# HUMAN-AI COLLABORATIVE UNCERTAINTY QUANTIFICATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

AI predictive systems are becoming integral to decision-making pipelines, shaping high-stakes choices once made solely by humans. Yet robust decisions under uncertainty still depend on capabilities that current AI lacks: domain knowledge not captured by data, long-horizon context, and the ability to reason and act in the physical world. This contrast has sparked growing efforts to design *collaborative* frameworks that combine the complementary strengths of both agents. This work advances this vision by identifying the fundamental principles of Human-AI collaboration in the context of uncertainty quantification—an essential component of any reliable decision-making pipeline. We introduce Human-AI Collaborative Uncertainty Quantification, a framework that formalizes how an AI model can refine a human expert's proposed prediction set with two goals in mind: *avoiding counterfactual harm*, ensuring the AI does not degrade the human's correct judgments, and *complementarity*, enabling the AI to recover correct outcomes the human missed. At the population level, we show that the optimal collaborative prediction set takes the form of an intuitive two-threshold structure over a single score function, extending a classical result in conformal prediction. Building on this insight, we develop practical offline and online calibration algorithms with provable *distribution free* finite-sample guarantees. The online algorithm adapts to *any* distribution shifts, including the interesting case of human behavior evolving through interaction with AI, a phenomenon we call "Human-to-AI Adaptation." We validate the framework across three modalities—image classification, regression, and text-based medical decision-making—using models from convolutional networks to LLMs. Results show that collaborative prediction sets consistently outperform either agent alone, achieving higher coverage and smaller set sizes across various conditions, including shifts in human behavior.

## 1 INTRODUCTION

Artificial intelligence has demonstrated extraordinary predictive power, enabling data-driven decision-making in high-stakes domains such as healthcare, law, and autonomous systems. These systems excel at extracting patterns from vast amounts of data, offering statistical accuracy and consistency at a scale unattainable by human reasoning alone. Yet, robust decision-making in such settings requires more than predictive accuracy. Human experts contribute domain knowledge beyond data (Hansen & Quinon, 2023), persistent memory for long-term planning and context (Bengio et al., 1994), and the ability to reason and act within the physical world in ways still inaccessible to current AI systems (Agrawal, 2010). These complementary strengths point to the importance of human-AI collaboration, where computational precision and human judgment can jointly guide decisions under uncertainty.

A central challenge in realizing this vision lies in uncertainty quantification (UQ). Precise characterization of uncertainty is fundamental to robust decision-making, as it allows decision-makers to weigh risks, assess reliability, and allocate trust between human and machine. While UQ has been extensively studied in the machine learning community, these efforts largely focus on AI systems in isolation. In collaborative settings, however, it is not clear what principles of UQ should be when humans and AI are jointly in the loop. Identifying these principles is essential for designing frameworks that achieve the best of both worlds: combining AI's predictive accuracy with human judgment to enable decisions more robust and effective than either could do alone. To this end, we ask:

*What should characterize a successful collaboration between a human expert and an AI system?*

Two principles naturally emerge. First, the expert must trust the collaboration to even be willing to engage: the AI's contribution should not degrade the quality of the human's input. In other words, collaborating with AI should not make the outcome worse in the worst case—a notion we refer to as *counterfactual harm*. Second, collaboration must offer clear benefits beyond what the expert could achieve alone. The AI should *complement* the human by addressing blind spots, identifying correct outcomes that may have been overlooked, and thereby strengthening the overall decision process. Together, these two principles, trust through non-degradation and benefit through complementarity, capture the essential properties of a meaningful human–AI collaborative framework.

In this work, motivated by recent advances in conformal prediction (Vovk et al., 2005; Lei et al., 2017; Romano et al., 2019; 2020; Angelopoulos et al., 2022), we develop a framework that instantiates these two principles in the context of collaborative prediction sets. This allows us to design distribution-free sets that respect both principles without assumptions on the behavior of the AI model or the human, making the approach particularly practical for modern applications. Additionally, recent work shows that conformal prediction sets are essential for risk-sensitive decision making, where decisions must account for predictive uncertainty in a principled way (Kiyani et al., 2025). This makes prediction sets an especially compelling subject of study for human–AI collaboration in high-stakes domains such as healthcare.

**Proposed Framework.** We propose a framework for *human–AI collaborative uncertainty quantification*, where the two agents jointly construct a prediction set. Formally, let $(X, Y) \sim \mathcal{P}$, where $X \in \mathcal{X}$ denotes the observed features and $Y \in \mathcal{Y}$ the corresponding label. The goal is to construct, for each input $x$, a set $C(x) \subseteq \mathcal{Y}$ that contains the true label $Y$ with high probability while remaining as small as possible.

In our collaborative setting, a human expert first proposes an initial set of plausible outcomes $H(x) \subseteq \mathcal{Y}$, based on their expertise. The AI system then refines this proposal by outputting a prediction set $C(x, H(x)) \subseteq \mathcal{Y}$, designed to complement the human input. For notational convenience, we drop the explicit dependence on $H(x)$ in what follows, and have $C(x) := C(x, H(x))$.

This modification is guided by two principles. The first is *low counterfactual harm*: the AI should not degrade the quality of the human proposal. Concretely, whenever the true label lies within the human's proposed set, the AI's refinement must preserve high coverage,

$$\mathbb{P}(Y \notin C(X) \mid Y \in H(X)) \; < \varepsilon.$$

The second is *complementarity*: the AI should add value precisely when the human misses the correct outcome. That is, with high probability, the AI's refinement recovers the true label whenever it is excluded from the human proposal,

$$\mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \; \geq 1 - \delta.$$

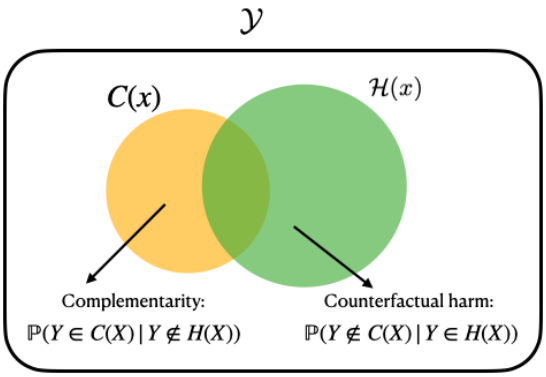



Figure 1: Schematic of the two guiding principles

These two principles are illustrated schematically in Figure 1. Together, they formalize a collaborative prediction strategy: the AI preserves the human's expertise while compensating for potential blind spots. They come together in the following optimization problem, which serves as the collaboration framework we study in this work:

**Human-AI Collaboration Optimization (HACO)**

Let $(X, Y) \sim \mathsf{P}$ and $H(x) \subseteq \mathcal{Y}$ be a human-proposed set. Let the prediction set returned by the AI be denoted as $C(x, H(x)) := C(x) \subseteq \mathcal{Y}$. The Human-AI Collaboration Optimization (HACO) problem is

$$\begin{aligned} \min_{C:\mathcal{X}\to 2^{\mathcal{Y}}} \quad & \mathbb{E}\,|C(X)| \\ \text{s.t.} \quad & \mathbb{P}(Y \notin C(X) \mid Y \in H(X)) \; < \varepsilon, \\ & \mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \; \geq \; 1 - \delta, \end{aligned} \tag{HACO}$$

where $\varepsilon$ and $\delta$ are two user-defined thresholds.

At a high level, the goal of prediction sets is to include the correct label with high probability while keeping the sets small — set size serving as the measure of efficiency in uncertainty quantification. Within our framework, the AI contributes in two complementary ways: pruning and augmentation. On the one hand, the AI prunes labels from the human proposal whenever possible, since smaller sets are more informative, but does so without violating the counterfactual harm constraint. On the other hand, the AI augments the set by adding likely labels that the human may have overlooked, thereby ensuring complementarity. The human contribution, in turn, is to provide the AI with a stronger starting point. When the initial human-proposed sets are of high quality, the AI's final sets achieve the same coverage level with significantly smaller size than what either could have produced in isolation.

**Preview of Results.**

- We characterize the optimal solution to HACO in Section 4. As we will show, the optimal solution takes the intuitive form of "two thresholds over one score function", one threshold for pruning labels in the human set, and the other guides the labels that we will add to the human set. We will then build upon this result to design conformity scores that will be used by our finite sample algorithm. In particular, for the case of regression, our score is a novel extension of conformalized quantile regression (Romano et al., 2019).

- In Section 4, we derive practical finite sample algorithms with provable distribution-free guarantees, in two settings of offline, where the calibration and test data are separated and exchangeable, and online, where the data is streamed one by one. Notably, in the online setting, our algorithm also captures the novel concept of "Human-to-AI Adaptation", which might be of its own interest and a promising subject for further studies.
- In Section 5, we evaluate our finite sample offline and online algorithms on three data modalities: image classification, text based medical diagnosis, and real-valued regression. Across all settings, we show that the parameters $\varepsilon$ and $\delta$ can be tuned such that the collaborative prediction set outperforms both human and AI-only baselines, achieving higher coverage, smaller size, or both. We vary human and AI strength to study each component's role and test robustness under various distribution shifts.

### 1.1 Related Works

We briefly discuss closely related works here and defer a full discussion to Appendix A. In the context of the human–AI collaboration, a growing line of work studies prediction sets as advice to experts (Straitouri & Rodriguez, 2024; Straitouri et al., 2023; Cresswell et al., 2024; Zhang et al., 2024; Paat & Shen, 2025). For instance, Straitouri et al. (2023) propose improving expert predictions with conformal prediction sets, Babbar et al. (2022) show empirically that set-valued advice can boost human accuracy, and Straitouri et al. (2024) analyze such systems through the lens of counterfactual harm. These works differ from ours in that they study how humans use AI-provided sets and evaluate downstream human accuracy, but do not construct a final *collaborative* prediction set that algorithmically integrates human feedback with AI. A complementary literature on *learning to defer* allocates instances between models and experts (Madras et al., 2018; Mozannar & Sontag, 2021; Okati et al., 2021; Verma & Nalisnick, 2022). This also differs from our goal in that we do not optimize who decides on each instance; instead, we collaboratively quantify uncertainty by combining the human's initial set with AI to return a single, joint prediction set with explicit safeguards (e.g., counterfactual harm and complementarity constraints).

## 2 Optimal Prediction Sets over Population

We begin by characterizing the optimal solution to the optimization problem HACO, the problem introduced in Section 1, in the infinite-sample regime, where the data distribution $\mathcal{P}$ is fully known. This characterization uncovers the statistical framework that we will later use to design finite-sample algorithms, enabling us to tune the dynamics of Human-AI collaboration with fine control over counterfactual harm and the complementarity rate of the collaboration procedure.

**Theorem 2.1.** *The optimal solution to HACO is of the form*

$$C^*(x) = \big\{\, y :\; 1 - p(y \mid x) \;\le\; a^*\, \mathbf{1}\{y \notin H(x)\} \;+ b^*\, \mathbf{1}\{y \in H(x)\} \,\big\}, \quad \textit{a.s. for any } x \in \mathcal{X},$$

*for some thresholds* $a^*, b^* \in \mathbb{R}$.

