Uncertainty Quantification for Generative and Collaborative AI

Sima Noorani

University of Pennsylvania

Nov 20th, 2025

Joint work with: Shayan Kiyani, George Pappas, Hamed Hassani

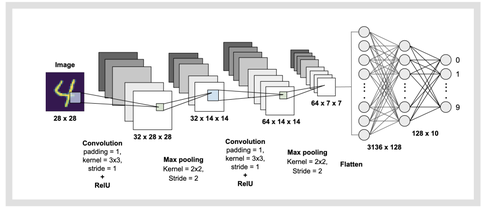


Penn
University of Pennsylvania

AI-powered Decision making pipeline

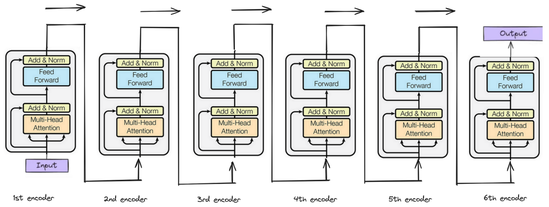
AI-powered Decision making pipeline

Inputs/prompts → AI model → Predictions/Recommendations → Decision Maker → Actions



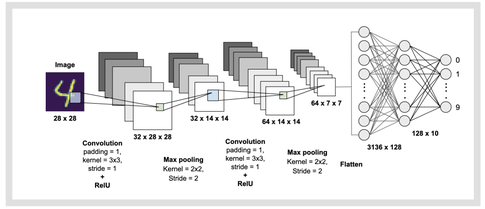
- (Heavily) Pre-trained

- Possibility of little tweaks using extra data (fine-tuning)



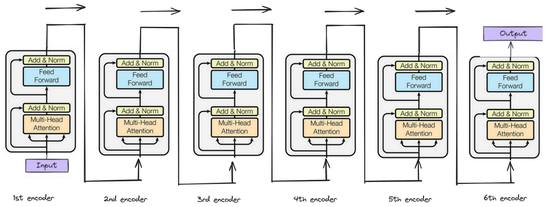
AI-powered Decision making pipeline

Inputs/prompts → AI model → Predictions/Recommendations → Decision Maker → Actions



- (Heavily) Pre-trained

- Possibility of little tweaks using extra data (fine-tuning)

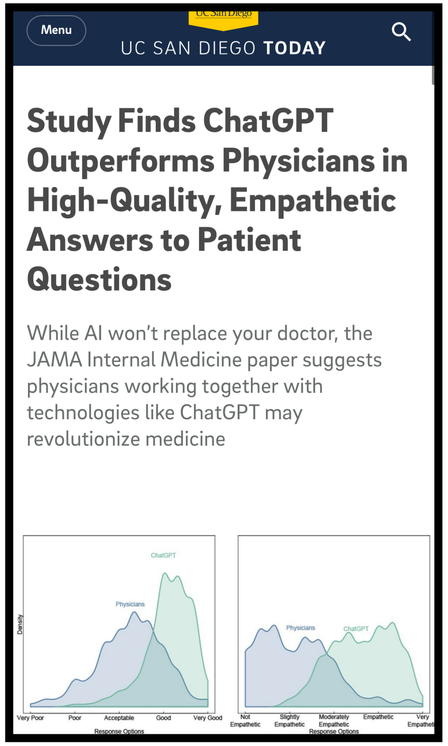


- Automated

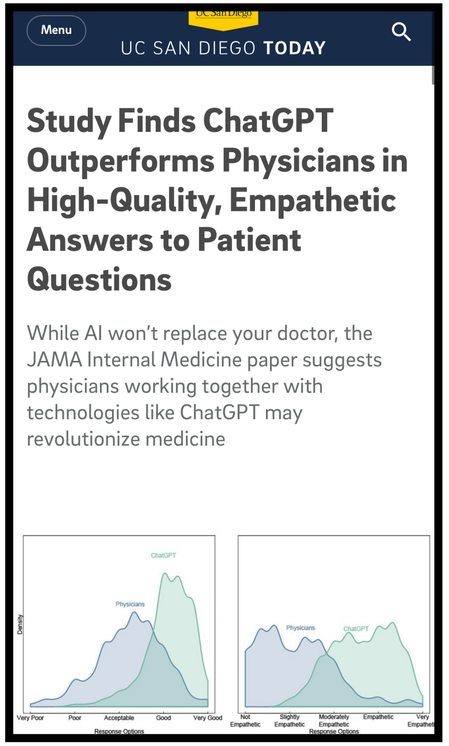
- Humans

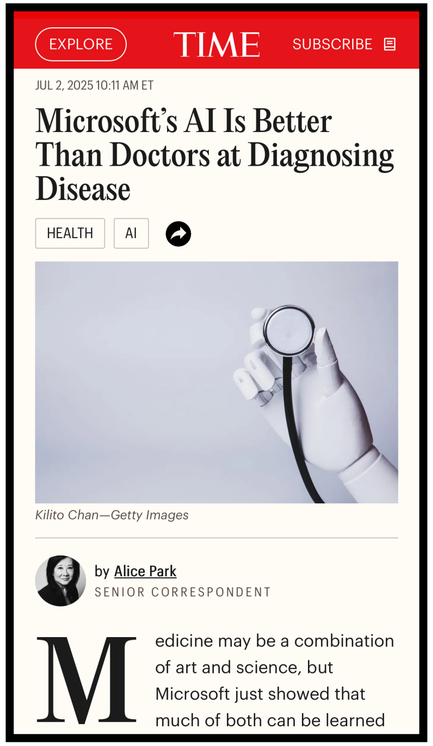
LLM Agents

AI-powered Decision making pipeline



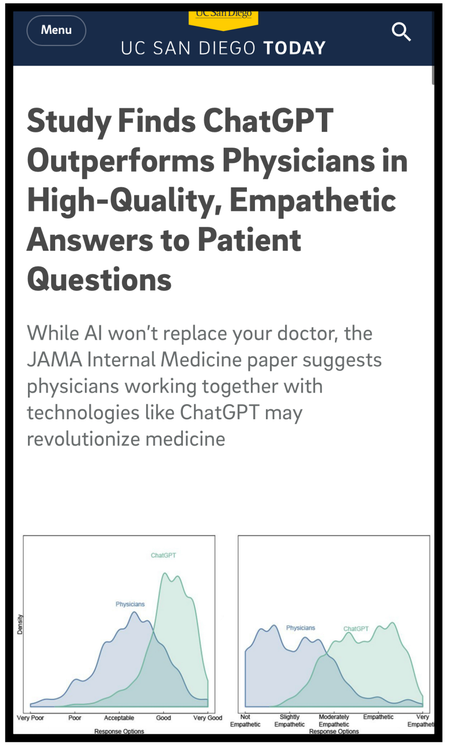
AI-powered Decision making pipeline

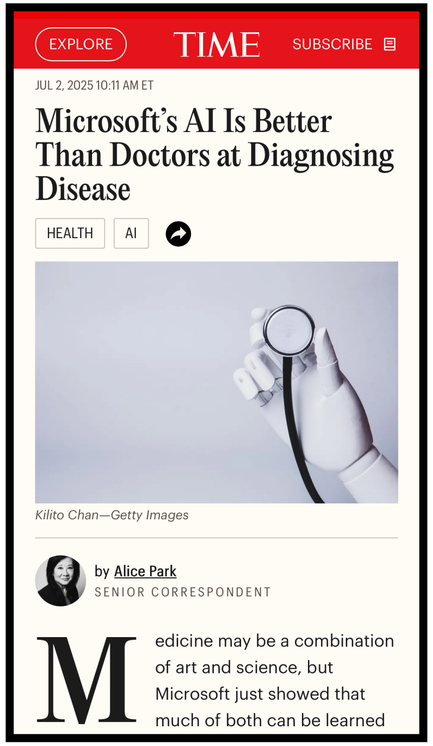


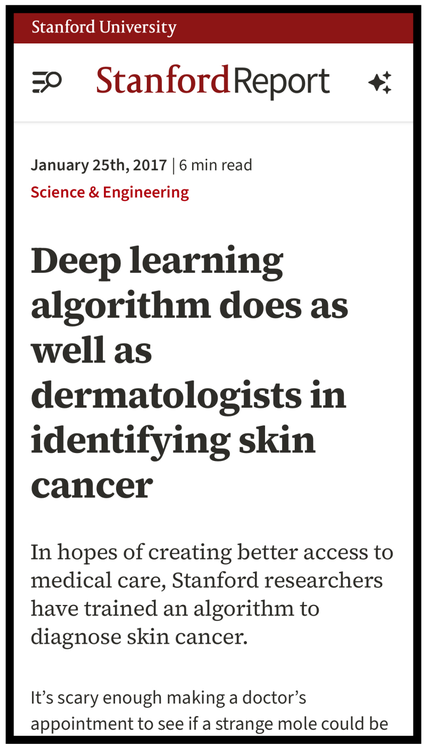


AI-powered Decision making pipeline

Inputs/prompts → AI model → Predictions/Recommendations → Decision Maker → Actions







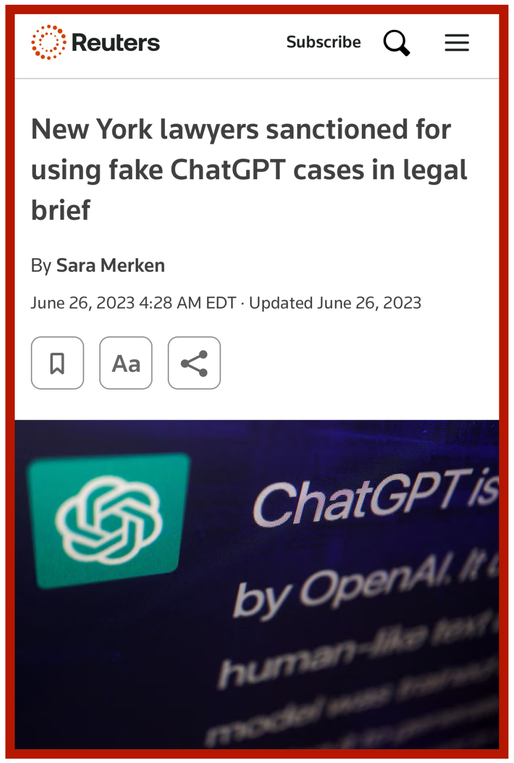
AI-powered Decision making pipeline

Menu
UC SAN DIEGO TODAY
Study Finds ChatGPT Outperforms Physicians in High-Quality, Empathetic Answers to Patient Questions
While AI won't replace your doctor, the JAMA Internal Medicine paper suggests physicians working together with technologies like ChatGPT may revolutionize medicine

EXPLORE
TIME
SUBSCRIBE
JUL 2, 2025 10:11 AM ET
Microsoft's AI Is Better Than Doctors at Diagnosing Disease
HEALTH
AI
Kilito Chan—Getty Images
by Alice Park
SENIOR CORRESPONDENT
Medicine may be a combination of art and science, but Microsoft just showed that much of both can be learned

