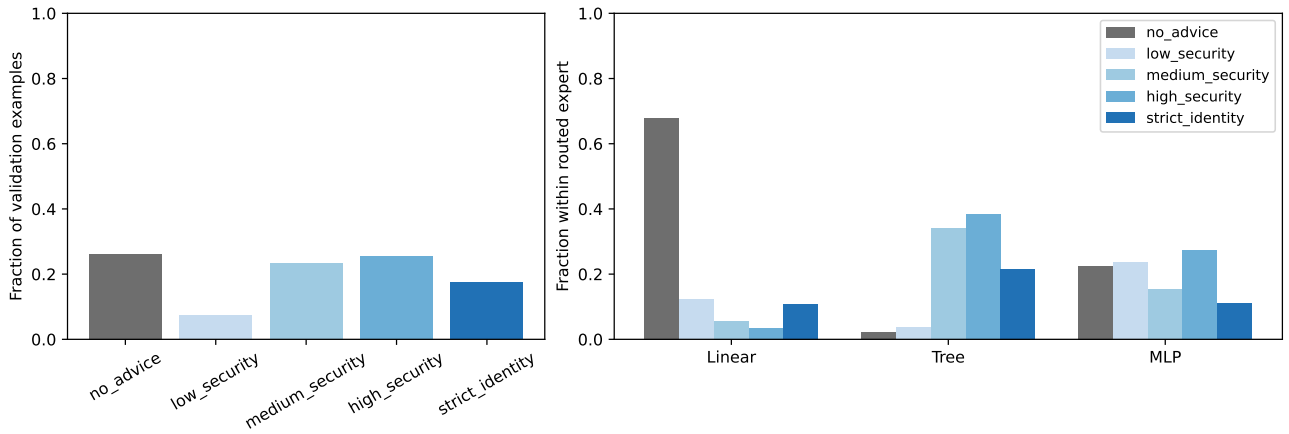

Figure 8. Deployed advice distribution of our method at $\lambda = 0.12$. The left panel shows the overall fraction of validation examples assigned to each advice action. The right panel conditions on the routed expert. At this moderate cost, the learned policy still uses queried advice frequently, but the preferred advice level already depends on the chosen expert.

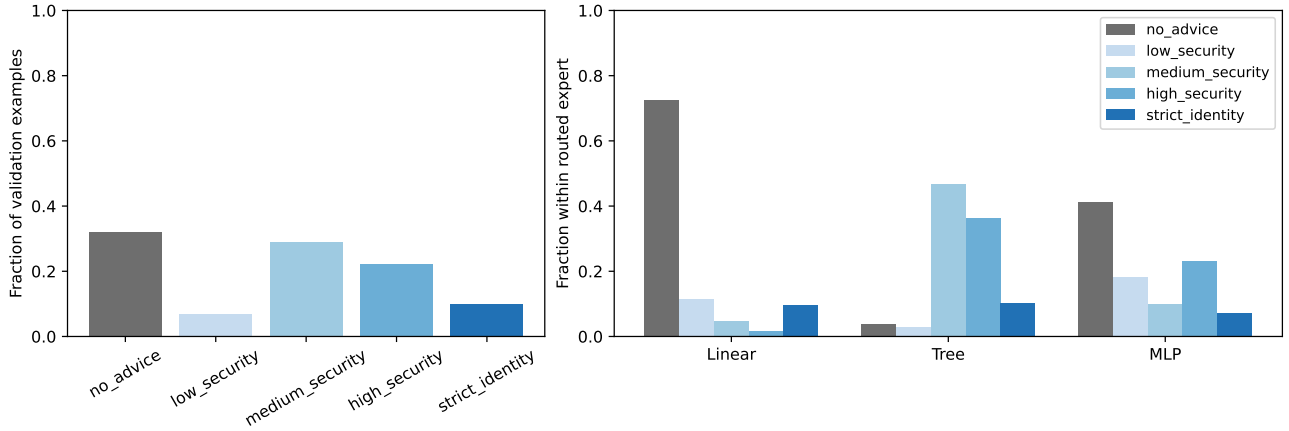

Figure 9. Deployed advice distribution of our method at $\lambda = 0.20$. Compared with Figure 8, the policy shifts further toward cheaper advice and no advice. The right panel shows that this transition is still expert-dependent rather than uniform across the routed experts.