The theorem shows that the optimal collaborative prediction set can be described by two thresholds: One, $b^*$, which is responsible for *pruning*, i.e., for the labels $y \in H(x)$, $b^*$ determines which ones we keep and which ones we exclude; And the other, $a^*$, which is responsible for *augmenting* new labels, i.e., for the labels $y \notin H(x)$, $a^*$ determines which ones to add to the final set. In other words, we include all labels whose $p(y \mid x)$ exceeds a threshold, and that threshold depends on whether the label was originally proposed by the human. If $y \in H(x)$, then the AI uses a threshold $b^*$, and if $y \notin H(x)$, the AI instead applies a different threshold $a^*$.

This theorem generalizes prior results on minimum set size conformal prediction (Sadinle et al., 2019; Kiyani et al., 2024). When the human set always includes all the labels or is always empty—essentially the two cases in which the human set carries no information about the true label—the optimal set reduces to a one-scalar characterization of the form $\big\{\, y :\; 1 - p(y \mid x) \;\le\; q^* \,\big\}$, which corresponds to minimum set size conformal prediction.

In what follows, we take advantage of the result of this theorem to design an algorithmic framework for Human-AI collaboration. In particular, in the characterization given by Theorem 2.1, there are three components that need to be approximated or estimated in a finite-sample setting: $p(y \mid x)$, $a^*$, and $b^*$. As we will see, the AI's role will be to provide an approximation of $p(y \mid x)$. In the next section, we will discuss this in the two different settings of classification and regression. We will then discuss debiasing strategies to estimate $a^*$ and $b^*$ from data.

## 3 Conformal Scoring Rules

Building on the results of Theorem 2.1, our goal is to construct prediction sets of the form

$$C^*(x) = \big\{\, y :\; s(x, y) \;\le\; a^*\, \mathbf{1}\{y \notin H(x)\} \;+ b^*\, \mathbf{1}\{y \in H(x)\} \,\big\},$$

where $s(x, y)$ is a **non-conformity score** that measures how unusual a label $y$ is for a given input $x$. In the infinite-sample regime, Theorem 2.1 shows that the optimal non-conformity score is $s(x, y) = 1 - p(y \mid x)$ where $p(y \mid x)$

is the true conditional distribution. However, since $p(y \mid x)$ is unknown in practice, we design a non-conformity score to approximate the behavior of the optimal score. Below, we describe how such scores can be constructed for both classification and regression settings.

**Classification** In classification tasks, predictive models typically output a probability distribution over labels, often obtained via a softmax layer. Formally, let $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ map each input $x \in \mathcal{X}$ to a $|\mathcal{Y}|$-dimensional vector of probabilities $\hat{p}(y \mid x)$, which approximates the true conditional probabilities $p(y \mid x)$. A widely used non-conformity score in classification (Sadinle et al., 2019) that we adopt in our framework is defined as

$$\hat{s}(x, y) = 1 - \hat{p}(y \mid x),$$

**Regression** In regression, the continuous label space makes it difficult to estimate the full conditional distribution $p(y \mid x)$, so directly approximating the optimal score $1 - p(y \mid x)$ is not straightforward. To circumvent this, we build upon *Conformalized Quantile Regression* (CQR) (Romano et al., 2019). The idea of CQR is to estimate lower and upper conditional quantiles of $Y$ given $X = x$ and then use them to construct a conformal score. Suppose we obtain an estimate $\hat{q}_{\alpha/2}$ of the $\alpha/2$ quantile of the distribution of $Y \mid X = x$, and an estimate $\hat{q}_{1-\alpha/2}$ for the $1 - \alpha/2$ quantile. We can then define the score

$$\hat{s}(x, y) \;=\; \max\Big\{\hat{q}_{\alpha/2}(x) - y,\;\; y - \hat{q}_{1-\alpha/2}(x)\Big\},$$

and use this to make prediction sets. One can verify that the resulting prediction sets are a calibrated version of $\big[\hat{q}_{\alpha/2},\; \hat{q}_{1-\alpha/2}\big]$ (either expanded or shrunk symmetrically). The intuition is that the CQR score remains small within the learned central quantile band and increases linearly into the tails. For common unimodal distributions, this ordering is approximately monotone with $1 - p(y \mid x)$, so thresholding the CQR score closely emulates the optimal rule. Prediction sets of this form have shown strong performance in terms of average set size in practice. We generalize the idea behind CQR to design a score function tailored to our two-threshold setting. The idea is to learn two distinct sets of quantile functions: one for counterfactual harm when $Y \in H(x)$ and one for complementarity when $Y \notin H(x)$. To achieve this, we learn two pairs of quantile functions, $(\hat{q}_{\varepsilon/2}, \hat{q}_{1-\varepsilon/2})$ for the counterfactual-harm constraint and $(\hat{q}_{\delta/2}, \hat{q}_{1-\delta/2})$ for the complementarity constraint. From these quantile estimates, we define the nonconformity score as

$$\hat{s}(x, y) := \begin{cases} \max\{\hat{q}_{\varepsilon/2}(x) - y,\; y - \hat{q}_{1-\varepsilon/2}(x)\}, & y \in H(x), \\ \max\{\hat{q}_{\delta/2}(x) - y,\; y - \hat{q}_{1-\delta/2}(x)\}, & y \notin H(x). \end{cases}$$

This score treats labels inside $H(x)$ differently from those outside it, applying a distinct CQR-style score to each in an intuitive manner: for $y \in H(x)$, the score is derived from the counterfactual-harm rate $\varepsilon$; for $y \notin H(x)$, it is derived from the complementarity rate $1 - \delta$.

## 4 Finite Sample Algorithms

So far, we have shown that optimal collaborative prediction sets are of the form $C^*(x) = \big\{ y : \; s(x, y) \;\leq\; a^* \, \mathbf{1}\{y \notin H(x)\} \; + b^* \, \mathbf{1}\{y \in H(x)\} \big\}$, where we have also discussed strategies for designing the score $s$ in both regression and classification. In this section, we fix the conformity score and focus on how to estimate the thresholds $a$ and $b$ from data. We introduce *Collaborative Uncertainty Prediction*-(CUP), our algorithmic framework for constructing collaborative prediction sets in finite samples. We consider two scenarios: (i) the offline setting, where calibration and test data are assumed exchangeable, and the task is to estimate thresholds on a held-out calibration set before evaluating on future points.; and (ii) the online setting, where data arrives sequentially and the underlying distribution may drift in arbitrary and unknown ways.

### 4.1 CUP - Offline

In the offline setting, we assume access to a held-out calibration dataset $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i, H(X_i))\}_{i=1}^{n}$ that is exchangeable with the test data $\mathcal{D}_{\text{test}} = \{(X_j, Y_j, H(X_j)))\}_{j=1}^{m}$. The goal is to estimate the thresholds $(\hat{a}, \hat{b})$ that implement the two-threshold structure of Theorem 2.1. For each calibration point $(x_i, y_i)$, we compute a non-conformity score $s_i = s(x_i, y_i)$, and separate the scores into two groups according to whether the true label lies in the human set or not. The thresholds are then obtained by taking empirical quantiles of these two groups:

$$\hat{b} = \text{Quantile}_{1-\varepsilon}\Big(\{\, s_i : Y_i \in H(X_i) \,\} \cup \{\infty\}\Big), \qquad \hat{a} = \text{Quantile}_{1-\delta}\Big(\{\, s_i : Y_i \notin H(X_i) \,\} \cup \{\infty\}\Big).$$

Given a new test input $x_{\text{test}}$, the collaborative prediction set is formed as

$$C(x_{\text{test}}) \;=\; \{\, y : s(x_{\text{test}}, y) \leq \hat{a} \cdot \mathbf{1}\{y \notin H(x_{\text{test}})\} + \hat{b} \cdot \mathbf{1}\{y \in H(x_{\text{test}})\}\}.$$

The following Proposition shows these sets satisfy finite-sample guarantees.

**Proposition 4.1** (Finite-Sample Offline Guarantees). *Let $(X_{n+1}, Y_{n+1}, H(X_{n+1}))$ be a new test point, exchangeable with the calibration data. Let $n_1$ be the number of calibration points where $Y_i \in H(X_i)$ and $n_2$ the number where $Y_i \notin H(X_i)$. The thresholds $\hat{a}$ and $\hat{b}$ satisfy:*

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1}) \mid Y_{n+1} \in H(X_{n+1})) \geq 1-\varepsilon \quad \textit{and} \quad \mathbb{P}(Y_{n+1} \in C(X_{n+1}) \mid Y_{n+1} \notin H(X_{n+1})) \geq 1-\delta.$$

*Additionally, if the conformity scores have continuous distribution, then:*

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1}) \mid Y_{n+1} \in H(X_{n+1})) < 1-\varepsilon+\frac{1}{n_1+1},$$
$$\mathbb{P}(Y_{n+1} \in C(X_{n+1}) \mid Y_{n+1} \notin H(X_{n+1})) < 1-\delta+\frac{1}{n_2+1}.$$

The assumption of exchangeability for lower bound and continuity for upper bounds are both common in the conformal prediction literature (e.g., Vovk et al. (2005)).

In practice the assumption of exchangeability is fragile and real-world deployments may inevitably face distribution shifts that undermine the validity of offline guarantees. Such shifts may stem from many sources, but in the context of long-term human-AI collaboration, a particularly salient one is what we call *Human-to-AI Adaptation*. As collaboration unfolds, humans may gradually adjust how they construct their proposed sets $H(x)$ in response to the AI's behavior. For instance, the human might learn over time which types of instances–such as which patients in a medical setting–the AI tends to be more knowledgeable about, and tune their proposals accordingly to be maximally helpful to the final set. In some cases, this may mean proposing larger sets to improve coverage, while in others it may mean offering smaller, more decisive sets to sharpen outcomes. Such feedback loops alter the distribution of test-time data in ways that violate exchangeability between calibration and test sets. This motivates the need for robustness to evolving distributions in collaborative settings. To address this challenge, we now turn to the *online setting*, which relaxes exchangeability and explicitly allows the data distribution to evolve over time.

### 4.2 CUP - Online

In this Section, we move to the online setting where data arrives sequentially, one sample at a time. At each round $t$, the test input $x_t$ and the human's proposed set $H(x_t)$ are provided to the AI, which must then output the final prediction set $C_t(x_t)$. Only after the final prediction set is announced is the true label $y_t$ revealed. Here, we make no assumptions about the distribution of the data stream, an assumption particularly natural for human-AI collaboration, where distribution shift is not merely accidental but may arise directly from the interaction itself.

We design an online algorithm, CUP–Online, that makes prediction sets of the form,

$$C_t(x_t) = \big\{ y \in \mathcal{Y} \mid s(x_t, y) \leq a_t \, \mathbf{1}\{y \notin H(x_t)\} + b_t \, \mathbf{1}\{y \in H(x_t)\} \big\},$$

where $s(.,.)$ is a fixed non-conformity score (look at Section 3), and $(a_t, b_t)$ are the two thresholds that we will update in an online fashion. Let us also define

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}\{y_t \notin C_t(x_t),\ y_t \in H(x_t)\}, \quad \mathrm{err}_t^{\mathrm{out}} := \mathbf{1}\{y_t \notin C_t(x_t),\ y_t \notin H(x_t)\}.$$

Then, fixing a learning rate $\eta > 0$, CUP–Online updates only one threshold at a time, depending on whether the human included the true label in their proposed set.

$$\begin{aligned} &if \quad y_t \in H(x_t): \quad b_{t+1} = b_t + \eta\,(\mathbf{1}\{s(x_t, y_t) > b_t\} - \varepsilon), \quad a_{t+1} = a_t \\ &if \quad y_t \notin H(x_t): \quad a_{t+1} = a_t + \eta\,(\mathbf{1}\{s(x_t, y_t) > a_t\} - \delta), \quad b_{t+1} = b_t \end{aligned}$$

Intuitively, if errors occur more often than expected, the threshold is relaxed to include more labels, if errors are too rare, the threshold is tightened. Over time this feedback process drives the empirical error rates toward their target values $\varepsilon$ and $\delta$. The choice of $\eta$ gives a tradeoff between adaptability and stability, while larger values will make the method more adaptive to observed distribution shifts (this will also show up in our guarantees) they also induce greater volatility in thresholds values, which may be undesirable in practice as it will allow the method to fluctuate between smaller sets to larger sets. Hence, in practice, a careful hyperparameter tuning for $\eta$ can enhance the performance of CUP–Online. We now outline the theoretical guarantees of our online algorithm.