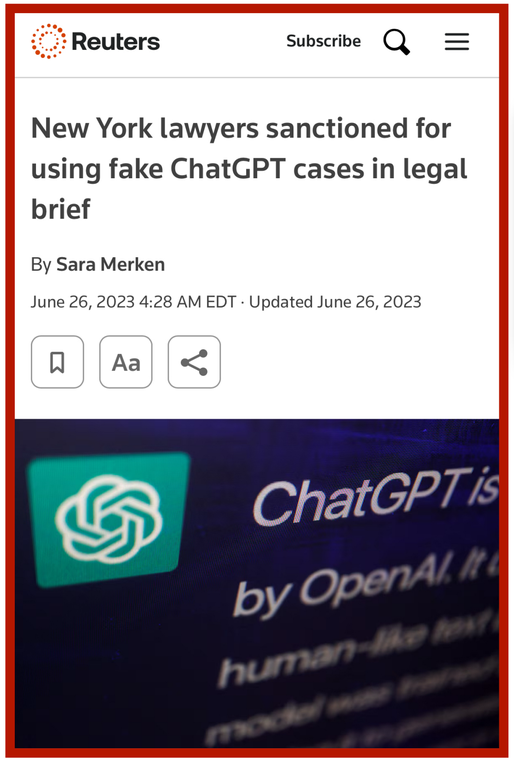
Stanford University
StanfordReport
January 25th, 2017 | 6 min read
Science & Engineering
Deep learning algorithm does as well as dermatologists in identifying skin cancer
In hopes of creating better access to medical care, Stanford researchers have trained an algorithm to diagnose skin cancer.
It's scary enough making a doctor's appointment to see if a strange mole could be

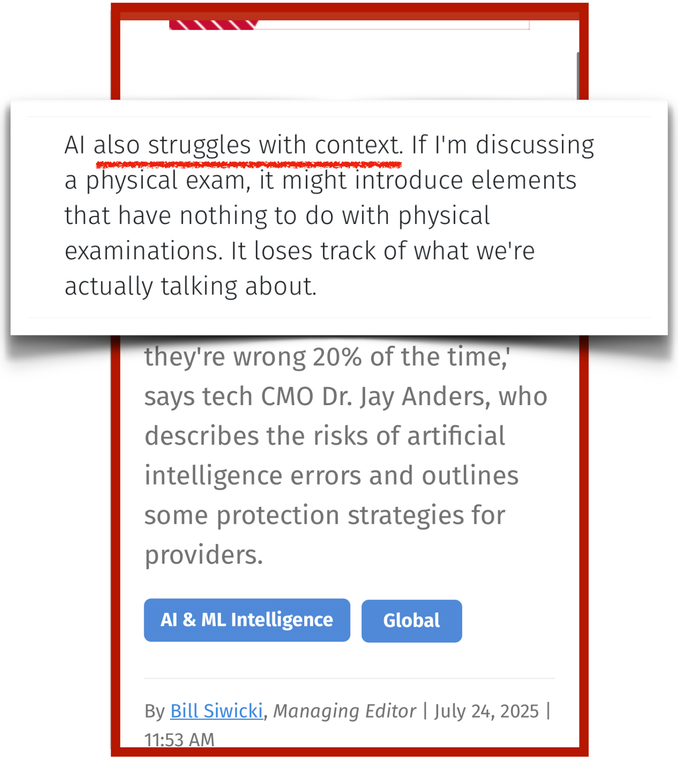
AI-powered Decision making pipeline



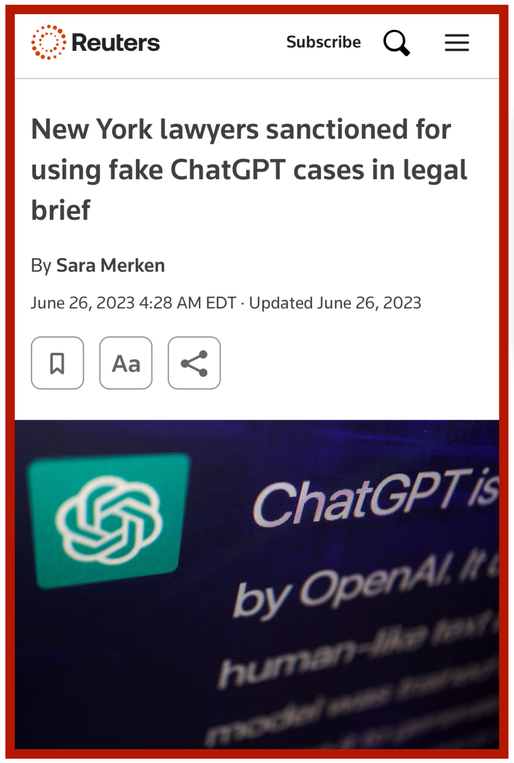
AI-powered Decision making pipeline

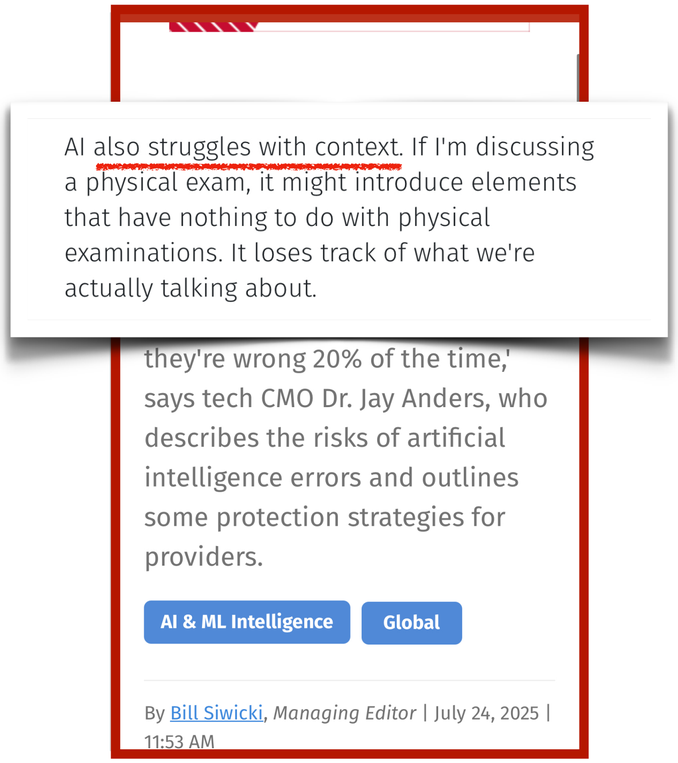
Inputs/prompts → AI model → Predictions/Recommendations → Decision Maker → Actions

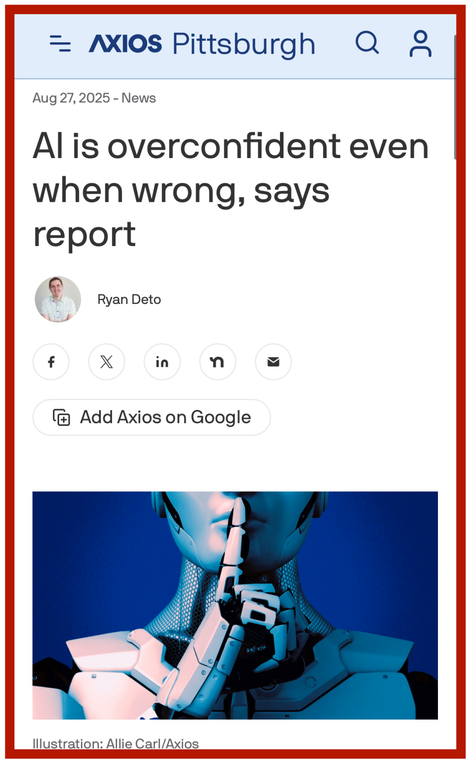




AI-powered Decision making pipeline

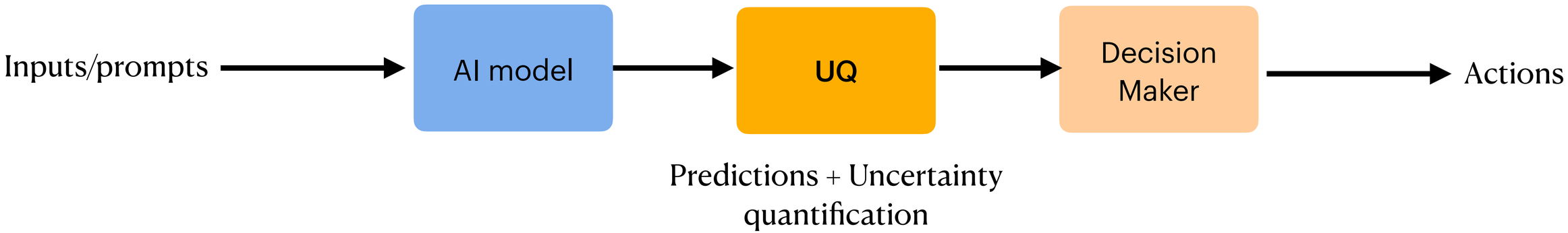




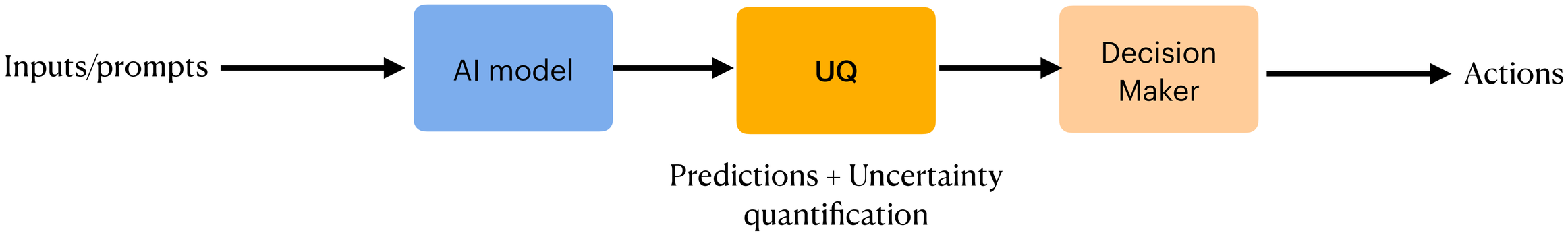


AI-powered Decision making pipeline

AI-powered Decision making pipeline

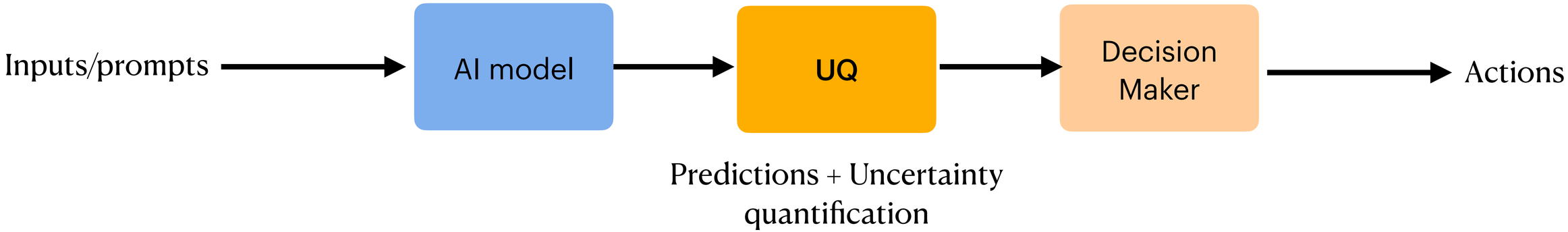


AI-powered Decision making pipeline



Robust Decision making requires precise Uncertainty Quantification

AI-powered Decision making pipeline



Part I: CP for Generative Models

Conformal Prediction Beyond the Seen: A Missing Mass Perspective for Uncertainty Quantification in Generative Models