C.5. CLIP Prompt-Escalation Experimental Details

The setting again matches the deferral-advice protocol, but advice now corresponds to increasingly rich prompt families for frozen CLIP (Radford et al., 2021) experts. The question is whether a joint expert-advice policy improves the true deployment objective when richer prompt sets may improve zero-shot accuracy but also incur an explicit consultation cost.

Data, action space, and metric. We use ImageNet-1k (Russakovsky et al., 2015). Each example is an image paired with one of 1000 class labels. A useful way to read this experiment is as a cost-sensitive zero-shot classification problem: given an image of, say, a goldfish, the system may either classify it with a single canonical prompt such as "a photo of a goldfish" or pay for a richer prompt family before sending the image to a CLIP expert. The retained run uses $39{,}222$ training examples and $14{,}288$ validation examples. With three experts and four advice actions, the composite action space has size 3×4. As in FEVER and sensitive-escalation, all policy comparisons follow the global protocol of Section C.1; in particular, the reported numbers are validation averages of the true deferral-advice loss.

Experts and advice construction. The expert pool contains three frozen CLIP models: CLIP-B/32, CLIP-B/16, and CLIP-L/14, with consultation costs $\beta = (0, 0.04, 0.08)$. Advice does not reveal new image features. Instead, it enlarges the prompt family used to build the text classifier for a given expert. The four advice actions are:

- $k = 0$: no advice, using a single canonical prompt.
- $k = 1$: a clean prompt ensemble with 5 closely related templates.
- $k = 2$: a medium-diversity family with 16 templates.

Table 12. Expert-by-advice true deferral-advice loss at $\lambda = 0$. Lower is better.

	k=0	k=1	k=2	k=3	k=4
Linear	0.325	0.322	0.316	0.308	0.308
Tree	0.317	0.301	0.277	0.273	**0.271**
MLP	0.369	0.382	0.360	0.377	0.372

Table 13. Expert-by-advice true deferral-advice loss at $\lambda = 0.12$. Lower is better.

	k=0	k=1	k=2	k=3	k=4
Linear	0.325	0.332	0.330	0.327	0.332
Tree	0.317	0.310	**0.291**	0.292	0.295
MLP	0.369	0.391	0.374	0.396	0.396

- $k = 3$: a large-diversity family with 32 templates.

These advice levels are nested in the practical sense that larger prompt families include and extend the simpler ones. Their base costs are $\gamma^{\text{base}} = (0, 0.03, 0.06, 0.09)$, multiplied by $\lambda \in \{0, 0.1, 0.15, 0.3, 0.4, 0.6, 0.7, 0.8, 1.0, 1.5, 10.0\}$. This produces a transparent trade-off: richer prompt families can improve zero-shot accuracy, but only at an additional advice cost.

Prompt examples. It is useful to make these advice actions concrete. For an image whose class is `goldfish`, the no-advice action uses only the canonical template "a photo of a goldfish." The clean-ensemble advice augments this with closely related variants such as "a clean photo of a goldfish," "a centered photo of a goldfish," and "a high quality photo of a goldfish." The medium-diversity advice expands further to prompts such as "a close-up photo of a goldfish," "a bright photo of a goldfish," "a blurry photo of a goldfish," and "a black and white photo of a goldfish." Finally, the large-diversity advice adds still broader views, for instance "a rendering of a goldfish," "a sketch of a goldfish," "a nighttime photo of a goldfish," and "a front-view photo of a goldfish." In this experiment, acquiring advice therefore means paying to evaluate the image against a richer prompt family before handing it to the selected CLIP expert.