**Proposition 4.2** (Finite-Sample Guarantees). *Assume the conformity is bounded, i.e, $s(x, y) \in [0, 1]$ and let $N_1(T) = \sum_{t=1}^{T} \mathbf{1}\{y_t \in H(x_t)\}$ and $N_2(T) = \sum_{t=1}^{T} \mathbf{1}\{y_t \notin H(x_t)\}$. For any $T \geq 1$:*

$$\left| \frac{1}{N_1(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} - \varepsilon \right| \leq \frac{1 + \eta \max(\varepsilon, 1-\varepsilon)}{\eta N_1(T)}, \quad \left| \frac{1}{N_2(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{out}} - \delta \right| \leq \frac{1 + \eta \max(\delta, 1-\delta)}{\eta N_2(T)}.$$

*In particular, if $N_1(T), N_2(T) \to \infty$, then almost surely*

$$\lim_{T\to\infty} \frac{1}{N_1(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} = \varepsilon, \qquad \lim_{T\to\infty} \frac{1}{N_2(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{out}} = \delta.$$

*Remark.* The boundedness assumption on the conformity score holds automatically in classification when $s$ is derived from a probability output (e.g., a softmax score, which lies in $[0, 1]$). In regression, where scores may be unbounded, this condition can be enforced by rescaling and clipping the score.

These types of update rules and guarantees are common in the online conformal prediction literature for controlling marginal coverage (Gibbs & Candès, 2021; Angelopoulos et al., 2023). We extend these ideas to simultaneously control counterfactual harm and complementarity rates. Our results show that over long intervals, CUP–online achieves the desired rates without any assumption on the data-generating distribution. In particular, the algorithm addresses human–to-AI adaptation, among other shifts, by decoupling validity from assumptions about human behavior. By tracking how human proposals interact with prediction set errors and adjusting its thresholds accordingly, CUP–online ensures that the target error rates are maintained, even as human strategies evolve over time.

## 5 Experiments

First, we outline our experimental setup, and then evaluate our framework across three distinct data modalities: (i) image classification, (ii) real-valued regression, and (iii) text based medical decision-making with large language models. For each modality, we study both the offline and online algorithms introduced in Section 4.

**Baselines.** We compare against the following natural baselines: *(i) Human alone.* which uses the human-proposed set $H(x)$ directly, without any AI refinement. We treat the human policy as a black box and make no assumptions about how the sets are generated. Coverage depends entirely on the provided sets and may vary with human expert quality These human sets are constructed using crowd-sourced annotations, rule-based diagnostic systems or synthetic noise, depending on the task. Full details are provided in each experiment subsection. *(ii) AI alone*, which uses the AI system without incorporating human input, reducing to standard conformal prediction based solely on the model scores. This provides a benchmark for how well the AI performs independently. Additionally, in the online setting we consider a fixed baseline that serves as a reference point for detecting and evaluating distribution shifts. This method uses a static set of thresholds computed from an initial subset of data (i.e early examples or a dedicated split), and then applies these thresholds over the online data stream without any further updates. This baseline provides a useful comparison to understand the value of adaptivity in the online setting.

**Evaluation metrics.** Across all experiments, we evaluate methods based on two key quantities: *marginal coverage*, the probability that the true label lies in the prediction set, and *average set size*, measured as cardinality in classification and interval length in regression. In the online setting, we use *running* versions of these metrics, defined at each time step $t$ as $\widehat{\mathrm{cov}}_t = \frac{1}{t}\sum_{i=1}^{t} \mathbf{1}\{y_i \in C(x_i)\}$ for marginal coverage and $\widehat{\mathrm{size}}_t = \frac{1}{t}\sum_{j=1}^{t} |C(x_j)|$ for average set size. These metrics capture the central tradeoff in uncertainty quantification: higher coverage is desirable, but must be balanced against set informativeness. Our algorithm does not explicitly enforce a fixed marginal coverage. Instead, the counterfactual harm parameter $\varepsilon$ and the complementarity parameter $\delta$ shape the resulting coverage and set size. By adjusting these parameters, we can navigate tradeoffs between the two metrics.

A successful human–AI collaboration should improve upon the human baseline in at least one dimension, coverage or set size, without significantly worsening the other. For example, it may increase coverage while avoiding large increases in set size, or shrink the set without losing coverage. In the best case, both metrics improve together. The better the AI model, the more effectively it should recover missed outcomes without unnecessarily inflating sets. Similarly, the stronger the human baseline, the better the collaborative procedure can perform, since it starts from a higher-quality initial proposal. Thus our framework reflects the complementary contributions of both human and AI, and we will explore this dependence on human and AI quality across our experimental tasks.

### 5.1 Classification: ImageNet-16H

Our first set of experiments use the ImageNet-16H dataset (Steyvers et al., 2022), which captures human prediction behavior under varying perceptual noise. It consists of 32,431 human predictions on 1,200 natural images, each annotated by multiple participants and perturbed with one of four noise levels $\omega \in \{80, 95, 110, 125\}$ that progressively increase task difficulty. The label space is restricted to a fixed set of 16 classes. For the AI component, we use a pre-trained VGG19 classifier (Simonyan & Zisserman, 2015) fine-tuned for 10 epochs. We evaluate our framework an offline setting and subsequently in an online setting, where we introduce various distribution shifts.

**Offline Setting.** We compare three approaches: *Human Alone*, *AI Alone*, and CUP-offline. Results are averaged over 10 random calibration/test splits. Table 1 reports coverage and set size under two representative noise levels,

$\omega = 95$ and $\omega = 125$. For the human baseline, we aggregate multiple annotations into empirical label frequencies and form top-$k$ sets by selecting the $k$ most frequently chosen labels. From the algorithm's perspective, only the sets—not raw annotations or confidences—are observed. The AI baseline applies standard conformal prediction without human input. Since conformal methods allow direct control over target coverage, we evaluate AI Alone at the same realized coverage achieved by CUP-offline. This ensures a fair comparison, where the only meaningful dimension for improvement is set size (i.e., if CUP achieves the same coverage with smaller sets, it shows that human input is being used effectively to tighten predictions). CUP-offline incorporates both sources, with coverage and size determined by $(\varepsilon, \delta)$ parameters that encode counterfactual harm and complementarity.

| $\omega$ = 125 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Human Alone** | | **CUP** | | | | **AI Alone** | |
| Strategy | Coverage | Size | Coverage | Size | $\varepsilon$ | $\delta$ | Coverage | Size |
| Top-2 | $0.8008 \pm 0.0090$ | $2.00 \pm 0.00$ | $\mathbf{0.9022 \pm 0.0083}$ | $\mathbf{1.49 \pm 0.04}$ | 0.05 | 0.70 | $0.9072 \pm 0.0138$ | $1.65 \pm 0.07$ |
| Top-1 | $0.7245 \pm 0.0103$ | $1.00 \pm 0.00$ | $\mathbf{0.8823 \pm 0.0134}$ | $1.36 \pm 0.07$ | 0.05 | 0.70 | $0.8828 \pm 0.0140$ | $1.48 \pm 0.05$ |

| $\omega$ = 95 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Human Alone** | | **CUP** | | | | **AI Alone** | |
| Strategy | Coverage | Size | Coverage | Size | $\varepsilon$ | $\delta$ | Coverage | Size |
| Top-2 | $0.9613 \pm 0.0061$ | $2.00 \pm 0.00$ | $\mathbf{0.9825 \pm 0.0066}$ | $\mathbf{1.77 \pm 0.44}$ | 0.01 | 0.80 | $0.9830 \pm 0.0061$ | $2.10 \pm 0.15$ |
| Top-1 | $0.9257 \pm 0.0060$ | $1.00 \pm 0.00$ | $\mathbf{0.9763 \pm 0.0076}$ | $1.43 \pm 0.07$ | 0.01 | 0.80 | $0.9755 \pm 0.0053$ | $2.27 \pm 0.21$ |

Table 1: **ImageNet-16H – Offline Results:** Comparison of Human, AI, and CUP under two noise levels. Reports marginal coverage and average set size (mean ± std over 10 splits). CUP uses calibration parameters $(\varepsilon, \delta)$.

We include two noise levels to evaluate performance under varying task difficulty for the human experts. As shown in Table 1, across both levels, our CUP-offline consistently improves on the human baseline. When the human sets are relatively large(e.g top -2), CUP-offline yields strict improvements in both dimensions, reducing set size while improving coverage. At $\omega = 125$, for example, human top-2 sets cover 80% of labels with size 2.0, whereas CUP-offline improves coverage to 90% while reducing size to 1.49. When human sets are very small (e.g., top-1), coverage improvements typically requires adding labels, slightly increasing set size. Even then, CUP-offline offers more efficient sets than AI Alone, leveraging human input to achieve better tradeoffs. At $\omega = 95$, for example, CUP-offline achieves 97.6% coverage with an average size of 1.43, whereas AI Alone requires size 2.27 for similar coverage. Overall, CUP-offline improves on raw human sets and produces tighter predictions than AI Alone, adapting to the strengths and limits of each source to provide a clear advantage over both baselines.

**Online Setting.** We now turn to the online setting, where the data arrives sequentially and distributional shifts may occur during deployment. We consider two types of shifts: a *noise shift*, where inputs are ordered from high to low noise levels ($\omega = 125 \to 95$), and a *human strategy shift*, where human prediction sets evolve from top-2 to top-3 strategies. The latter serves as a concrete instance of what we term *Human-to-AI Adaptation* which is this case is how humans might adapt their behavior in response to increasing task difficulty or AI feedback.