Sima Noorani[*1], Shayan Kiyani[*1], George Pappas[1], and Hamed Hassani[1]

[1]University of Pennsylvania

Part II: Collaborative CP

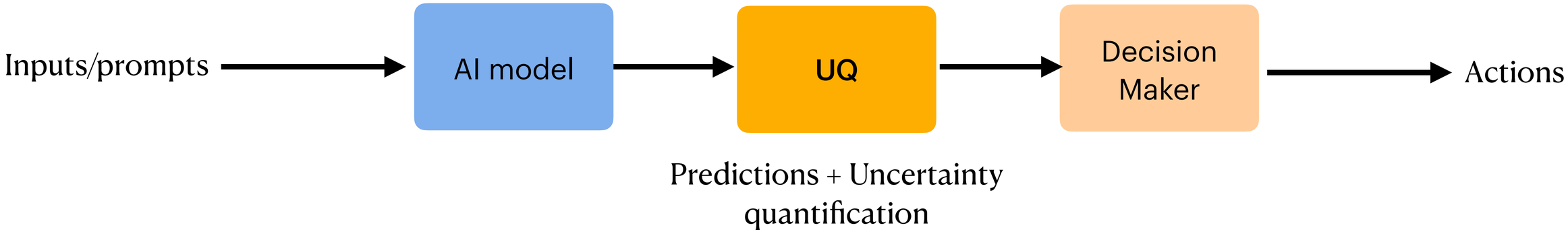
Human AI Collaborative Uncertainty Quantification

Sima Noorani[*1], Shayan Kiyani[*1], George Pappas[1], and Hamed Hassani[1]

[1]University of Pennsylvania

Arxiv - Oct '25

AI-powered Decision making pipeline



Part I: CP for Generative Models

Conformal Prediction Beyond the Seen: A Missing Mass Perspective for Uncertainty Quantification in Generative Models

Sima Noorani[*1], Shayan Kiyani[*1], George Pappas[1], and Hamed Hassani[1]

[1]University of Pennsylvania

Part II: Collaborative CP

Human AI Collaborative Uncertainty Quantification

Sima Noorani[*1], Shayan Kiyani[*1], George Pappas[1], and Hamed Hassani[1]

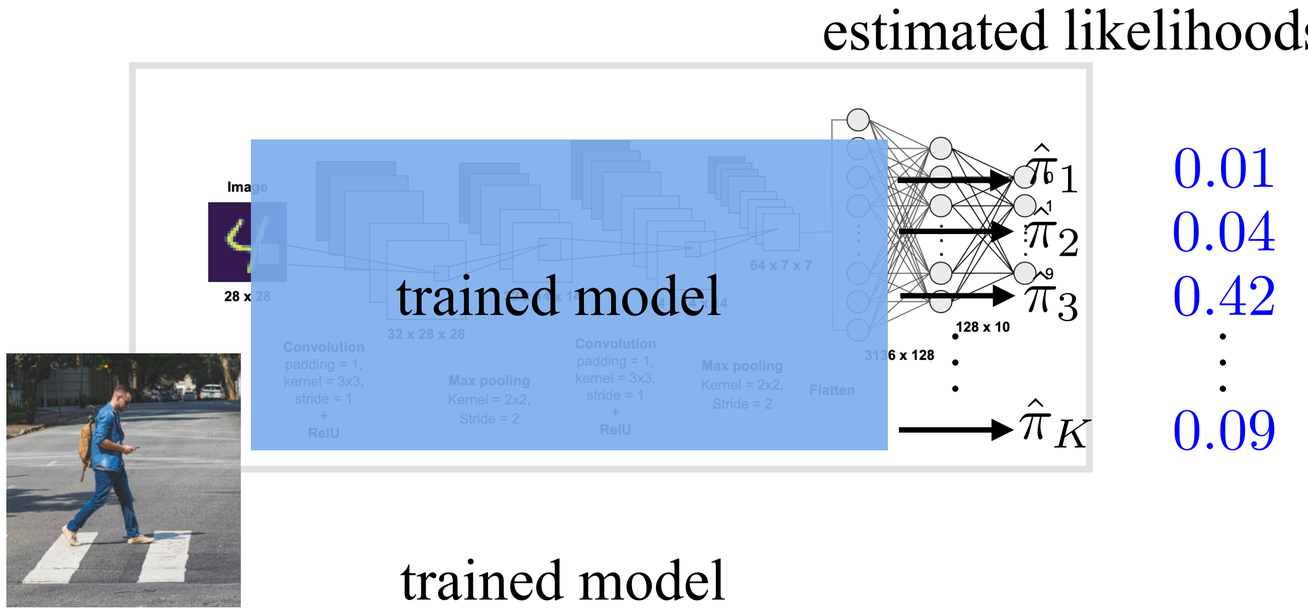
[1]University of Pennsylvania

Arxiv - Oct '25

How to do uncertainty quantification in the generative setting (LLMs) ?

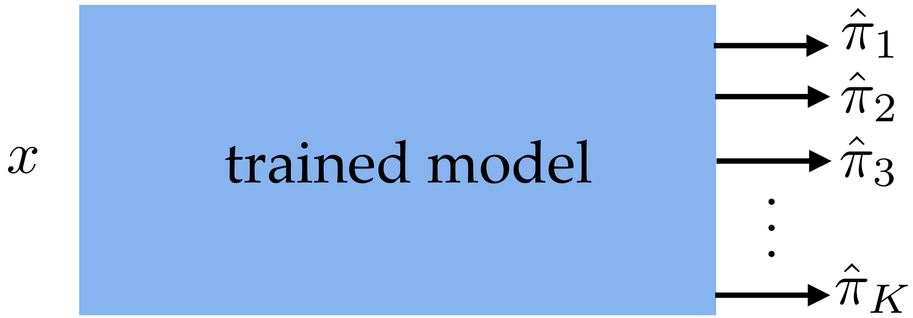
(Quick) Review of Conformal Prediction

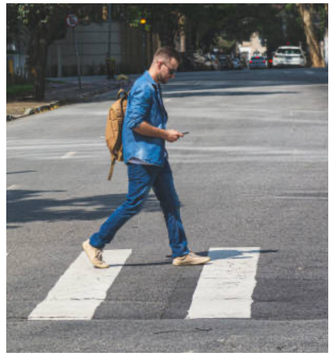
Conformal Prediction



- prediction is then based on the class that has maximum likelihood

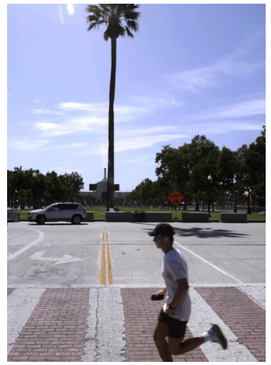
Conformal Prediction





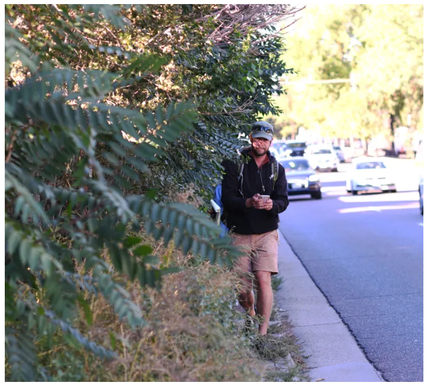
{ human 0.99 }

easy



{ human, tree, bin
0.90 0.08 0.02 }

less easy



{ Tree, human, trash can
0.70 0.20 0.05 }

difficult

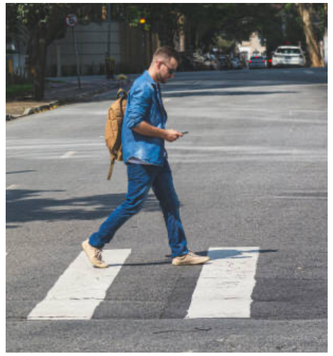
model's prediction: human ✅ human ✅ tree ❌

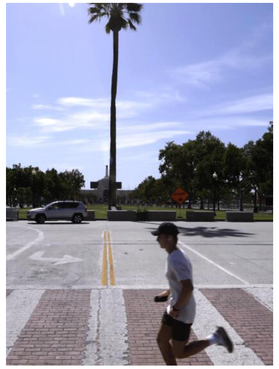
➔ the trained model provides estimated likelihoods (a notion of uncertainty)

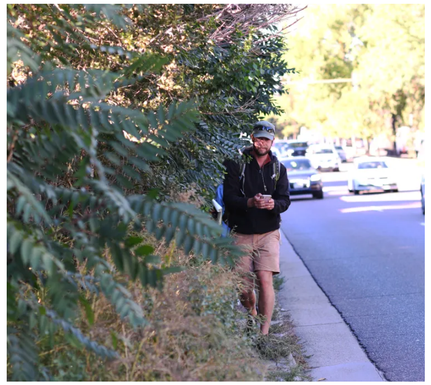
➔ these likelihoods are informative but not always correct

Conformal Prediction

x







{ human 0.99 }

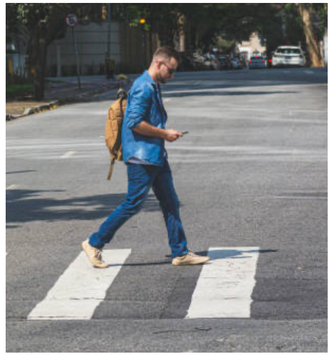
{ human 0.90, tree 0.08, bin 0.02 }

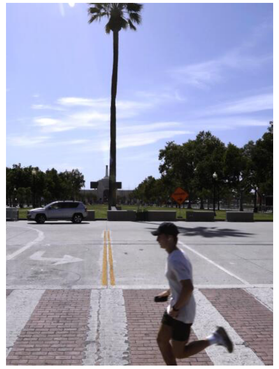
{ Tree 0.70, human 0.20, trash can 0.05 }

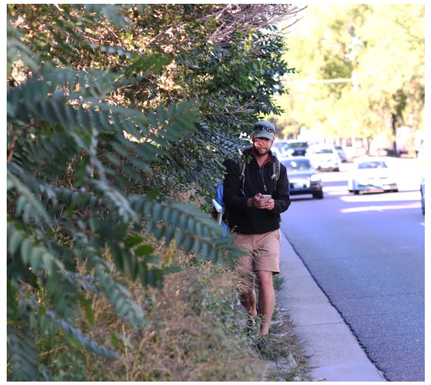
model's prediction: human ✅ human ✅ tree ❌

Conformal Prediction

x







{ human 0.99 } { human , tree , bin 0.90 0.08 0.02 } { Tree , human , trash can 0.70 0.20 0.05 }

$C(x)$

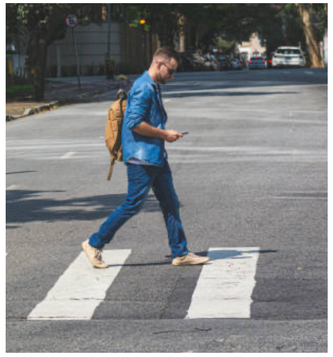
(model's conformal prediction)