Policy training and baselines. Input features are frozen CLIP-B/32 image embeddings (dimension 512), standardized with z-score normalization. All learned methods — our augmented surrogate and the L2D baseline — share the same backbone: a two-layer MLP with hidden dimensions $(512, 256)$, ReLU activations, and no dropout. They are trained with AdamW (weight decay 10^{-3}, gradient clipping at norm 10) for 400 epochs with batch size 256 and a cosine learning-rate scheduler with 5% linear warmup. We use a learning rate of 10^{-2} for L2D, and 10^{-3} for our approach. The augmented comp-sum surrogate uses $\tau = 1$; the L2D baseline uses the same architecture but is restricted to the k=0 slice and trained with the standard cross-entropy deferral surrogate. For every image, every expert, and every prompt family, we precompute the expert prediction and the realized cost, so all methods are trained and evaluated on exactly the same expert-advice outcome tensor. The best-fixed, partial-randomization, and joint-random baselines are exactly those defined in Section C.1, instantiated on this CLIP expert pool and prompt-family advice set.

Reading the appendix CLIP tables. The tables below mirror those of the previous subsections. Table 15 first shows how each CLIP expert's preferred advice action changes with the advice cost. Table 16 then evaluates full policies on the true deployment metric, and the detailed tables at $\lambda = 0$, $\lambda = 0.15$, and $\lambda = 10$ make the cost-dependent expert-advice optimum explicit.

Table 15 shows the same structural pattern as the FEVER and tabular experiments. The best advice action depends on both the expert and the cost regime. When advice is cheap, all three experts prefer queried prompt families, although not the same ones: the two smaller CLIP models move through the medium-diversity family, whereas CLIP-L/14 already switches to the cheaper clean ensemble at $\lambda = 0.1$ and $\lambda = 0.15$. Once the advice cost increases further, all three experts revert to the no-advice action, but not at the same point. This is again the regime in which a joint expert-advice policy is most natural.

Table 16 shows a clear cost-dependent transition. Our method attains the lowest mean true loss throughout the low- and moderate-cost regime, improving on both L2D and the best fixed pair. As λ increases, the gap narrows, which is expected because the optimal behavior approaches the no-advice regime. At $\lambda = 10$, our method ties L2D, consistent with the fact that querying has become too expensive to justify and the learned policy ceases to acquire advice. The partial-randomization baselines support the same conclusion as in the other experiments: the gain comes from coordinating expert choice and advice choice rather than from either component in isolation.

Figures 10 and 11 show how the queried mass is distributed across prompt families. As the advice cost increases, the policy

Table 14. Expert-by-advice true deferral-advice loss at $\lambda = 5$. Lower is better.

	k=0	k=1	k=2	k=3	k=4
Linear	0.325	0.722	0.916	1.108	1.308
Tree	**0.317**	0.701	0.877	1.073	1.271
MLP	0.369	0.782	0.960	1.177	1.372

Table 15. Best advice action for each expert across the retained CLIP cost grid. Each cell reports the best advice index k and the corresponding true loss. Here $k = 0$ denotes the canonical no-advice prompt, $k = 1$ the clean ensemble, $k = 2$ the medium-diversity prompt family, and $k = 3$ the large-diversity prompt family.

Expert	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.15$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 10$
CLIP-B/32	k=3 / 0.391	k=2 / 0.398	k=2 / 0.401	k=0 / 0.403	k=0 / 0.403	k=0 / 0.403	k=0 / 0.403	k=0 / 0.403	k=0 / 0.403	k=0 / 0.403	k=0 / 0.403
CLIP-B/16	k=3 / 0.378	k=2 / 0.386	k=2 / 0.389	k=0 / 0.394	k=0 / 0.394	k=0 / 0.394	k=0 / 0.394	k=0 / 0.394	k=0 / 0.394	k=0 / 0.394	k=0 / 0.394
CLIP-L/14	k=3 / **0.351**	k=1 / **0.359**	k=1 / **0.361**	k=1 / **0.365**	k=0 / **0.366**	k=0 / **0.366**	k=0 / **0.366**	k=0 / **0.366**	k=0 / **0.366**	k=0 / **0.366**	k=0 / **0.366**

moves toward cheaper prompt families and then toward no advice, and it does so differently for different routed experts.