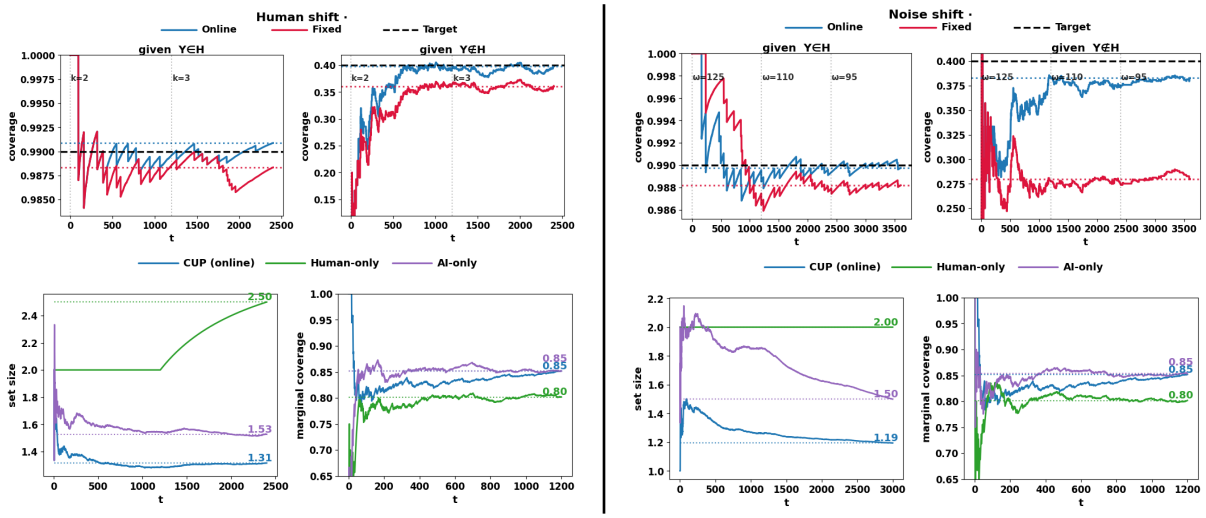



Figure 2: **ImageNet-16H- Online Results:** Performance under human strategy shift (left) and noise shift (right). Top: running coverage for CUP-online vs fixed baseline. Bottom: CUP-online vs human-only and AI-only baselines on running set size and marginal coverage.

We first compare CUP-online to the fixed baseline tuned on a separate segment of the data stream. For instance, in the noise shift setting, we tune $(a, b)$ on $\omega = 80$, and for the human shift, on top-1 human prediction sets. To evaluate, we track *constraint-specific coverage* over time. At each time step $t$, we compute $\mathrm{cov}_t^* = 1 - 1/t \sum_{i=1}^{t} 1 - \mathrm{err}_i^*$, where $\mathrm{err}_i^*$ is either a counterfactual harm error or a complementarity error defined in Section 4.2. Intuitively, this metric tracks how well the algorithm maintains the target coverage level over time. When the algorithm is effective, this running estimate converges to the nominal targets $1 - \varepsilon$ and $1 - \delta$.

Figure 2 (top row) shows the results for both forms of distribution shift: human strategy shift (left) and noise shift (right). In both cases, the online algorithm remains close to the target coverage levels throughout the stream, while the fixed baseline drifts away and fails to recover from the changes in the underlying distribution. In the bottom row of Figure 2, we compare CUP-online with human-only and AI-only baselines, using the running marginal coverage and set size metrics defined earlier. For a fair comparison, we run the AI-only baseline at a target coverage level matched to the realized coverage achieved by CUP-online across the full stream. The results show that CUP-online consistently improves over the human baseline by achieving higher coverage while keeping the prediction sets small. Compared to AI alone, where coverage is matched by design, CUP-online produces more compact sets. These trends mirror those seen in the offline setting, showing that our online collaborative procedure maintains the advantages of the framework under distribution shift.

### 5.2 LLMs for Medical Diagnosis Decision Making

Our second set of experiments evaluates the framework in the text modality of data, focusing on a medical decision-making task using the DDXPlus dataset (Fansi Tchango et al., 2022). This dataset contains synthetic patient records generated from a medical knowledge base and rule-based diagnostic system. Each record includes demographics, symptoms, and antecedents linked to an underlying condition, along with a differential diagnosis list. From this list we form human prediction sets using a top-$k$ strategy, where the human provides the $k$ most likely diagnoses. For the AI component, we use two language models with contrasting accuracy: **GPT-5**, which performs strongly, and **GPT-4o**, which is weaker and often falls below the human baseline. This contrast highlights how the quality of the AI model shapes the trade-offs of collaboration.

**Offline Setting.** Tables 2 summarize the results for GPT-4o and GPT-5, respectively, under two different human strategies. Across all settings, CUP-offline improves on the human baseline by raising coverage, as the procedure explicitly augments human sets when the true label is missing. Naturally, this may increase set size, but when the AI is sufficiently strong, as with GPT-5, the algorithm is able to both prune away incorrect human labels and add the correct label when necessary more efficiently. This yields prediction sets that improve across both dimensions: achieving higher coverage *and* smaller size, outperforming both baselines.

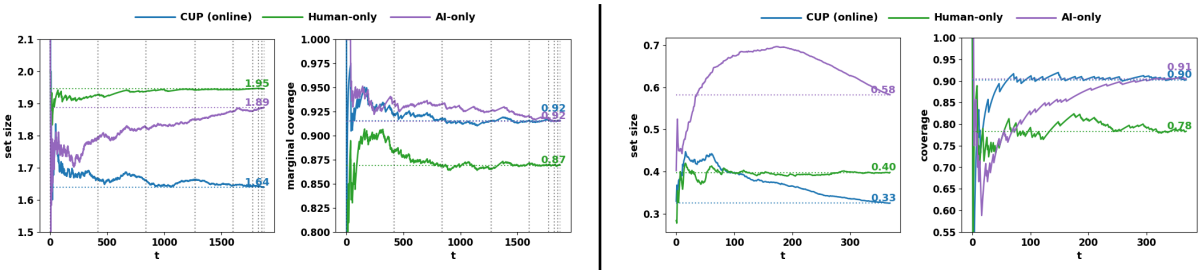
| | **Human** | **GPT-4o** | | | **GPT-5** | | |
|---|---|---|---|---|---|---|---|
| **Strategy** | **C/S** | **CUP C/S** | $(\varepsilon, \delta)$ | **AI C/S** | **CUP C/S** | $(\varepsilon, \delta)$ | **AI C/S** |
| Top-1 | 0.71 / 1.00 | 0.90 / 2.84 | (0.02, 0.70) | 0.88 / 4.64 | 0.91 / 1.59 | (0.02, 0.70) | 0.91 / 1.76 |
| Top-2 | 0.87 / 1.95 | 0.93 / 3.14 | (0.01, 0.45) | 0.90 / 9.12 | 0.93 / 1.65 | (0.02, 0.45) | 0.93 / 1.95 |

Table 2: **LLMs-Offline Results** Entries report coverage/size (C/S). Calibration parameters $(\varepsilon, \delta)$ shown for CUP.

With a weaker model such as GPT-4o, coverage gains may come at the cost of slightly larger sets, reflecting that the model is less capable of efficient pruning or complementarity. This does not undermine the approach but rather illustrates the role of the AI component in determining the ultimate efficiency of the collaborative sets. Still, CUP-offline produces smaller sets than AI alone at comparable coverage levels, showing that human knowledge is being used productively. Taken together, these results demonstrate that CUP-offline yields consistent benefits over both the human and AI baselines, while the degree to which coverage and set size can be simultaneously optimized depends on the strength of the AI model.

**Online Setting.** We next evaluate the CUP-online algorithm in the medical setting, using GPT-5 as the AI model. Human prediction sets follow a top-$k = 2$ strategy, and distribution shift is induced by ordering test patients by age, from younger to older groups. We also include a non-adaptive baseline with fixed thresholds tuned on the earliest segment (ages 1–30) to isolate the effect of adaptivity in the face of demographic change, however due to space constraints direct comparison with this baseline is deferred to the Appendix C.2.

As in the ImageNet experiments, we benchmark CUP-online against human-only and AI-only baselines. Figure 3 shows results under demographic shift. The pattern is consistent with the offline setting: CUP improves over the human baseline in both coverage and set size, and against AI Alone it achieves smaller sets at matched coverage.




**LLM (DDXPlus) under age shift**

**Regression - (Communities & Crime) under demographic shift**

Figure 3: Online results: LLM (left) and regression (right), comparing CUP–online with baselines Human and AI.

### 5.3 Regression: Communities & Crime

Our final set of experiments evaluates the framework in a regression setting using the UCI *Communities & Crime* dataset (Redmond, 2002), where the goal is to predict the violent crime rate per community. To simulate human input, we generate intervals centered around noisy point estimates of the ground truth. Specifically, we perturb each true label with Gaussian noise to form $\hat{y}(x)$, then construct the interval $H(x) = [\,\hat{y}(x) - w(x)/2,\ \hat{y}(x) + w(x)/2\,]$, where $w(x)$ is a base width also subject to noise. By varying the noise levels, we simulate human experts of differing quality. As in earlier experiments, the algorithm only observes the final set $H(x)$, not how it was generated. The AI model builds on the setting explained in Section 3.We train two MLPs using the pinball loss to estimate conditional quantiles: one for predicting $(\hat{q}_{\varepsilon/2}, \hat{q}_{1-\varepsilon/2})$, and one for $(\hat{q}_{\delta/2}, \hat{q}_{1-\delta/2})$. Each model shares a backbone with two output heads. The resulting four quantiles define the CQR-style score in Section 3, to which we apply the CUP-offline procedure to obtain the two thresholds used at test time.

We compare *Human Alone*, which uses the raw intervals $H(x)$; *AI Alone*, which applies standard conformalized quantile regression without access to the human sets; and CUP-offline, which combines both sources via the proposed collaborative algorithm. Results are reported in Table 3 for two human experts of different quality. First,

| **Human A** | | | | **Human B** | | | |
|---|---|---|---|---|---|---|---|
| **Human A C/S** | **CUP C/S** | $(\varepsilon,\ 1-\delta)$ | **AI C/S** | **Human B C/S** | **CUP C/S** | $(\varepsilon,\ 1-\delta)$ | **AI C/S** |
| 0.760 / 0.581 | 0.862 / 0.380 | (0.10, 0.70) | 0.862 / 0.394 | 0.872 / 0.618 | 0.948 / 0.528 | (0.05, 0.90) | 0.948 / 0.608 |
| 0.760 / 0.581 | 0.825 / 0.326 | (0.15, 0.70) | 0.825 / 0.337 | 0.872 / 0.618 | 0.953 / 0.558 | (0.05, 0.95) | 0.953 / 0.588 |

Table 3: Regression–Offline Results: Coverage/size (C/S) under two human expert settings.

we note that CUP improves upon the human baseline in terms of both coverage and interval width. Second, the results highlight the complementary role of human, with greater gains observed over AI Alone when initial human input is of higher quality. This complements the medical diagnosis results, where we varied the AI instead of the human. Together, the two experiments show that the collaboration efficiency depends on the quality of both parties.

**Online Setting.** We evaluate CUP-online under a controlled distribution shift based on community demographics. Test examples are ordered by the proportion of residents identified by a randomly selected race-coded variable. Figure 3 reports the running marginal coverage and average set size for AI, Human, and CUP-online. Consistent with the previous experiments, we again observe that CUP-online improves upon both baselines: compared to Human Alone, it increases coverage without inflating intervals; compared to AI Alone, it reduces interval width while preserving coverage. The non-adaptive baseline with fixed thresholds is tuned on the earliest portion of the stream (i.e., communities with the lowest demographic proportion). Results for this baseline are deferred to Appendix 7. All together, the results underscore the robustness of the collaborative approach across modalities and under shifting data distributions.

## 6 Conclusion

We introduced a framework for constructing prediction sets collaboratively between humans and AI, grounded in two core principles: avoiding counterfactual harm and enabling complementarity. We showed that the optimal sets take a simple two-threshold form, and developed finite-sample algorithms for both offline and online settings. Across diverse domains and agents (human or AI) strengths, our methods consistently leverage the collaboration capabilities of human and AI to produce sets that outperform either alone. This framework offers a principled and practical approach to structured collaboration under uncertainty.

## REFERENCES


Kush Agrawal. To study the phenomenon of the moravec's paradox. *arXiv preprint arXiv:1012.3148*, 2010.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction, 2022. URL https://arxiv.org/abs/2009.14193.