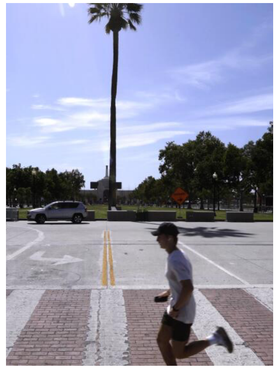
{ human } { human } { tree , human }

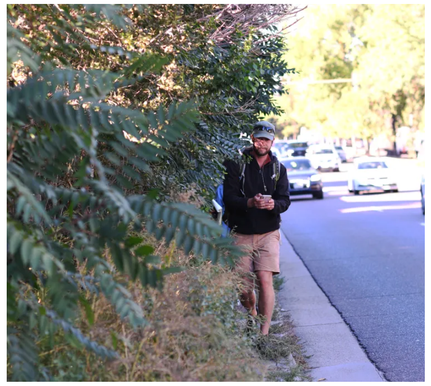
✅ ✅ ✅

Single prediction → set of predictions

x







{ human 0.99 } { human 0.90, tree 0.08, bin 0.02 } { Tree 0.70, human 0.20, trash can 0.05 }

$C(x)$

(model's conformal prediction)

{ human } ✅ { human } ✅ { tree, human } ✅

Marginal Coverage

$$\Pr\{Y_{\text{test}} \in C(X_{\text{test}})\} \geq 1 - \alpha$$

User-specified value; e.g. $1 - \alpha = 0.9$

Conformal Prediction

Marginal Coverage

$$\Pr\{Y_{\text{test}} \in C(X_{\text{test}})\} \geq 1 - \alpha$$

Classification

x

AI model

CP

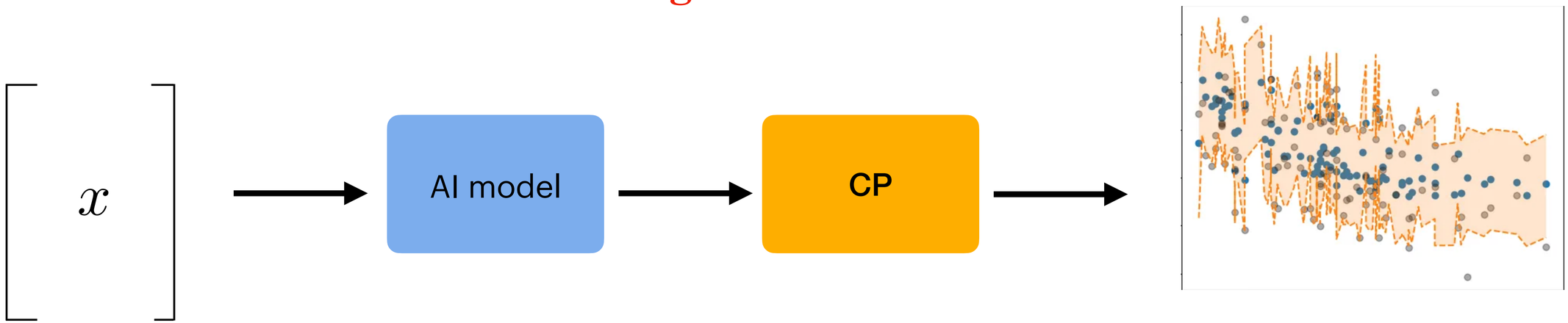
{Tree, human, trash can}
0.70 0.20 0.05

Conformal Prediction

Marginal Coverage

$$\Pr\{Y_{\text{test}} \in C(X_{\text{test}})\} \geq 1 - \alpha$$

Regression



Conformal Prediction

Marginal Coverage

$$\Pr\{Y_{\text{test}} \in C(X_{\text{test}})\} \geq 1 - \alpha$$

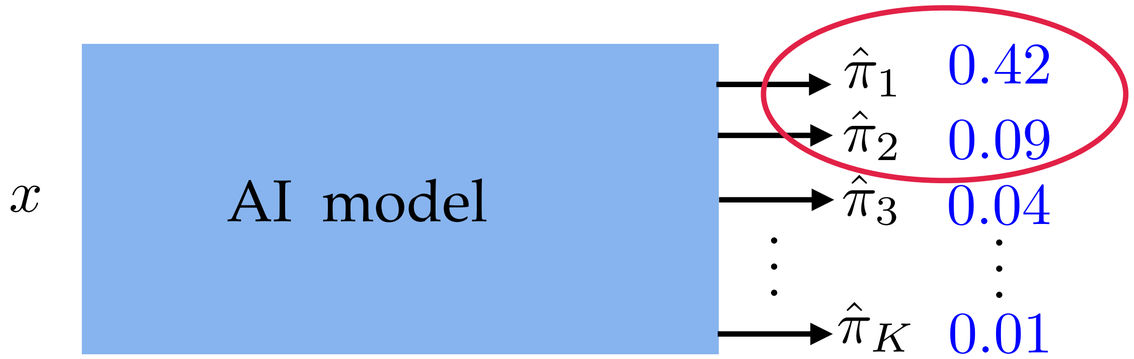
the threshold

$$C(x) = \{y \in \mathcal{Y} : \quad S(x, y) \leq q\}$$

difficulty of the input ↑ models' uncertainty about the label ↑ size of the prediction set ↑

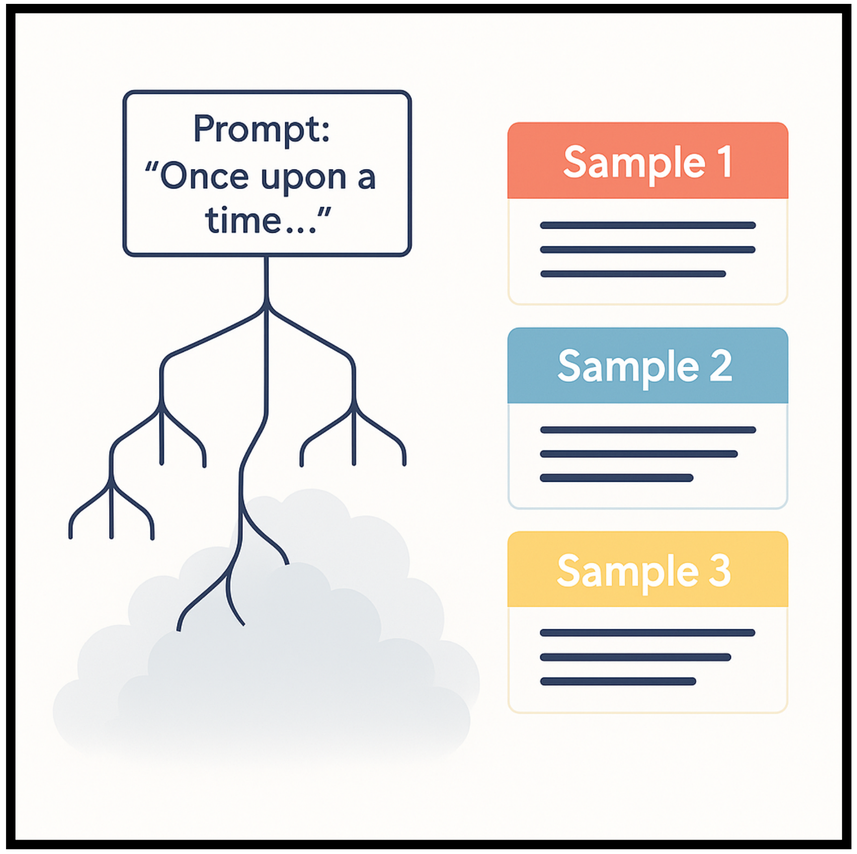
The Challenges of Extending CP to Generative Models