Table 17. Expert-by-advice true deferral-advice loss at $\lambda = 0$. Lower is better.

	k=0	k=1	k=2	k=3
CLIP-B/32	0.403	0.397	0.392	**0.391**
CLIP-B/16	0.394	0.385	0.380	**0.378**
CLIP-L/14	0.366	0.356	0.354	**0.351**

Table 18. Expert-by-advice true deferral-advice loss at $\lambda = 0.15$. Lower is better.

	k=0	k=1	k=2	k=3
CLIP-B/32	0.403	0.402	**0.401**	0.405
CLIP-B/16	0.394	0.390	**0.389**	0.391
CLIP-L/14	0.366	**0.361**	0.363	0.364

Table 19. Expert-by-advice true deferral-advice loss at $\lambda = 10$. Lower is better.

	k=0	k=1	k=2	k=3
CLIP-B/32	**0.403**	0.697	0.992	1.291
CLIP-B/16	**0.394**	0.685	0.980	1.278
CLIP-L/14	**0.366**	0.656	0.954	1.251

Tables 17, 18, and 19 show how the deployment-optimal prompt family changes with the advice cost. At $\lambda = 0$, all three experts prefer the richest prompt family. At $\lambda = 0.15$, the optimum has already shifted to smaller prompt sets. By $\lambda = 10$, the no-advice action is uniformly optimal in true loss. The relevant decision is therefore the prompt family with smallest deferral-advice loss, not the richest family available.

Table 16. Validation true deferral-advice loss in the CLIP prompt-escalation experiment. Entries are mean ± standard deviation over 4 random seeds. Lower is better.

λ	Our	L2D	Best fixed pair	Learned expert, random advice	Random expert, learned advice	Random (j, k)
0	**0.337 ± 0.000**	0.352 ± 0.001	0.351	0.343 ± 0.001	0.374 ± 0.002	0.379 ± 0.003
0.1	**0.345 ± 0.000**	0.352 ± 0.001	0.359	0.348 ± 0.001	0.381 ± 0.002	0.383 ± 0.003
0.15	**0.348 ± 0.001**	0.352 ± 0.001	0.361	0.350 ± 0.001	0.384 ± 0.002	0.386 ± 0.003
0.3	**0.350 ± 0.002**	0.352 ± 0.001	0.365	0.357 ± 0.001	0.387 ± 0.002	0.392 ± 0.003
0.4	**0.350 ± 0.002**	0.352 ± 0.001	0.366	0.361 ± 0.001	0.388 ± 0.003	0.397 ± 0.003
0.6	**0.350 ± 0.002**	0.352 ± 0.001	0.366	0.370 ± 0.001	0.388 ± 0.002	0.406 ± 0.003
0.7	**0.350 ± 0.002**	0.352 ± 0.001	0.366	0.374 ± 0.001	0.388 ± 0.002	0.410 ± 0.003
0.8	**0.351 ± 0.002**	0.352 ± 0.001	0.366	0.379 ± 0.001	0.388 ± 0.002	0.415 ± 0.003
1	**0.351 ± 0.001**	0.352 ± 0.001	0.366	0.388 ± 0.001	0.388 ± 0.002	0.424 ± 0.003
1.5	**0.351 ± 0.001**	0.352 ± 0.001	0.366	0.410 ± 0.001	0.388 ± 0.002	0.446 ± 0.003
10	0.352 ± 0.001	0.352 ± 0.001	0.366	0.795 ± 0.003	0.388 ± 0.002	0.829 ± 0.004