Anastasios N. Angelopoulos, Emmanuel J. Candes, and Ryan J. Tibshirani. Conformal pid control for time series prediction, 2023. URL https://arxiv.org/abs/2307.16895.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2025. URL https://arxiv.org/abs/2208.02814.

Adrian Arnaiz-Rodriguez, Nina Corvelo Benz, Suhas Thejaswi, Nuria Oliver, and Manuel Gomez-Rodriguez. Towards human-ai complementarity in matching tasks, 2025. URL https://arxiv.org/abs/2508.13285.

Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, November 1976. doi: 10.1214/aos/1176343654. URL https://doi.org/10.1214/aos/1176343654.

Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams, 2022. URL https://arxiv.org/abs/2205.01411.

Tim Bary, Benoît Macq, and Louis Petit. No need for "learning" to defer? a training free deferral framework to multiple experts through conformal prediction, 2025. URL https://arxiv.org/abs/2509.12573.

Sander Beckers, Hana Chockler, and Joseph Y. Halpern. Quantifying harm, 2022. URL https://arxiv.org/abs/2209.15111.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. doi: 10.1109/72.279181.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2972–3005. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/charusaie22a.html.

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, pp. 1532–1543, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718222. URL https://doi.org/10.1145/3717823.3718222.

Santiago Cortes-Gomez, Carlos Patiño, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification, 2025. URL https://arxiv.org/abs/2410.01767.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making, 2024. URL https://arxiv.org/abs/2401.13744.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. Towards human-ai complementarity with prediction sets. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 31380–31409. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/37d4d4413b7c7558cc27a6d3d42ea998-Paper-Conference.pdf.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31306–31318. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cae73a974390c0edd95ae7aeae09139c-Paper-Datasets_and_Benchmarks.pdf.

Joel Feinberg. Wrongful life and the counterfactual element in harming. *Social Philosophy and Policy*, 4(1):145–178, 1986. doi: 10.1017/S0265052500000467.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification, 2024. URL https://arxiv.org/abs/2208.12084.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift, 2021. URL https://arxiv.org/abs/2106.00170.

Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, July 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108496. URL http://dx.doi.org/10.1016/j.patcog.2021.108496.

J. Hansen and P. Quinon. The importance of expert knowledge in big data and machine learning. *Synthese*, 201 (35), 2023. doi: 10.1007/s11229-023-04041-5.

Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction, 2024. URL https://arxiv.org/abs/2406.18814.

Shayan Kiyani, George Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents, 2025. URL https://arxiv.org/abs/2502.02561.

Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: A survey of empirical studies, 2021. URL https://arxiv.org/abs/2112.11471.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2017. URL https://arxiv.org/abs/1604.04173.

Jordan Lekeufack, Anastasios N. Angelopoulos, Andrea Bajcsy, Michael I. Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions, 2024. URL https://arxiv.org/abs/2310.05921.

Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. Trustworthy policy learning under the counterfactual no-harm criterion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20575–20598. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23ay.html.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J. Pappas. Safe planning in dynamic environments using conformal prediction, 2023. URL https://arxiv.org/abs/2210.10254.

David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, 1969.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018. URL https://arxiv.org/abs/1711.06664.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert, 2021. URL https://arxiv.org/abs/2006.01862.

Sima Noorani, Orlando Romero, Nicolo Dal Fabbro, Hamed Hassani, and George J. Pappas. Conformal risk minimization with variance reduction, 2025. URL https://arxiv.org/abs/2411.01696.

Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage, 2021. URL https://arxiv.org/abs/2103.08902.

Helbert Paat and Guohao Shen. Conformal set-based human-ai complementarity with multiple experts, 2025. URL https://arxiv.org/abs/2508.06997.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pp. 345–356. Springer, 2002.

Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2002. DOI: https://doi.org/10.24432/C53W3X.

Jonathan G. Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm, 2022. URL https://arxiv.org/abs/2204.12993.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression, 2019. URL https://arxiv.org/abs/1905.03222.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage, 2020. URL https://arxiv.org/abs/2006.02544.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pp. 722–726, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

Henry Scheffé and John W. Tukey. Non-parametric estimation. i. validation of order statistics. *Annals of Mathematical Statistics*, 16(2):187–192, jun 1945. doi: 10.1214/aoms/1177731119. URL https://doi.org/10.1214/aoms/1177731119.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022. doi: 10.1073/pnas.2111547119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2111547119.

Eleni Straitouri and Manuel Gomez Rodriguez. Designing decision support systems using counterfactual prediction sets, 2024. URL https://arxiv.org/abs/2306.03928.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32633–32653. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/straitouri23a.html.

Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. Controlling counterfactual harm in decision support systems based on prediction sets. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 129443–129479. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e9e3e5bdb017bbc887271c6f6de5353f-Paper-Conference.pdf.

David Stutz, Krishnamurthy, Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers, 2022. URL https://arxiv.org/abs/2110.09192.

Harini Suresh, Natalie Lao, and Ilaria Liccardi. Misplaced trust: Measuring the interference of machine learning in human decision-making. In *12th ACM Conference on Web Science*, WebSci '20, pp. 315–324. ACM, July 2020. doi: 10.1145/3394231.3397922. URL http://dx.doi.org/10.1145/3394231.3397922.

Michael Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8:2293–2303, 2024. doi: 10.1038/s41562-024-02024-1. URL https://doi.org/10.1038/s41562-024-02024-1.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers, 2022. URL https://arxiv.org/abs/2202.03673.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

Zixi Wei, Yuzhou Cao, and Lei Feng. Exploiting human-ai dependence for learning to defer. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Samuel S Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300509. URL https://doi.org/10.1145/3290605.3300509.

Dongping Zhang, Angelos Chatzimparmpas, Negar Kamali, and Jessica Hullman. Evaluating the utility of conformal prediction sets for ai-advised image labeling, 2024. URL https://arxiv.org/abs/2401.08876.

# Appendix

## TABLE OF CONTENTS

# A LITERATURE REVIEW

**Conformal Prediction** The idea of constructing prediction regions can be traced back to classical work on tolerance intervals in statistics (Wilks, 1941; Scheffé & Tukey, 1945). Modern conformal prediction (CP), introduced by Vovk et al. (1999); Saunders et al. (1999); Vovk et al. (2005), builds on this foundation to provide distribution-free, finite-sample validity: given a desired confidence level, CP guarantees that the constructed prediction set contains the true outcome with the prescribed marginal probability.

Over the past two decades, CP has become a standard tool in machine learning for both classification and regression tasks (Papadopoulos et al., 2002; Lei et al., 2017; Romano et al., 2019; 2020), with a large literature on improving *efficiency* (shrinking set size while preserving coverage) (Fisch et al., 2024; Gupta et al., 2022; Kiyani et al., 2024; Stutz et al., 2022; Noorani et al., 2025). A growing body of work extends CP beyond marginal coverage to control more general notions of risk Angelopoulos et al. (2025) introduced conformal risk control, showing how prediction sets can be calibrated to satisfy monotone risk measures rather than coverage alone. Lindemann et al. (2023) apply these principles to safe planning in dynamic environments, demonstrating how conformal methods can enforce operational safety constraints. Lekeufack et al. (2024) developed a conformal decision-theoretic framework where decisions are parameterized by a single scalar and calibrated to control risk. Cortes-Gomez et al. (2025) expand on this view by developing utility-directed conformal prediction, which constructs sets that both retain standard coverage guarantees and minimize downstream decision costs specified by a user-defined utility function.More broadly, Kiyani et al. (2025) show that prediction sets can be viewed as a natural primitive for risk-sensitive decision making: they communicate calibrated uncertainty in a form well-suited for risk-averse decision makers operating in high-stakes domains. This perspective makes conformal prediction sets particularly relevant for human–AI collaboration, where reliable uncertainty estimates are essential for enabling trust and complementarity between human expertise and machine predictions.

**Human-AI collaboration** Human–AI decision-making has attracted growing interest across the machine learning community and social sciences. Yet, realizing true complementarity where the joint system outperforms either the human or the AI-alone, or both, remains challenging. Interestingly, a recent meta analysis by Vaccaro et al. (2024) found out that, on average, human–AI teams underperform the stronger individual agent. These findings underscore persistent difficulties around coordination, trust, and communication between machine and human, motivating the need for algorithmic frameworks that can systematically structure collaboration.

**Learning to Defer**. One approach is the *learning to defer* (L2D) paradigm, where the AI model learns when to predict on its own and when to defer to a human expert. Earlier work Madras et al. (2018) framed this as a mixture-of-experts problem, jointly training a classifier with a deferral mechanism. Wilder et al. (2021) extended this with a decision-theoretic formulation, training models to complement human strengths rather than maximize accuracy alone.

Subsequent work studied the design of surrogate losses for deferral, for example Mozannar & Sontag (2021) showed that standard training objectives can fail to produce optimal deferral policies and proposed a consistent surrogate loss that guarantees Bayes-optimal deferral. Extensions address various settings: Verma & Nalisnick (2022) and Charusaie et al. (2022) studied deferral with multiple experts, while Wei et al. Wei et al. (2024) emphasized that humans and models are not independent and introduced dependent Bayes optimality to exploit correlations between them. Okati et al. (2021) formulated differentiable learning under triage, providing exact optimality guarantees for multi-expert deferral. Most recently Bary et al. (2025) proposed a training-free deferral framework that leverages conformal prediction to allocate decisions among multiple experts. And most recently, along these ideas Arnaiz-Rodriguez et al. (2025) introduced a collaborative matching system that selectively defers to humans to maximize overall performance.

Overall, the L2D literature focuses on *who decides* on each instance: the model or the human. These methods improve team performance by abstention or delegation, which is inherently different than our approach. We start from the human's proposed *set* and ask how to refine it with AI. The goal is to always produce a combined prediction set, particularly one that is simultaneously more reliable and more informative than either agent alone. In this sense, our approach complements deferral-based methods but addresses a different question: not *who decides*, but *how to decide together*.

**Agrement protocols** While more distant from our framework, another line of work views collaboration as an interactive process through *agreement protocols*, where humans and models iteratively exchange feedback until consensus is reached (Aumann, 1976; Collina et al., 2025).

**Prediction Sets for Human-AI decision support** A more related and recent strand of work has explored prediction sets as a structured interface for collaboration and human decision making support. Straitouri et al. (2023) formalized the problem of improving expert predictions with conformal prediction in multiclass classification. In their setting, the AI provides a subset of candidate labels for each instance, from which the human selects, ensuring that the advice is structured but does not override the expert's agency. In parallel, Babbar et al. (2022) empirically evaluated prediction sets in human–AI teams, and showed that set-valued advice can improve human accuracy

compared to single-label predictions. However, they also found that large prediction sets may confuse or slow down human decision-making. To mitigate this, they introduced Deferral-CP (D-CP), where the AI is allowed to abstain entirely on instances for which no sufficiently small set can be produced, deferring the decision back to the human. Other works have studied how to design prediction sets specifically tailored for human use, for example De Toni et al. (2024) proposed a greedy algorithm for constructing prediction sets and showed empirically that it can improve average human accuracy compared to standard conformal sets.