The Challenges of Extending CP to Generative Models

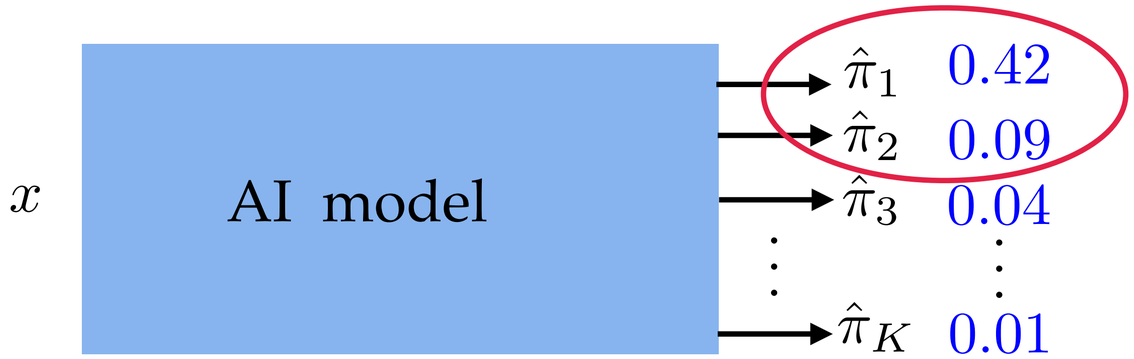


$$C(x) = \{y \in \mathcal{Y} : S(x, y) \leq q\}$$

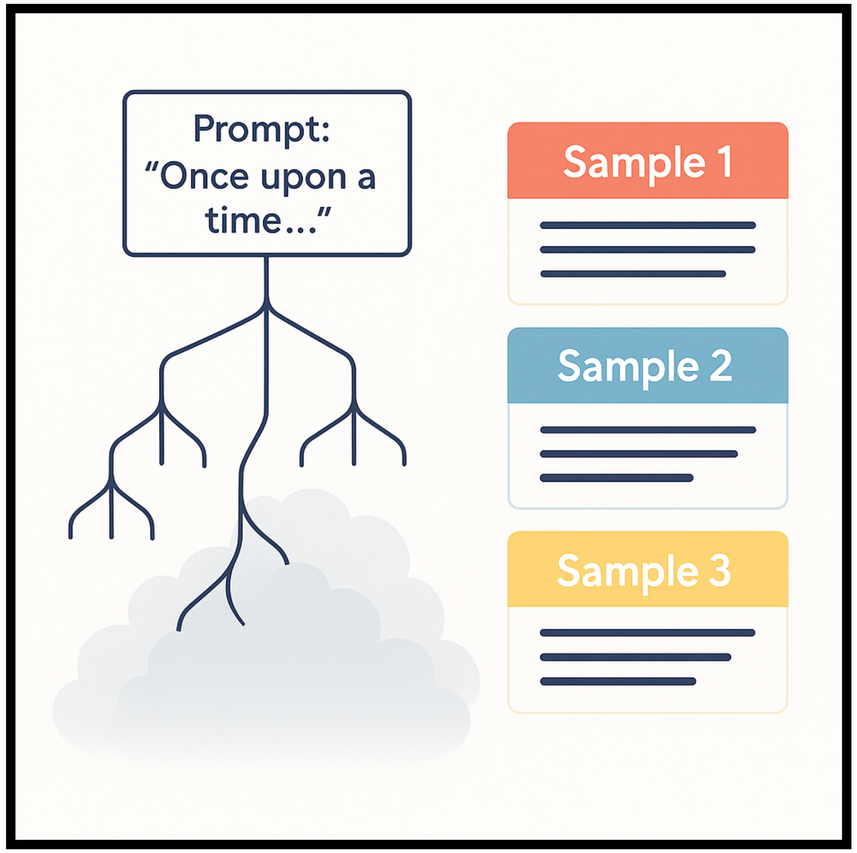
- Enumerate? ✗
- Represent compactly? ✗



The Challenges of Extending CP to Generative Models



$$C(x) = \{y \in \mathcal{Y} : S(x, y) \leq q\}$$

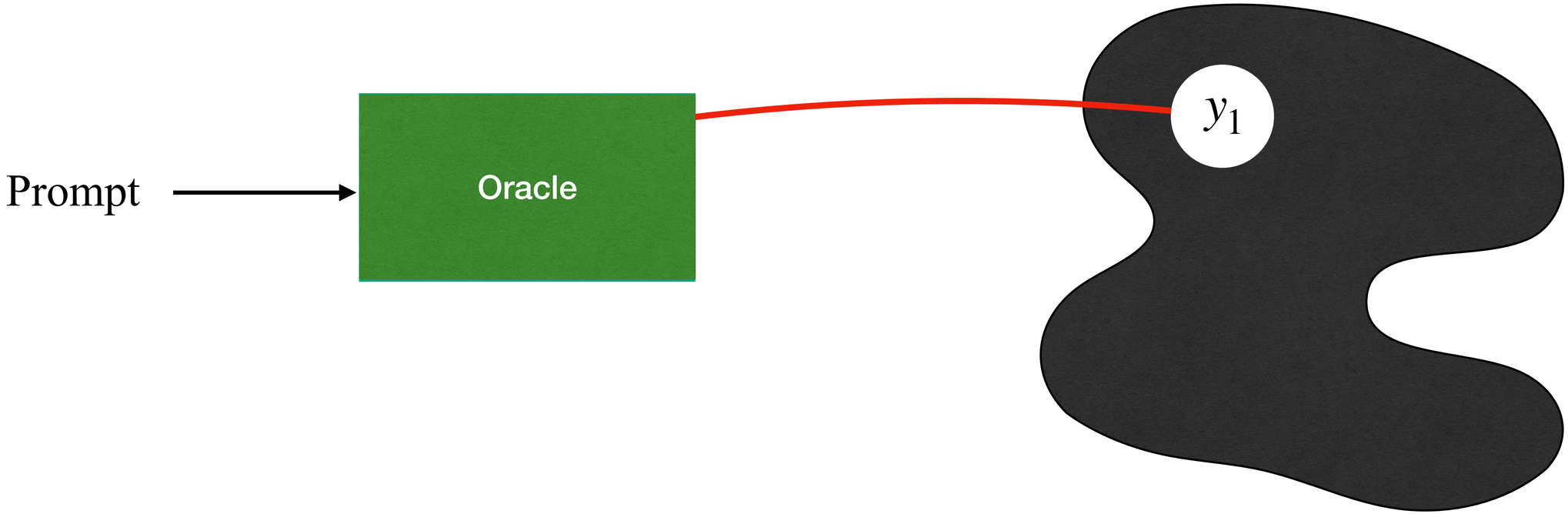


Challenge 1 : We are dealing with an unstructured + unbounded output space

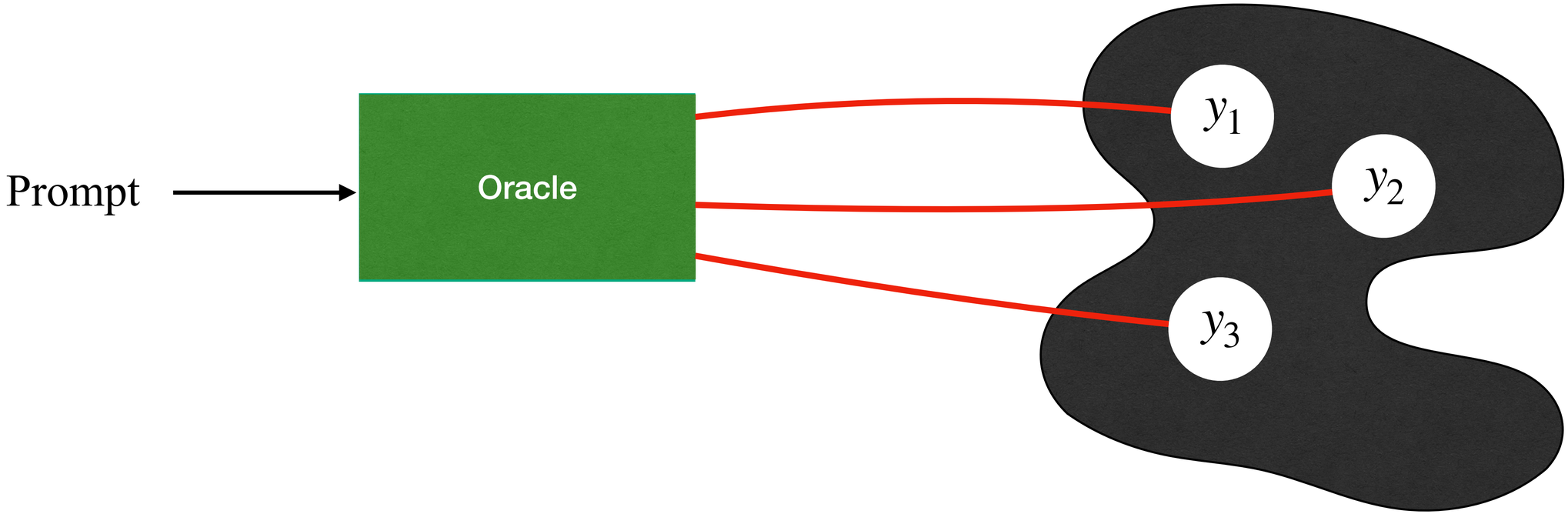
The Challenges of Extending CP to Generative Models

The Challenges of Extending CP to Generative Models

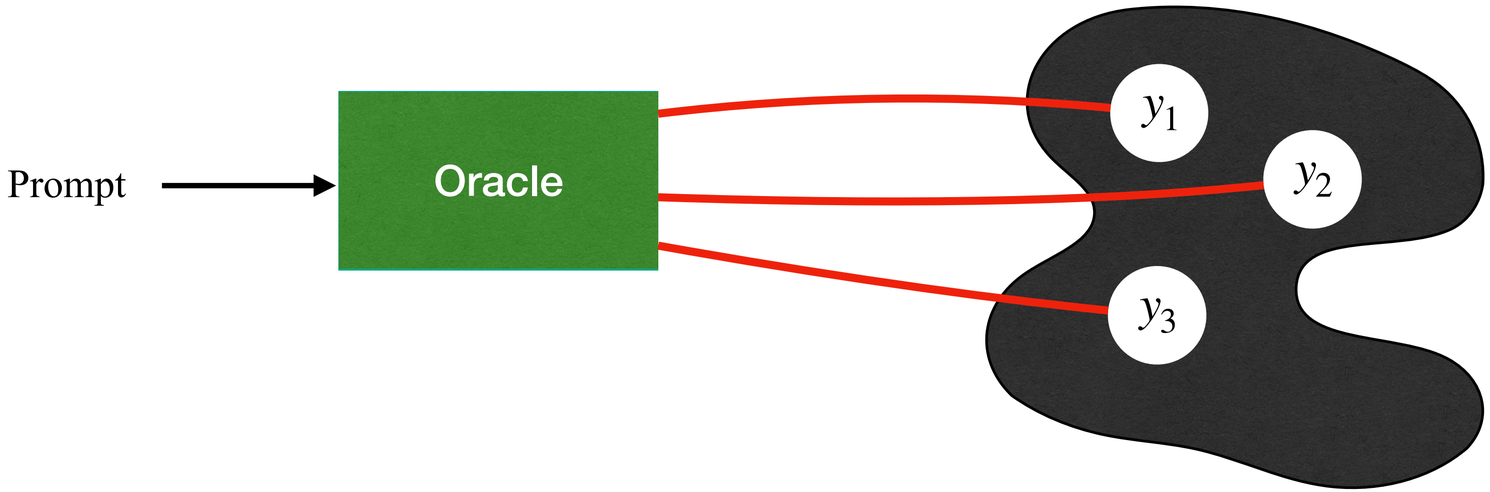
The Challenges of Extending CP to Generative Models