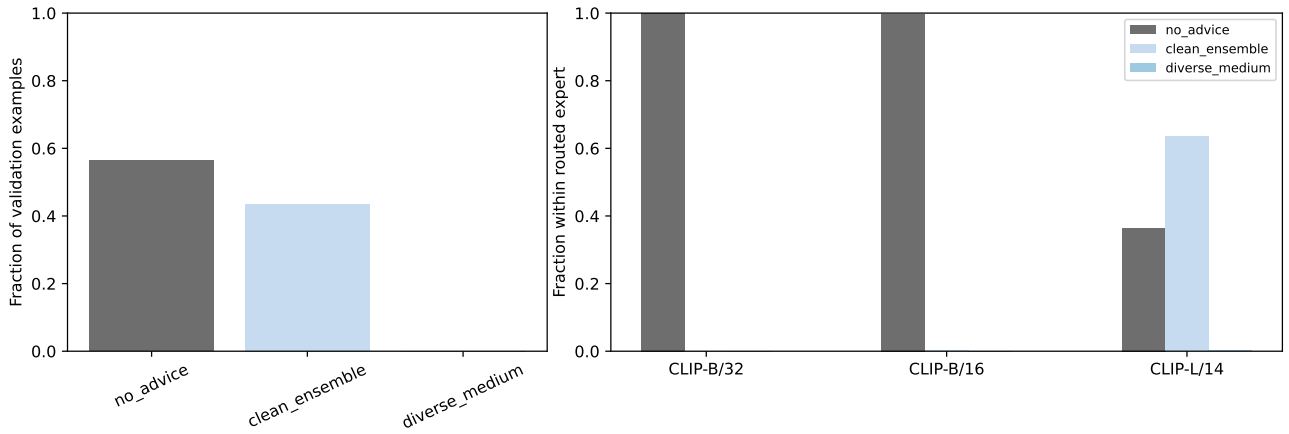

Figure 10. Deployed advice distribution of our method at $\lambda = 0.3$ in the CLIP prompt-escalation experiment. The left panel shows the overall fraction of validation images assigned to each advice action. The right panel conditions on the routed expert. At this cost level, the policy still queries frequently, but it already favors cheaper prompt families than in the free-advice regime.

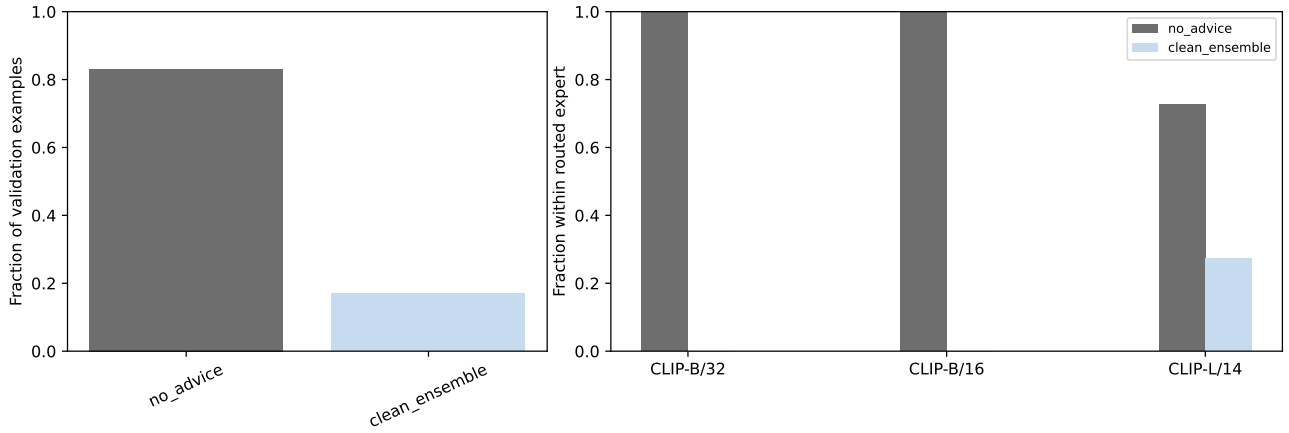

Figure 11. Deployed advice distribution of our method at $\lambda = 0.4$ in the CLIP prompt-escalation experiment. Compared with Figure 10, the policy already shifts further toward no advice and the cheapest prompt families. The expert-conditional panel shows that this transition remains expert-dependent rather than uniform across routed experts.