**Counterfactual harm and complementarity** In recent years, there has been growing concern about the unintended consequences of decision support systems using machine learning algorithms in high stakes domains. (Richens et al. (2022); Li et al. (2023); Beckers et al. (2022))To this end, Straitouri et al. (2024) analyze decision-support systems based on prediction sets through the lens of *counterfactual harm* (Feinberg, 1986).Their concern is that requiring humans to always select from a machine-provided set may, in some cases, harm performance: a human who would have been correct unaided might be misled by the system. Using structural causal models, they formally defined and quantified this notion of harm, and under natural monotonicity assumptions, provided methods to estimate or bound how frequently harm may occur without deploying the system. While closely related to our framework, their setting differs from ours in that they study systems where the AI supplies sets from scratch, and their definition of counterfactual harm focuses on the subsequent prediction accuracy of human, whereas we start from sets already provided by the human and ask how to refine them collaboratively, and our definition of counterfactual harm is a direct measure of the quality of the refining procedure.

On the other hand, in the broader human-AI collaboration literature, complementarity is typically defined as whether the combined system achieves higher average accuracy than either the human or the model alone (Yin et al., 2019; Suresh et al., 2020; Lai et al., 2021), and it is currently still unclear how to guarantee this. Our formulation of complementarity is different: it is set-based rather than accuracy-based. Instead of asking whether joint predictions improve overall accuracy, we require that the collaborative prediction set recovers outcomes the human initially missed, while simultaneously avoiding counterfactual harm. This shifts the focus from point-prediction accuracy to the tradeoff between set-based coverage and set size. Crucially, by defining complementarity in this way, our framework provides a principled way to formalize and guarantee it, with clear tradeoffs between the two central metrics of uncertainty quantification.

# B PROOFS

## B.1 PROOF OF THEOREM 2.1 ( OPTIMAL PREDICTION SETS )

*Proof.* The primary optimization problem is given by:

$$\begin{aligned} \min_{C:\mathcal{X}\to 2^{\mathcal{Y}}} \quad & \mathbb{E}\,|C(X)| \\ \text{s.t.} \quad & \mathbb{P}(Y \in C(X) \mid Y \in H(X)) \;\geq\; 1-\varepsilon, \\ & \mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \;\geq\; \delta. \end{aligned} \tag{P}$$

**Part 1** (LP Relaxation)**.** *The original problem involves optimizing over a space of discrete sets, which is a combinatorial and generally NP-hard problem. To make it more tractable, we can formulate an equivalent problem using a continuous relaxation. Let $C(x, y) \in [0, 1]$ be a variable indicating the degree to which $y$ is included in the set for instance $x$. With this relaxation, the objective function, which is the expected size of the set, can be rewritten as $\mathbb{E}\left[\int_{\mathcal{Y}} C(X, y)dy\right]$. Next, we can use the definition of conditional probability, i.e $P(A|B) = \frac{\mathbb{E}[\mathbf{1}_A \mathbf{1}_B]}{\mathbb{E}[\mathbf{1}_B]}$ to rewrite the constraints in terms of expectations, which leads to the following:*

$$\begin{aligned} \min_{C:\mathcal{X}\times\mathcal{Y}\to[0,1]} \quad & \mathbb{E}\left[\int_{\mathcal{Y}} C(X,y)dy\right] \\ \textit{s.t.} \quad & \mathbb{E}[C(X,Y)\mathbf{1}_{Y\in H(X)}] \geq (1-\varepsilon)\mathbb{E}[\mathbf{1}_{Y\in H(X)}]\,, \\ & \mathbb{E}[C(X,Y)\mathbf{1}_{Y\notin H(X)}] \geq \delta\mathbb{E}[\mathbf{1}_{Y\notin H(X)}]\,. \end{aligned} \tag{$\mathcal{P}_{rel}$}$$

*Note that $\mathbb{E}_{X,Y}[\mathbf{1}_{Y\in H(X)}] = P[Y \in H(X)]$. This problem is a linear program, where both the objective functions and the constraints are linear with respect to the decision variable $C$. Since the objective is linear ( and thus convex ) and the feasible region is a convex set, this is a convex optimization problem. Therefore strong duality holds (see Theorem 1 Section 8.3 of Luenberger (1969)).*

**Part 2** (Minimax Formulation)**.** *We formulate the Lagrangian for the relaxed problem by introducing Lagrange multipliers $\lambda_1, \lambda_2 \geq 0$ for the two constraints:*

$$g(\lambda_1, \lambda_2, C) = \mathbb{E}_X \left[ \int_{\mathcal{Y}} C(X, y) dy \right] - \lambda_1 \left( \mathbb{E}_{X,Y}[C(X, Y)\mathbf{1}_{Y \in H(X)}] - (1 - \varepsilon)\mathbb{E}_{X,Y}[\mathbf{1}_{Y \in H(X)}] \right) - \lambda_2 \left( \mathbb{E}_{X,Y}[C(X, Y)\mathbf{1}_{Y \notin H(X)}] - \delta \mathbb{E}_{X,Y}[\mathbf{1}_{Y \notin H(X)}] \right)$$

*The minimax problem is:*

$$\min_C \max_{\lambda_1, \lambda_2 \geq 0} g(\lambda_1, \lambda_2, C)$$

*By strong duality, we can swap the order of the maximization and minimization:*

$$\max_{\lambda_1, \lambda_2 \geq 0} \min_C g(\lambda_1, \lambda_2, C)$$

*To perform the inner minimization over $C$, lets first rewrite the Lagrangian in integral form over the joint probability distribution $p(x, y)$.*

$$g = \int_{\mathcal{X}} \int_{\mathcal{Y}} C(x, y) \left[ p(x) - \lambda_1 \mathbf{1}_{y \in H(x)} p(x, y) - \lambda_2 \mathbf{1}_{y \notin H(x)} p(x, y) \right] dxdy + \text{constant.}$$

*To minimize this integral, we can minimize the integrand for each point $(x, y)$ independently, and thus the inner minimization over $C(x, y) \in [0, 1]$is pointwise. By using the relationship $p(x, y) = p(x)p(y|x)$ and factoring out $p(x)$, the term multiplying $C(x, y)$ becomes:*

$$p(x) \left[ 1 - \lambda_1 \mathbf{1}_{y \in H(x)} p(y|x) - \lambda_2 \mathbf{1}_{y \notin H(x)} p(y|x) \right]$$

*Since $p(x) \geq 0$, the choice of $C(x, y) \in [0, 1]$ that minimizes the expression depends on the sign of the term in brackets. The minimum is attained at the boundaries by setting $C(x, y) = 1$ if the term is negative and $C(x, y) = 0$ if it's positive. This results in an optimal solution $C^*(x, y)$ that is naturally binary-valued*

$$C^*(x, y) = \mathbf{1} \left\{ 1 - \lambda_1 \mathbf{1}_{y \in H(x)} p(y|x) - \lambda_2 \mathbf{1}_{y \notin H(x)} p(y|x) \leq 0 \right\}$$

*and thus the continuous relaxation is tight, as the optimal solution to the relaxed problem is guarantees to be a valid solution for the original problem over the discrete set where $C(x, y) \in \{0, 1\}$.*

**Part 3** (Deriving the Final Form)**.** *The condition for including $y$ in the set $C^*(x)$ can be rewritten as:*

$$1 \leq \lambda_1 \mathbf{1}_{y \in H(x)} p(y|x) + \lambda_2 \mathbf{1}_{y \notin H(x)} p(y|x)$$

*This inequality can be simplified by considering the two mutually exclusive cases for any outcome $y \in \mathcal{Y}$:*

- *If $y \in H(x)$, the condition is $1 \leq \lambda_1 p(y|x) \iff p(y|x) \geq 1/\lambda_1$.*
- *If $y \notin H(x)$, the condition is $1 \leq \lambda_2 p(y|x) \iff p(y|x) \geq 1/\lambda_2$.*

*Combining these two conditions, we can express the optimal set $C^*(x)$ as a single thresholding rule on the conditional probability $p(y|x)$. Let $\lambda_1^*$ and $\lambda_2^*$ be the optimal Lagrange multipliers. We define the optimal thresholds as $b^* = 1 - 1/\lambda_1^*$ and $a^* = 1 - 1/\lambda_2^*$. The condition for including $y$ in the optimal set becomes:*

$$1 - p(y|x) \leq a^* \cdot \mathbf{1}\{y \notin H(x)\} + b^* \cdot \mathbf{1}\{y \in H(x)\}$$

*The optimal set can then be written compactly as a single thresholding rule on the score:*

$$C^*(x) = \left\{ y \in \mathcal{Y} \mid s(y|x) \leq a^* \cdot \mathbf{1}\{y \notin H(x)\} + b^* \cdot \mathbf{1}\{y \in H(x)\} \right\}$$

*Since the optimal solution to the relaxed problem is binary and takes this form, it is also the optimal solution to the original problem.*

This completes the proof. □

### B.2 Proof of Proposition 4.1 ( Offline Algorithm - Coverage Validity )

*Proof.* Given the calibration set $D_{cal} = \{(X_i, Y_i)\}_{i=1}^n$, let $H : \mathcal{X} \to 2^{\mathcal{Y}}$ be the human set map and $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a non-conformity score function. Assume without loss of generality that the (conditional) distribution of the scores is continuous without ties, however in practice this condition is not important as we can always add a vanishing amount of noise to the scores.

For each point define $s_i := S(X_i, Y_i)$ and $h_i = \mathbf{1}\{Y_i \in H(X_i)\}$. Split the indices of the calibration set into two disjoint groups $\mathcal{D}_1 = \{i \leq n : h_i = 1\}$ with size $n_1$ and $\mathcal{D}_2 = \{j \leq n : h_j = 0\}$ with size $n_2$. Given a calibration

set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, let $H : \mathcal{X} \to 2^{\mathcal{Y}}$ denote the human set map, and $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a nonconformity score function. For each example, define the conformity score $s_i := s(X_i, Y_i)$ and let $h_i := \mathbf{1}\{Y_i \in H(X_i)\}$ indicate whether the true label was covered by the human set.