The Challenges of Extending CP to Generative Models

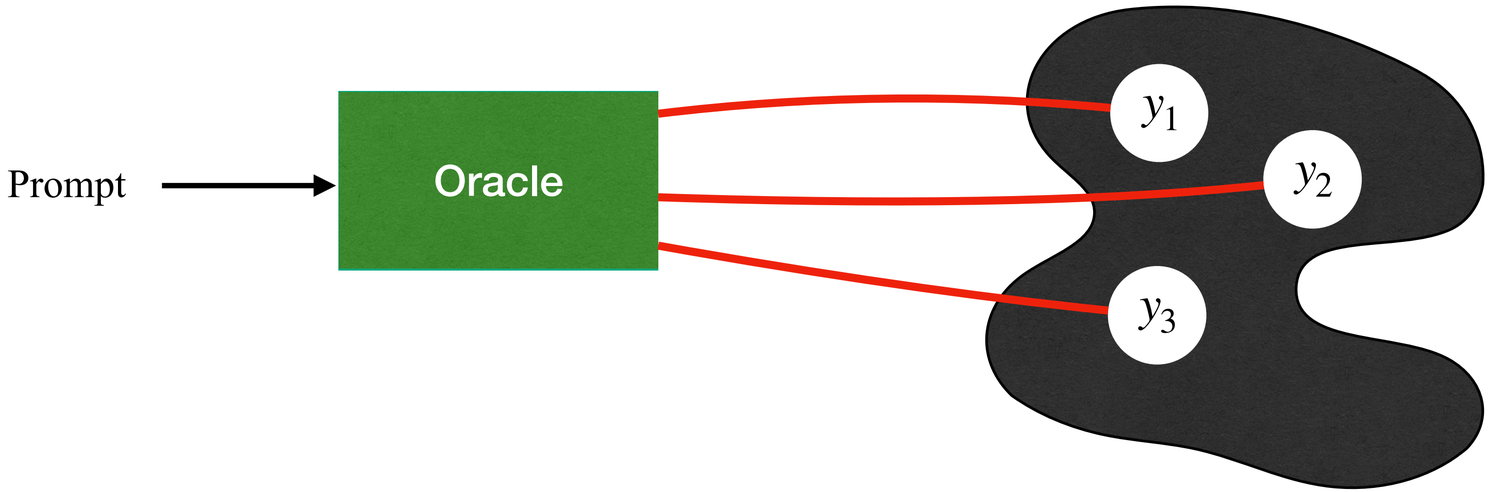


The Challenges of Extending CP to Generative Models



$$x \longrightarrow C(x) \subseteq \{\text{set of representative outputs}\}$$

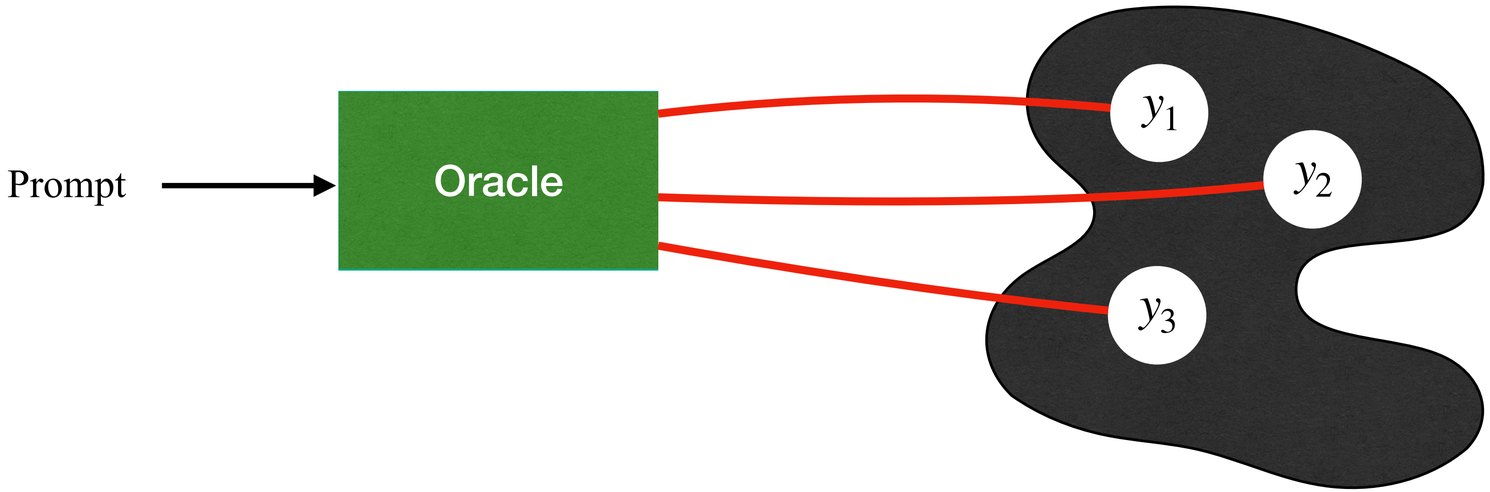
The Challenges of Extending CP to Generative Models



$$x \longrightarrow C(x) \subseteq \{\text{set of representative outputs}\}$$

$$\subseteq \{y_1, y_2, \cdots, y_t\}$$

The Challenges of Extending CP to Generative Models



$$x \longrightarrow C(x) \subseteq \{\text{set of representative outputs}\}$$

$$\subseteq \{y_1, y_2, \cdots, y_t\}$$

Challenge 2 : query only access to the output space

Conformal Prediction for Generative Models (Setting)

$(X, Y) \in \mathcal{X} \times \mathcal{Y}$

$(x, y) \sim p(x, y)$

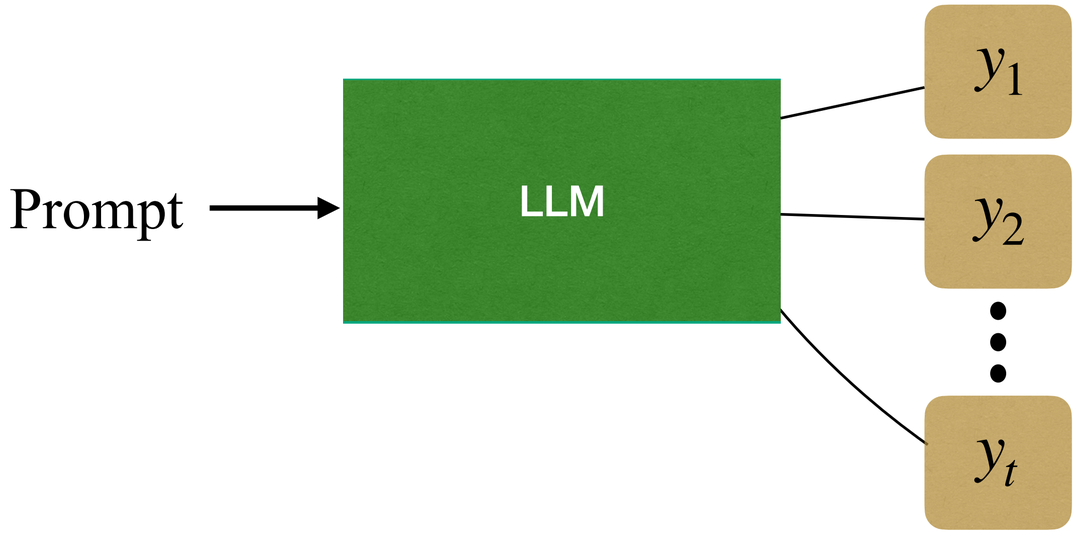
Query Oracle: $y \sim \pi(y \mid x)$

Conformal Prediction for Generative Models (Setting)

$(X, Y) \in \mathcal{X} \times \mathcal{Y}$

$(x, y) \sim p(x, y)$

Query Oracle: $y \sim \pi(y \mid x)$

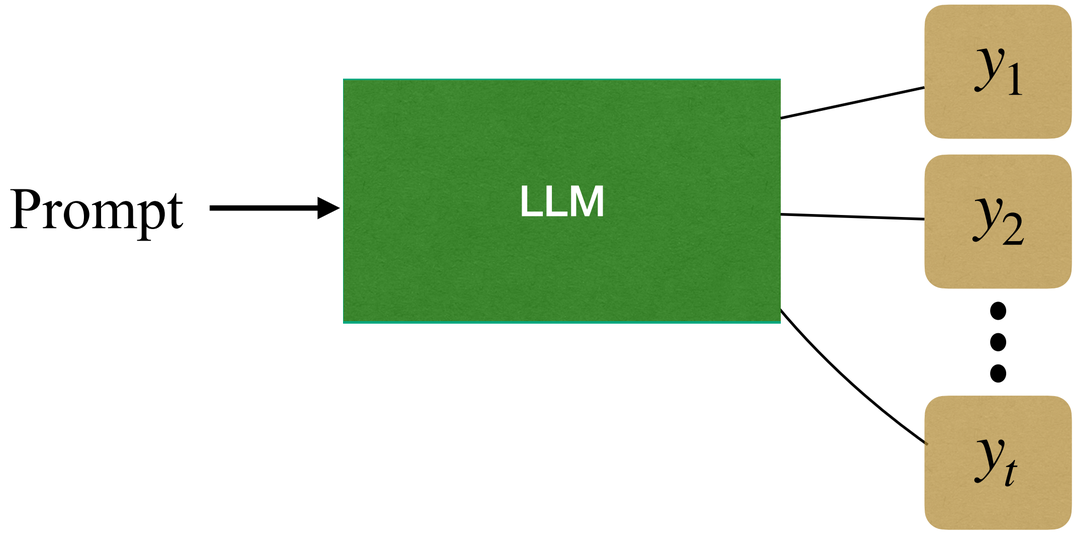


Conformal Prediction for Generative Models (Setting)

$(X, Y) \in \mathcal{X} \times \mathcal{Y}$

$(x, y) \sim p(x, y)$

Query Oracle: $y \sim \pi(y \mid x)$



We query the oracle for a finite amount of times

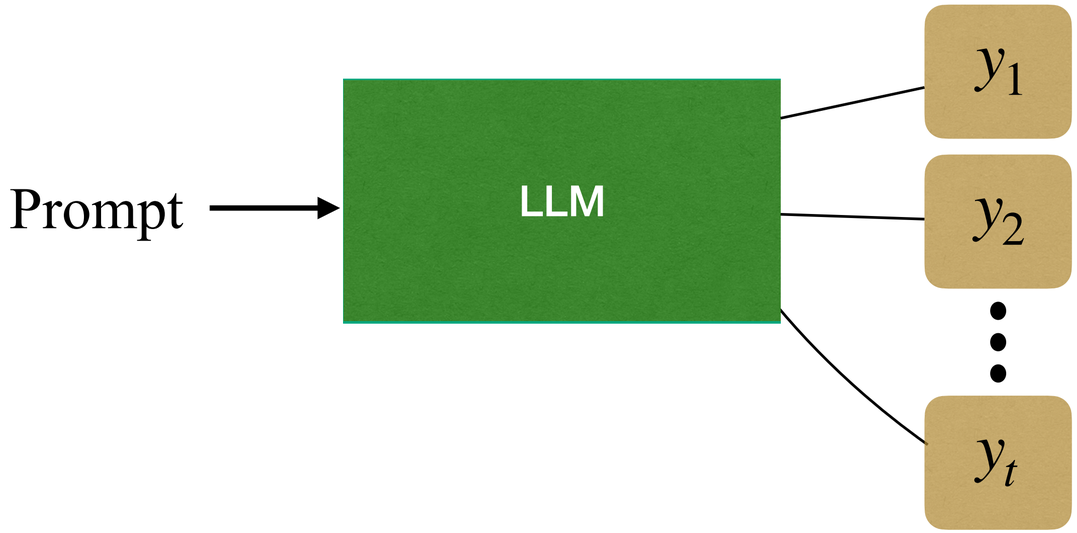
$$Z_t(x) = \{y_1^x, \ldots, y_t^x\}$$

Conformal Prediction for Generative Models (Setting)

$(X, Y) \in \mathcal{X} \times \mathcal{Y}$

$(x, y) \sim p(x, y)$

Query Oracle: $y \sim \pi(y \mid x)$



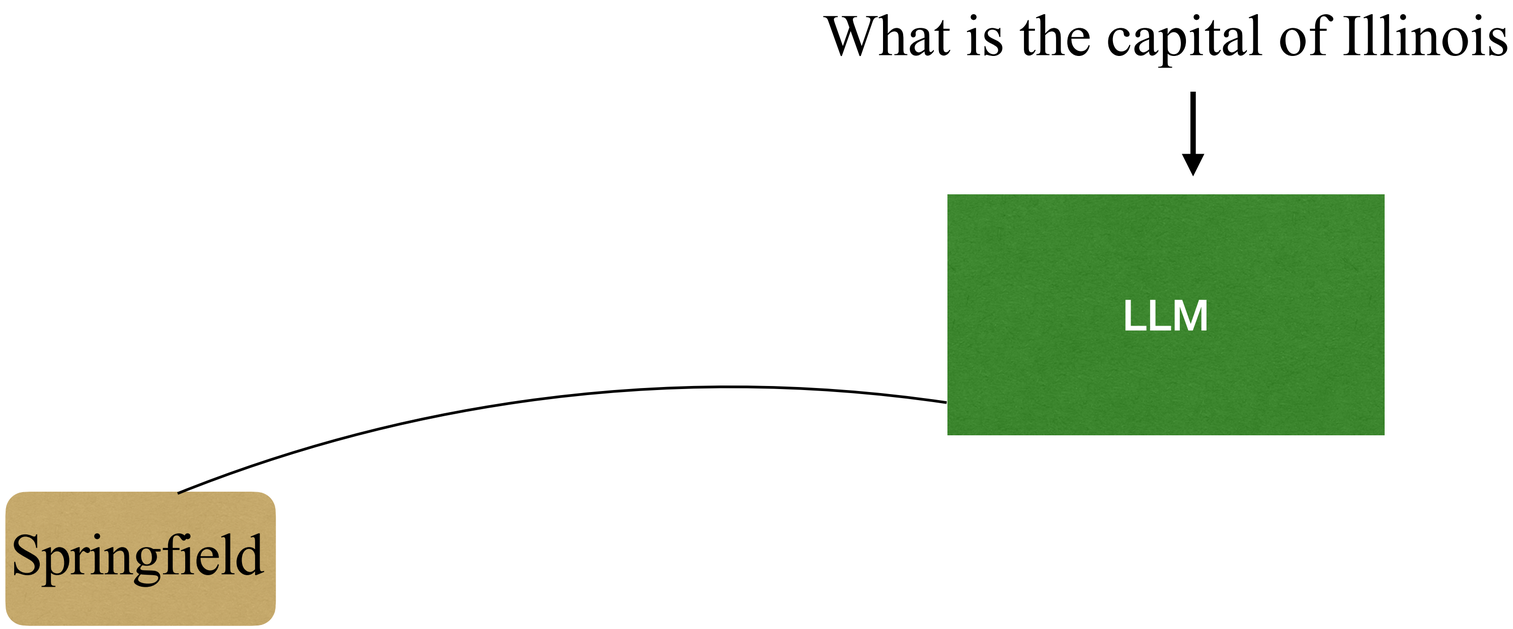
We query the oracle for a finite amount of times