A new test point $(X_{test}, Y_{test})$ is assumed to be exchangeable with the full calibration set. This overall exchangeability of the full set of points implied that, conditioned on the event $h_{test} = 1$, the test point is exchangeable with the set of points in $\mathcal{D}_1$. Similarly, conditioned on the event $h_{test} = 0$, the test point is exchangeable with the set of points in $\mathcal{D}_2$. The prediction set is then defined as

$$C(X_{test}) = \{\, y : \; s(X_{test}, y) \;\le\; \hat{a}\,\mathbf{1}\{y \notin H(X_{test})\} \;+ \hat{b}\,\mathbf{1}\{y \in H(X_{test})\}\,\},$$
$$\text{where} \quad \begin{cases} \hat{a} := \text{Quantile}_{1-\varepsilon}\big(\{s_i : i \in \mathcal{D}_1\} \cup \{\infty\}\big), \\ \hat{b} := \text{Quantile}_{1-\delta}\big(\{s_i : i \in \mathcal{D}_2\} \cup \{\infty\}\big). \end{cases}$$

We now derive a chain of equalities and inequalities for the case of $Y_{test} \in H(X_{test})$:

$$\Pr[Y_{\text{test}} \in C(X_{\text{test}}) \mid h_{\text{test}} = 1] \overset{(a)}{=} \Pr[s_{\text{test}} \le \hat{a} \mid h_{\text{test}} = 1] = \Pr\Big[s_{\text{test}} \le \text{Quantile}_{1-\varepsilon}\Big(\{s_i : i \in \mathcal{D}_{\text{in}}\} \cup \{\infty\}\Big) \;\Big|\; h_{\text{test}} = 1\Big]$$
$$\overset{(b)}{=} \mathbb{E}\left[\frac{1}{n_1+1} \sum_{i \in \mathcal{D}_1 \cup \{\text{test}\}} \mathbf{1}\Big\{s_i \le \text{Quantile}_{1-\varepsilon}\Big(\{s_j : j \in \mathcal{D}_1\} \cup \{s_{\text{test}}\}\Big)\Big\} \;\middle|\; h_{\text{test}} = 1\right] \overset{(c)}{\ge} 1 - \varepsilon.$$

and analogously the case if $Y_{test} \notin H(X_{test})$:

$$\Pr[Y_{\text{test}} \in C(X_{\text{test}}) \mid h_{\text{test}} = 0] \overset{(a)}{=} \Pr\Big[s_{\text{test}} \le \hat{b} \;\Big|\; h_{\text{test}} = 0\Big] = \Pr\Big[s_{\text{test}} \le \text{Quantile}_{1-\delta}\Big(\{s_i : i \in \mathcal{D}_2\} \cup \{\infty\}\Big) \;\Big|\; h_{\text{test}} = 0\Big]$$
$$\overset{(b)}{=} \mathbb{E}\left[\frac{1}{n_2+1} \sum_{i \in \mathcal{D}_2 \cup \{\text{test}\}} \mathbf{1}\Big\{s_i \le \text{Quantile}_{1-\delta}\Big(\{s_j : j \in \mathcal{D}_2\} \cup \{s_{\text{test}}\}\Big)\Big\} \;\middle|\; h_{\text{test}} = 0\right] \overset{(c)}{\ge} 1 - \delta.$$

**where**

(a) By definition of $C(\cdot)$ and the thresholds $\hat{a}$ and $\hat{b}$ when: $h_{\text{test}} = 1$ (resp. 0), inclusion is the event $s_{\text{test}} \le \hat{a}$ (resp. $s_{\text{test}} \le \hat{b}$).

(b) By exchangeability within the corresponding group: conditional on $h_{\text{test}}$, the set of scores $\{s_i : i \in \mathcal{D}_1\} \cup \{s_{\text{test}}\}$ (or $\mathcal{D}_2$) is exchangeable, so we average the indicator over the $n_{\text{group}} + 1$ equally likely ranks.

(c) By the definition of the empirical quantile: at least a $1 - \varepsilon$ (resp. $\delta$) fraction of the $n_{\text{group}} + 1$ values are $\le$ that quantile.

Therefor thus far we have established that

$$\Pr[Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} \in H(X_{\text{test}})] \;\ge\; 1 - \varepsilon, \qquad \Pr[Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} \notin H(X_{\text{test}})] \;\ge\; 1 - \delta.$$

And now for the upper bounds: for the case $Y_{test} \in H(X_{test})$:

$$\Pr[Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} \in H(X_{\text{test}}] \overset{(d)}{=} \frac{\lceil (1-\varepsilon)(n_1+1) \rceil}{n_1+1} \overset{(e)}{<} 1 - \varepsilon + \frac{1}{n_1+1}.$$

Similarly, for the case $Y_{test} \notin H(X_{test})$, we have:

$$\Pr[Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} \notin H(X_{\text{test}}] = \frac{\lceil (1-\delta)(n_2+1) \rceil}{n_2+1} < (1-\delta) + \frac{1}{n_2+1}.$$

**where**

(d) Given that the groupwise score distributions are continuous, the rank of $s_{\text{test}}$ among the scores in the corresponding group ($\mathcal{D}_1 \cup \{s_{\text{test}}\}$) is uniformly distributed. This makes the probability equal to the exact proportion of scores less than or equal to the quantile.

(e) By the property of the ceiling function $\lceil x \rceil < x + 1$

□

### B.3 Proof of Proposition 4.2 ( Online Algorithm - Gaurantees )

First we present the following lemma that states that the thresholds $a_t$ and $b_t$ remain bounded at all time steps:

**Lemma B.1** (Parameter Boundedness). *Let $s(x, y) \in [0, 1]$ be the non conformity scores. For any learning rate $\eta > 0$, the sequences $\{a_t\}$ and $\{b_t\}$ are bounded. Specifically, for all $t > 1$:*

$$b_t \in [-\eta\varepsilon,\ 1 + \eta(1 - \varepsilon)], \qquad a_t \in [-\eta\delta,\ 1 + \eta(1 - \delta)].$$

*Proof.* We prove the result for $b_t$; the proof for $a_t$ is symmetric.

Let $I_b = [-\eta\varepsilon, 1 + \eta(1 - \varepsilon)]$. We show by induction that once $b_t$ enters $I_b$, it never leaves.

First, since $b_1 \in [0, 1]$, the first update ensures $b_2 \in I_b$. Now, assume $b_t \in I_b$ for some $t > 1$.

**Upper Bound**: $b_{t+1}$ is maximized if the update is positive, which requires $\mathbf{1}\{s_t > b_t\} = 1$. This implies $s_t > b_t$, so $b_t$ must be less than $s_t \leq 1$. The update is $b_{t+1} = b_t + \eta(1-\varepsilon)$. Since this increase only happens when $b_t < 1$, we have $b_{t+1} < 1 + \eta(1 - \varepsilon)$. If the update is negative, $b_{t+1} < b_t$, so it is also below the upper bound. Thus, $b_{t+1} \leq 1 + \eta(1 - \varepsilon)$.

**Lower Bound**: $b_{t+1}$ is minimized if the update is negative, which requires $\mathbf{1}\{s_t > b_t\} = 0$. This implies $s_t \leq b_t$, so $b_t$ must be greater than $s_t \geq 0$. The update is $b_{t+1} = b_t - \eta\varepsilon$. Since this decrease only happens when $b_t \geq 0$, we have $b_{t+1} \geq 0 - \eta\varepsilon = -\eta\varepsilon$. If the update is positive, $b_{t+1} > b_t$, so it is also above the lower bound. Thus, $b_{t+1} \geq -\eta\varepsilon$.

We have shown by induction that the parameters remain in their respective intervals for all $t > 1$. □

Now first, lets restate the proposition

**Proposition B.2** (Finite-Sample Guarantees). *Let $N_1(T) = \sum_{t=1}^{T} \mathbf{1}\{Y_t \in H(X_t)\}$ and $N_2(T) = \sum_{t=1}^{T} \mathbf{1}\{Y_t \notin H(X_t)\}$. For any $T \geq 1$:*

$$\left| \frac{1}{N_1(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} - \varepsilon \right| \leq \frac{1 + \eta \max(\varepsilon, 1 - \varepsilon)}{\eta N_1(T)}, \quad \left| \frac{1}{N_2(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{out}} - \delta) \right| \leq \frac{1 + \eta \max(\delta, 1 - \delta)}{\eta N_2(T)}.$$

We prove the first bound, and the second statement is symmetric. Let $I_1(T) = \{t \leq T \mid Y_t \in H(X_t)\}$ be the set of indices where the true label lies within the human proposed set. The number of such examples is $N_1(T) = |I_1(T)|$. As per algorithm, we only update the threshold $b_t$ for such points, and the update is given by:

$$b_{t+1} - b_t = \eta(\mathrm{err}_t^{\mathrm{in}} - \varepsilon)$$

where $err_{CH,t} = \mathbf{1}\{Y_t \notin C(X_t) \mid Y_t \in H(X_t)\}$ Note that this update only occurs to times $t \in I_{in}(T)$. If you sum over all relevant time steps where the update occurs we form a telescoping sum:

$$b_{T+1} - b_0 = \sum_{t \in I_1(T)} \eta(\mathrm{err}_t^{\mathrm{in}} - \varepsilon) = \eta \left( \sum_{t \in I_1(T)} \mathrm{err}_t^{\mathrm{in}} - \sum_{t \in I_1(T)} \varepsilon \right)$$

Since $err_t^1 = 0$ for all $t \notin I_1(T)$, we can expand the sum over all time steps and rearrange to get:

$$\frac{b_{T+1} - b_1}{\eta} = \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} - \varepsilon N_1(T)$$

and rearranging again and taking the absolute value we obtain:

$$\left| \frac{1}{N_1(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} - \varepsilon \right| = \left| \frac{b_{T+1} - b_1}{\eta N_1(T)} \right|$$

Using Lemma B.1, we can bound the numerator. The maximum value of $b_t$ is $1 + \eta(1 - \varepsilon)$ and the minimum is $-\eta\varepsilon$, which gives the bound $|b_{T+1} - b_1| \leq 1 + \eta \max(1 - \varepsilon, \varepsilon)$. Substituting this directly we obtain our finite-sample inequality

$$\left| \frac{1}{N_1(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} - \varepsilon \right| \leq \frac{1 + \eta \max(1 - \varepsilon, \varepsilon)}{\eta N_1(T)}$$

## C Additional Experimental Results

This section presents supplementary results for each of the data modalities studied in the main paper. These results were omitted due to space constraints but are fully consistent with the findings we discussed earlier. All additional experiments are complementary and serve to reinforce the core claims of the paper.

### C.1 CLASSIFICATION: IMAGENET-16H

Beyond the noise and human strategy shifts discussed in the main text, we also study a class label shift on ImageNet-16H. In this setting, calibration is performed on a restricted subset of classes, while evaluation takes place on a disjoint set of unseen classes. For instance, calibration may use only dog and cat images, while the test stream consists of bird images. This creates a particularly challenging shift, since the labels encountered at test time are entirely absent from calibration. We report the same metrics as in the main experiments. Figure 4 compares our adaptive online algorithm with the fixed baseline in terms of running coverage. As before, the adaptive method tracks the target levels closely, while the fixed baseline drifts and does not recover. Figure 5 then compares CUP against the Human-only and AI-only baselines. The pattern is consistent with other shifts: Relative to the human baseline, CUP raises coverage while also pruning incorrect labels from overly conservative sets, resulting in sharper and more informative predictions. Relative to AI alone, CUP achieves the same level of coverage with smaller sets, showing human input is being efficiently incorporated in the resulting prediction sets.

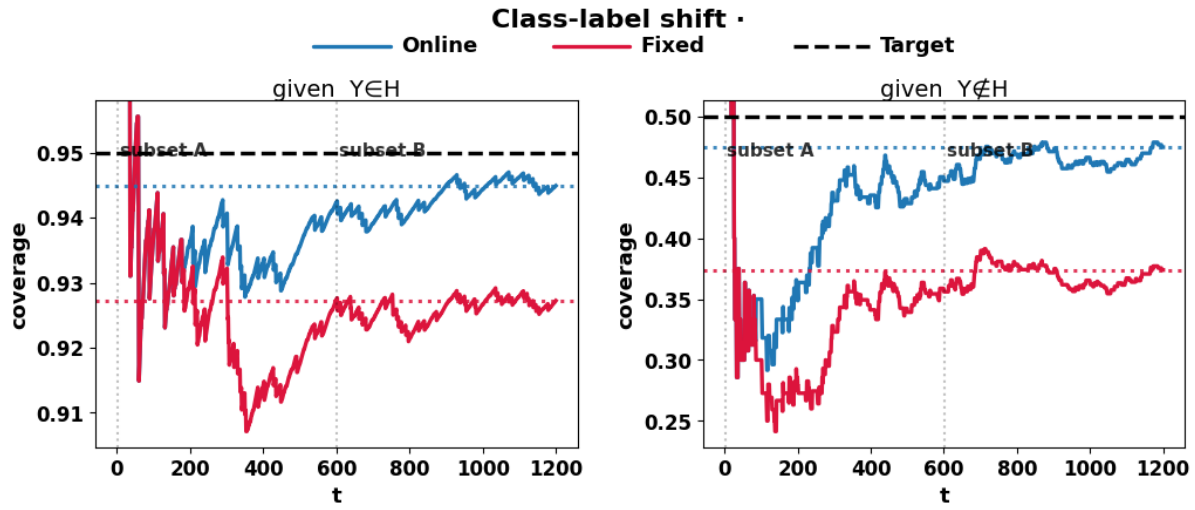



Figure 4: **Fixed vs. online CUP on ImageNet-16H under class label shift.** The online algorithm remains close to target coverage, while the fixed baseline drifts and fails to recover.

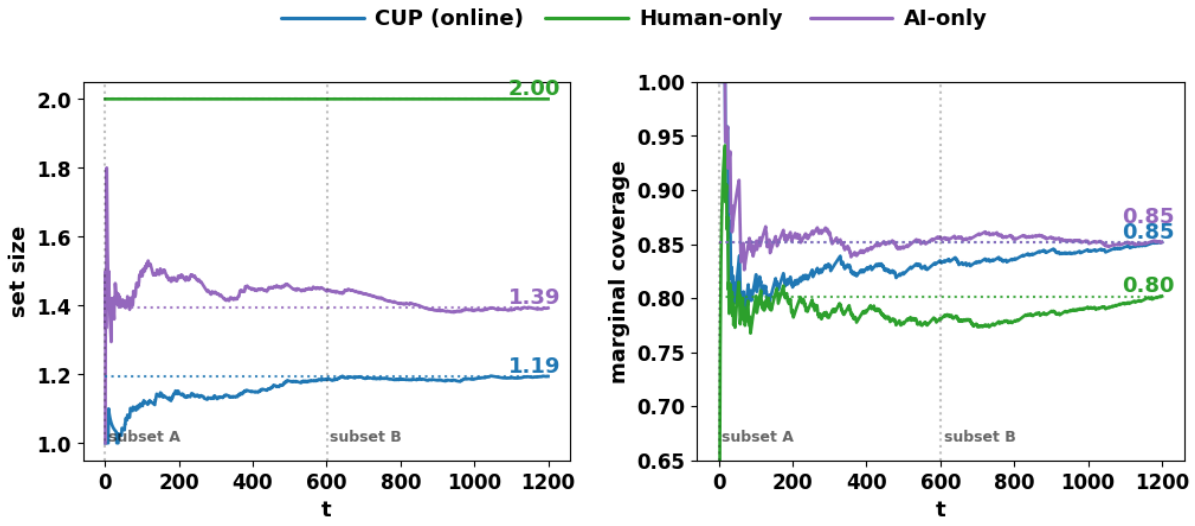



Figure 5: **Comparison against baselines on ImageNet-16H under class label shift.** CUP outperforms both Human-only and AI-only baselines, achieving higher coverage with smaller prediction sets.

### C.2 LLMs FOR MEDICAL DIAGNOSIS DECISION MAKING

We begin with additional offline results on the DDXPlus dataset, exploring a wider range of calibration parameters $(\varepsilon, \delta)$. Table 4 reports coverage and set size for Human-only, AI-only, and CUP (ours) across both GPT-4o and GPT-5. These extra configurations make it possible to see how tuning the targets affects performance.

Across all reported settings, CUP improves upon the human baseline in at least one dimension, coverage or set size, with the magnitude of the gain depending on the specific $(\varepsilon, \delta)$ configuration. With the stronger model (GPT-5), CUP is often able to achieve simultaneous improvements in both dimensions: pruning away incorrect human labels while adding the correct one when needed, leading to higher coverage and smaller sets. With the weaker model (GPT-4o), coverage improvements are still observed, but they often come with larger set sizes, reflecting the model's more limited ability to prune. In all cases, however, CUP achieves smaller sets than AI alone at matched coverage levels, confirming that human input is effectively incorporated.

| | Human | GPT-4o | | | GPT-5-mini | | |
|---|---|---|---|---|---|---|---|
| **Strategy** | **C/S** | **CUP C/S** | **$(\varepsilon, \delta)$** | **AI C/S** | **CUP C/S** | **$(\varepsilon, \delta)$** | **AI C/S** |
| Top-1 | 0.71 / 1.00 | 0.89 / 2.56 | (0.01, 0.65) | 0.88 / 4.58 | 0.87 / 1.27 | (0.02, 0.55) | 0.88 / 1.54 |
| Top-1 | 0.71 / 1.00 | 0.88 / 2.51 | (0.02, 0.65) | 0.88 / 4.40 | 0.88 / 1.36 | (0.01, 0.55) | 0.89 / 1.59 |
| Top-1 | 0.71 / 1.00 | 0.85 / 1.77 | (0.01, 0.50) | 0.85 / 3.69 | 0.85 / 1.19 | (0.02, 0.50) | 0.86 / 1.42 |
| Top-1 | 0.71 / 1.00 | 0.90 / 2.84 | (0.02, 0.70) | 0.88 / 4.64 | 0.91 / 1.59 | (0.02, 0.70) | 0.91 / 1.76 |
| Top-2 | 0.87 / 1.95 | 0.90 / 2.47 | (0.01, 0.30) | 0.88 / 4.57 | 0.95 / 2.31 | (0.01, 0.70) | 0.95 / 2.79 |
| Top-2 | 0.87 / 1.95 | 0.93 / 3.14 | (0.01, 0.45) | 0.90 / 9.12 | 0.94 / 1.73 | (0.02, 0.55) | 0.94 / 2.10 |
| Top-2 | 0.87 / 1.95 | 0.94 / 3.69 | (0.01, 0.55) | 0.93 / 22.41 | 0.91 / 1.51 | (0.02, 0.40) | 0.91 / 1.83 |
| Top-2 | 0.87 / 1.95 | 0.93 / 3.41 | (0.01, 0.50) | 0.91 / 13.55 | 0.93 / 1.65 | (0.02, 0.45) | 0.93 / 1.95 |

Table 4: **Additional configurations for offline setting: results on DDXPlus.** Human sets (shared across models) use a top-$k$ strategy. We compare the Human alone against CUP (ours) and AI Alone for two models side-by-side. Entries report *Coverage/Size* (C/S) with calibration parameters $(\varepsilon, \delta)$ shown for CUP.

We also report online results for DDXPlus under the same age-based distribution shift described in the main paper. Calibration is performed on younger patients, while testing proceeds on streams of older patients. These results were omitted earlier due to space but are included here for completeness.

Figure 6 shows the running coverage of CUP (online) compared to a fixed, non-adaptive variant. As in other modalities, the adaptive updates keep CUP close to the target levels throughout the stream, while the fixed baseline drifts and fails to recover.

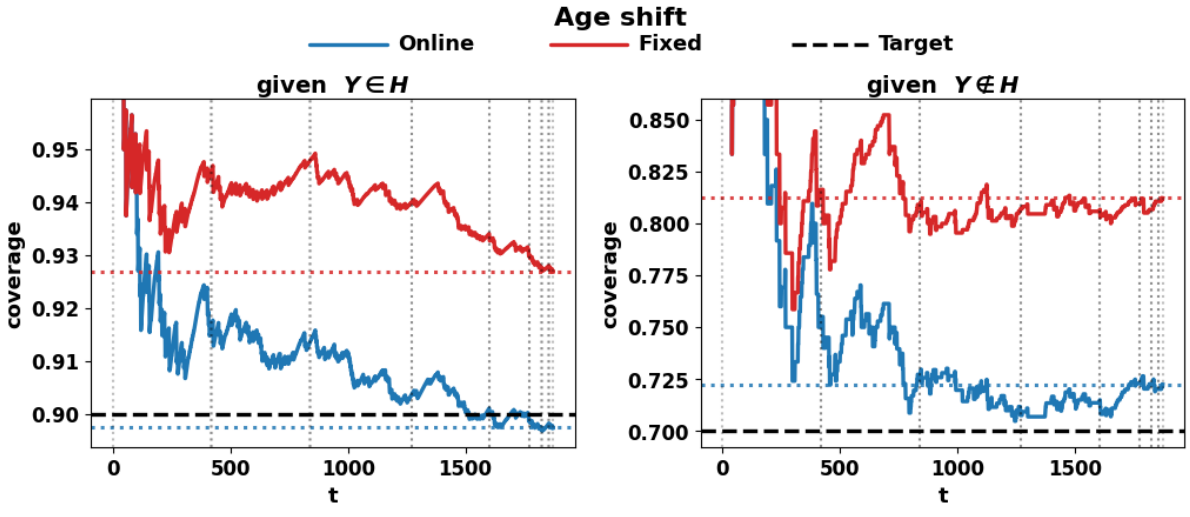



Figure 6: **Fixed vs. online CUP on DDXPlus under an age-based shift.** Same setting as in the main paper, included here for completeness. The online algorithm remains close to target coverage, while the fixed baseline drifts and fails to recover.

### C.3 Regression: Communities & Crime

As in the main paper for ImageNet, and in the appendix for LLMs, we evaluate CUP-online we against a fixed, non-adaptive variant in the regression setting with the UCI Communities & Crime dataset (Redmond, 2002) in Figure 7

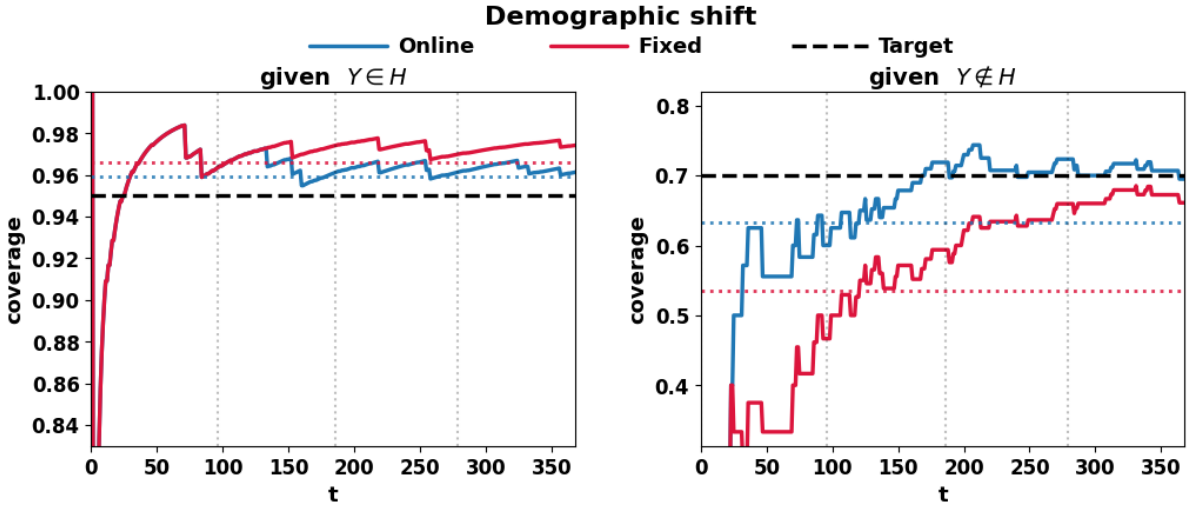



Figure 7: Fixed vs online CUP on Crime & Communitites dataset