$$Z_t(x) = \{y_1^x, \ldots, y_t^x\}$$

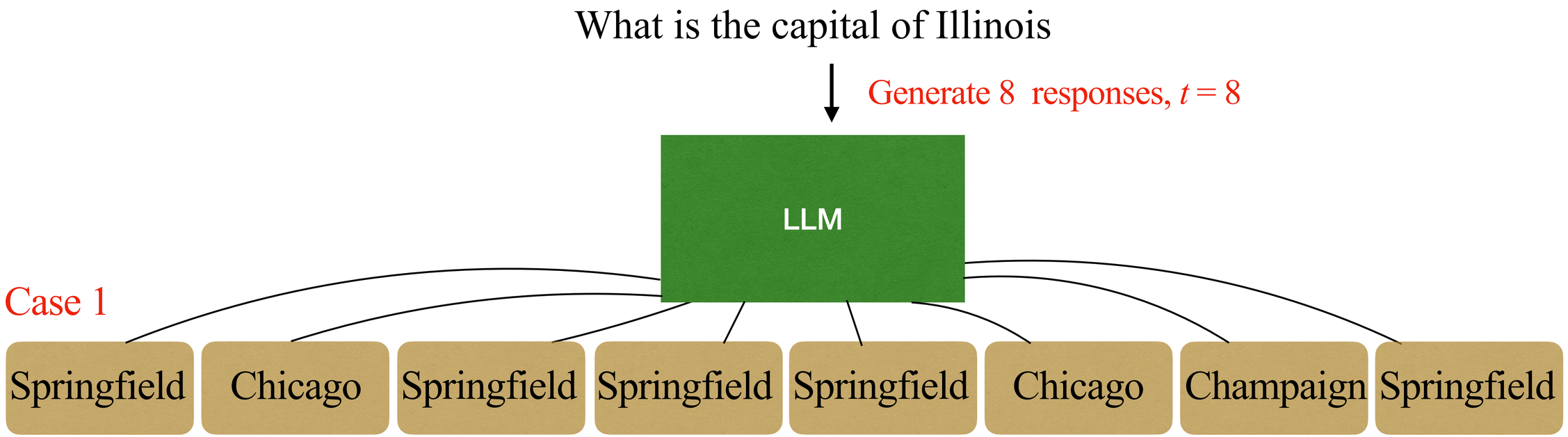
Caveat: What if the true label Y is not amount the queried outputs?

“Toy” Running Example

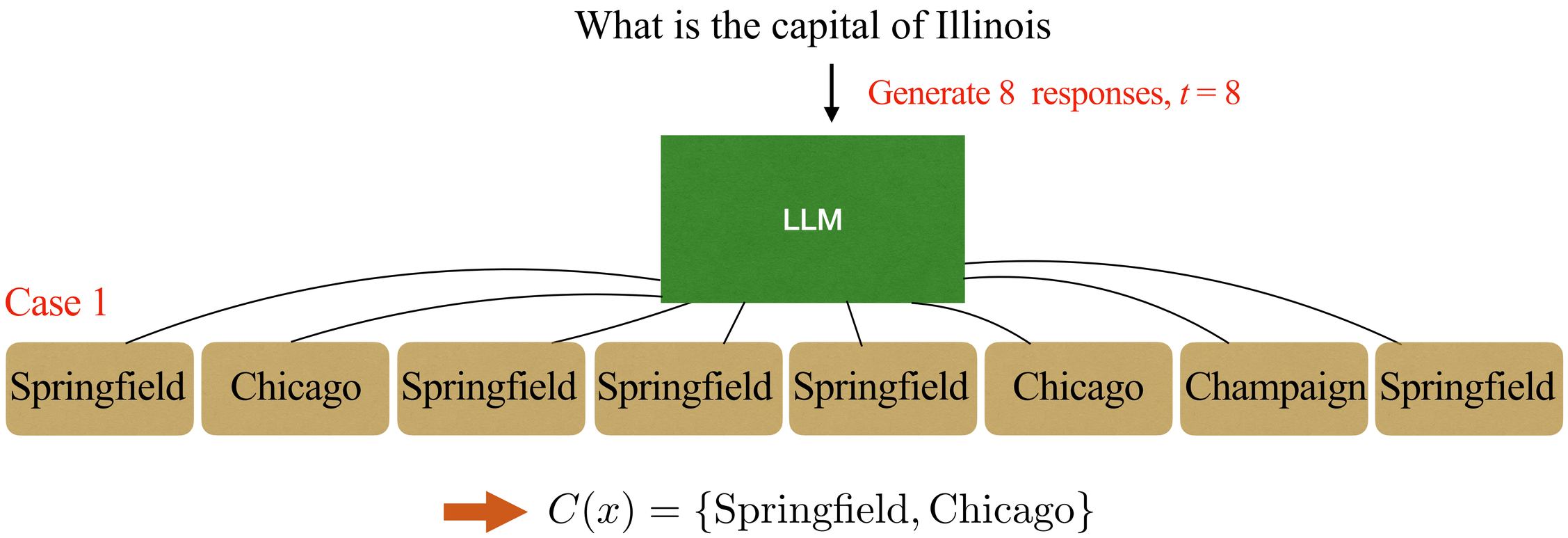
“Toy” Running Example



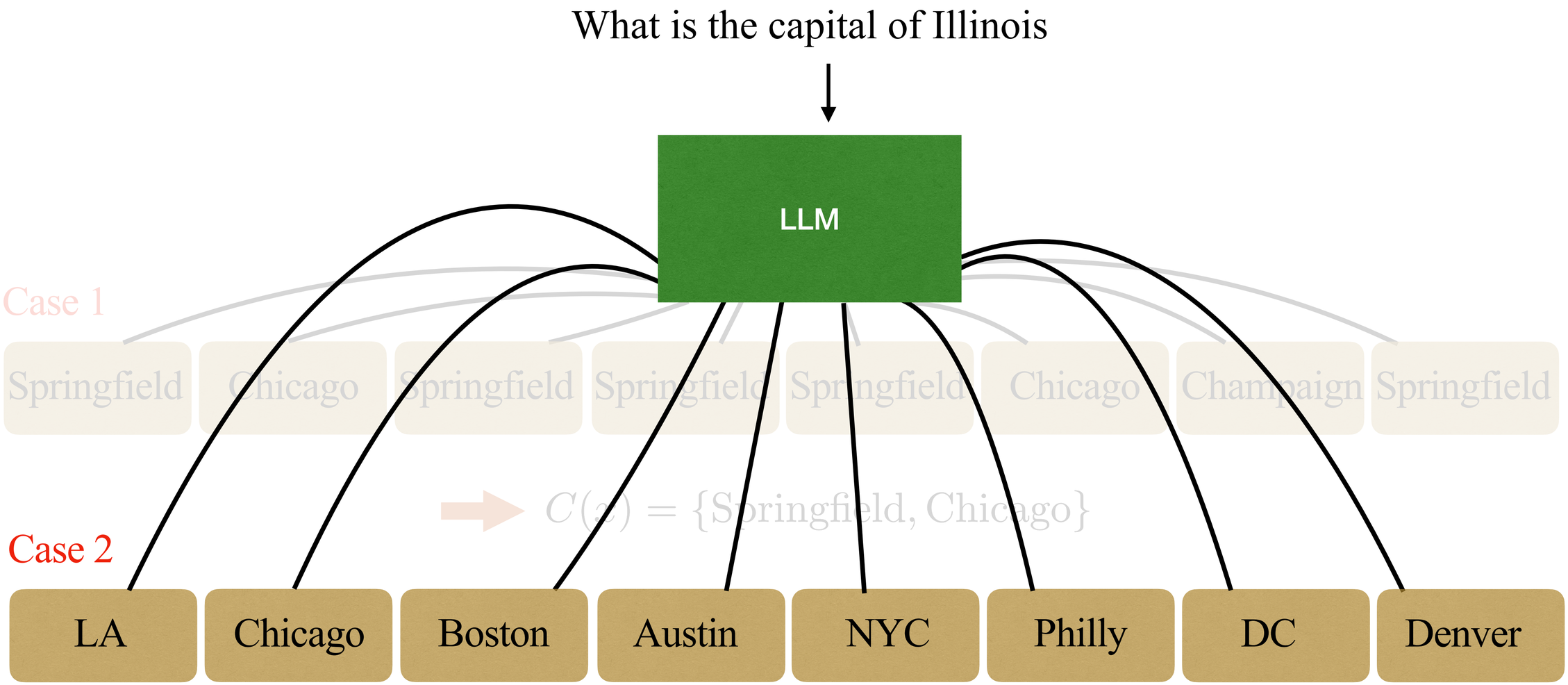
“Toy” Running Example



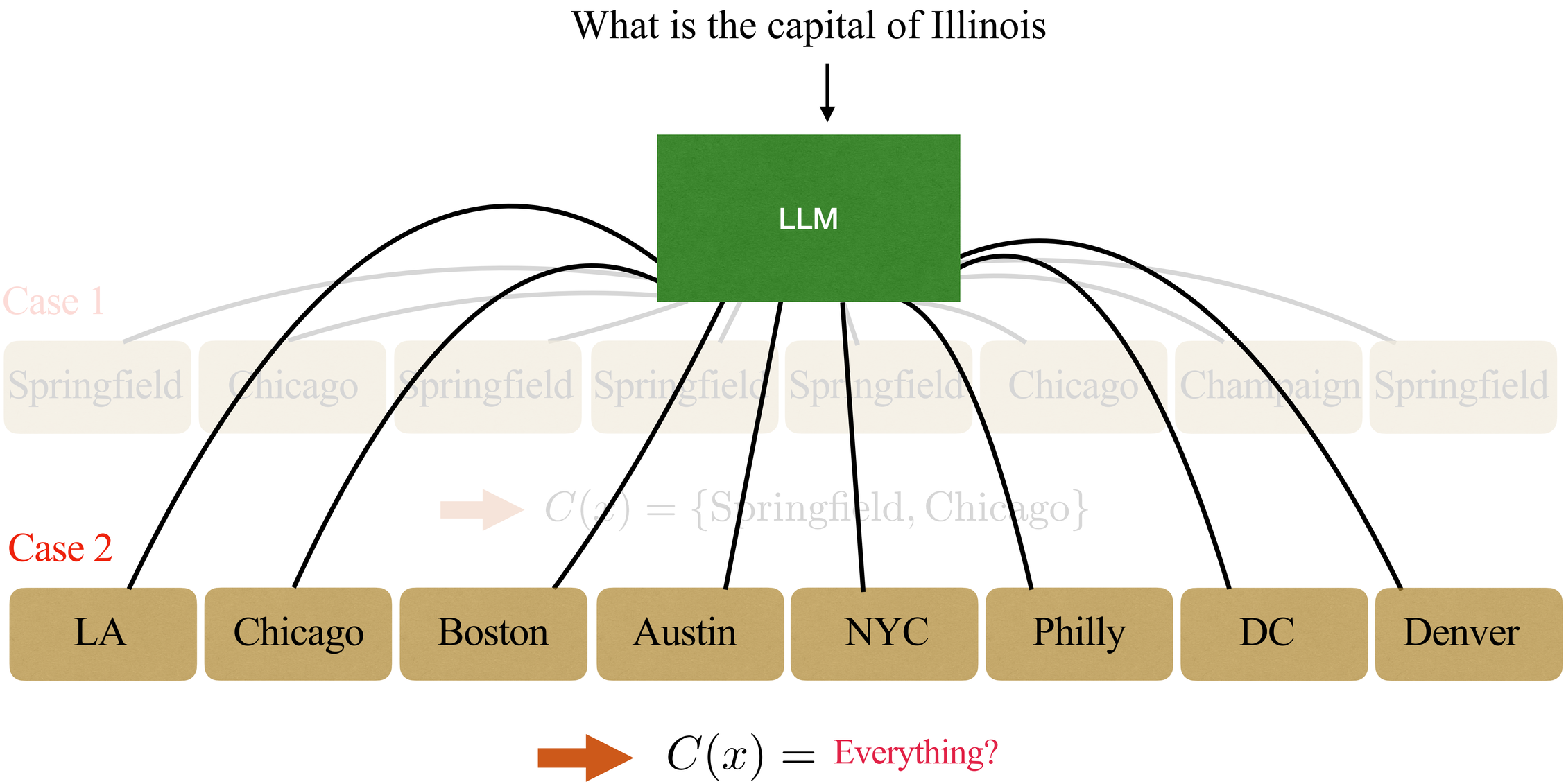
“Toy” Running Example



“Toy” Running Example



"Toy" Running Example



Everything Else : “EE”

We want to capture the full label space in finite queries

$$Z_t(x) = \{y_1^x, \ldots, y_t^x\} \cup \mathcal{Y} \backslash Z_t(x)$$

Everything Else : “EE”

We want to capture the full label space in finite queries

$$Z_t(x) = \{y_1^x, \ldots, y_t^x\} \cup \mathcal{Y} \backslash Z_t(x)$$

We have not seen!

Everything Else : “EE”

We want to capture the full label space in finite queries

$$Z_t(x) = \{y_1^x, \ldots, y_t^x\} \cup \text{EE}$$

“Everything Else” in the output space

Everything Else : “EE”

We want to capture the full label space in finite queries

$$Z_t(x) = \{y_1^x, \ldots, y_t^x\} \cup \text{EE}$$

$$C(x) \subseteq Z_t(x) \cup \text{EE}$$

Everything Else : “EE”

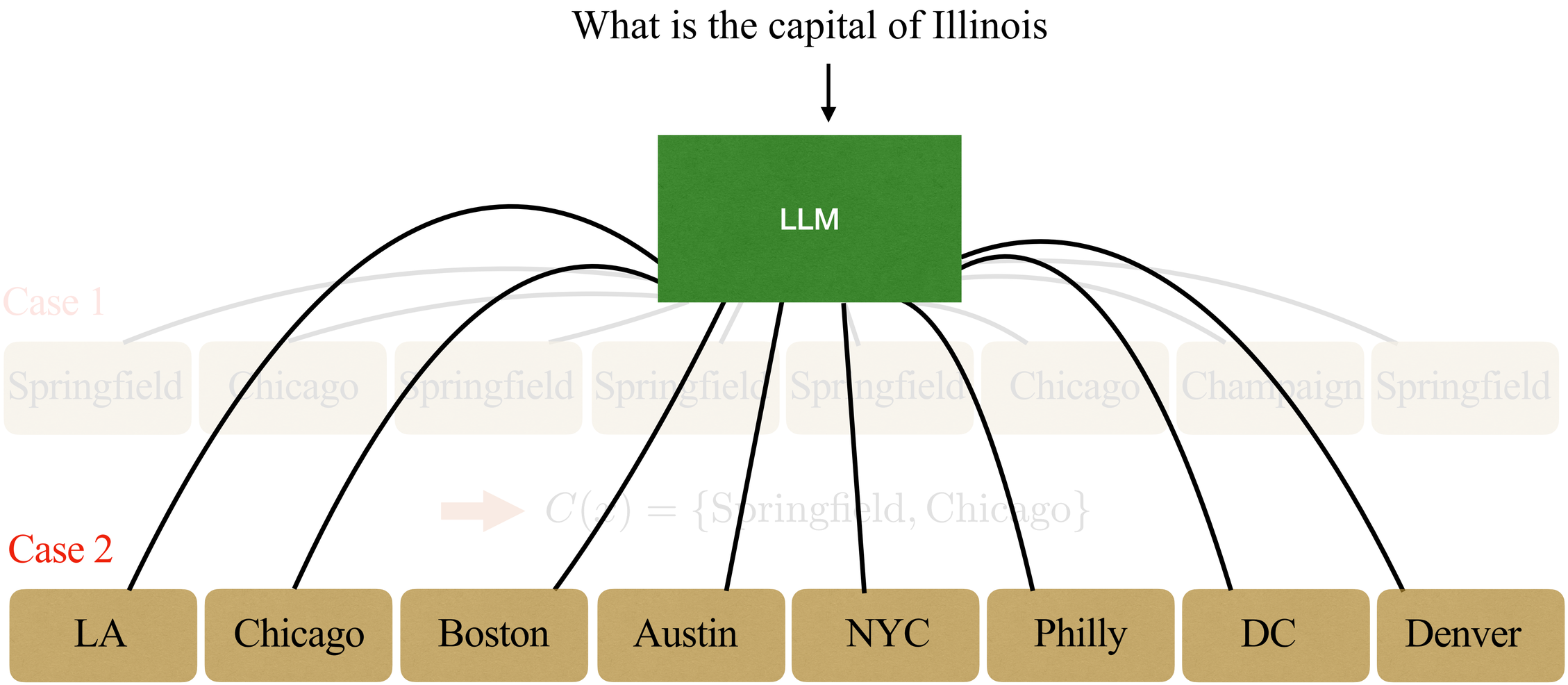
We want to capture the full label space in finite queries

$$C(x) \subseteq Z_t(x) \cup \mathsf{EE}$$

$$\text{If } \mathrm{EE} \in C(x) \rightarrow \text{Covered}$$

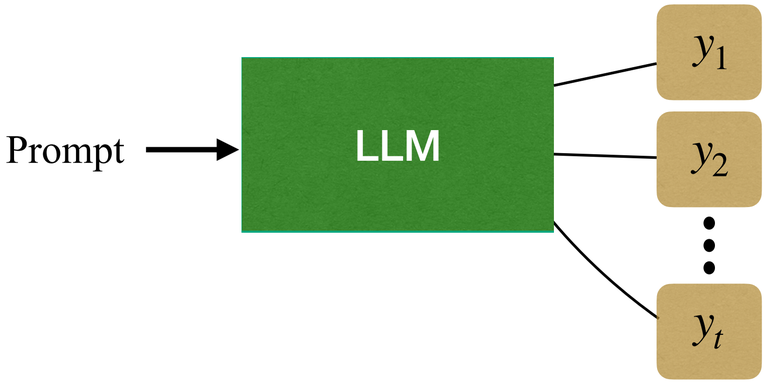
Key challenge: avoid including EE as much as possible

"Toy" Running Example



So far …

Our setting:



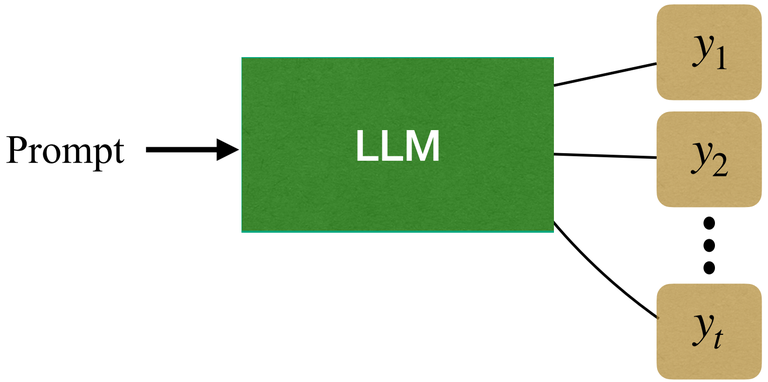
$$Z_t(x) = \{y_1^x, \cdots y_t^x\}$$

$$C(x) \subset Z_t(x) \cup \{\text{EE}\}$$

Trade-offs

So far …

Our setting:



$$Z_t(x) = \{y_1^x, \cdots y_t^x\}$$

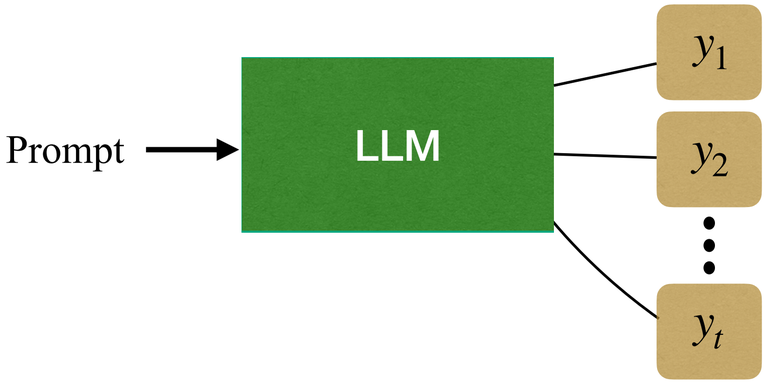
$$C(x) \subset Z_t(x) \cup \{\text{EE}\}$$

Trade-offs

Finite number of queries to explore the output space (budget)

So far …

Our setting:



$$Z_t(x) = \{y_1^x, \cdots y_t^x\}$$

$$C(x) \subset Z_t(x) \cup \{\text{EE}\}$$

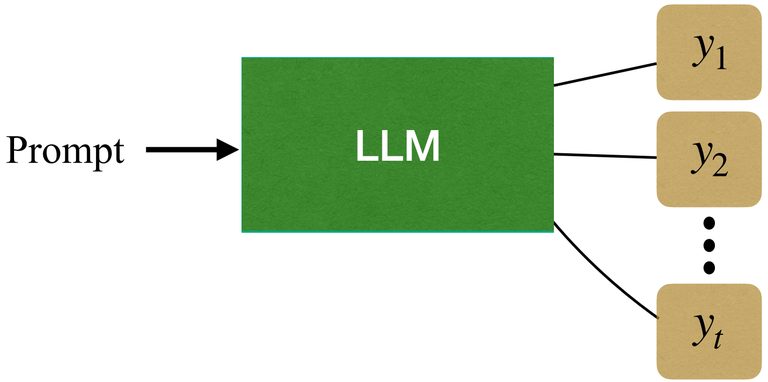
Trade-offs

Finite number of queries to explore the output space (budget)

Avoid EE as much as possible (Informativeness)

So far …

Our setting:



$$Z_t(x) = \{y_1^x, \cdots y_t^x\}$$

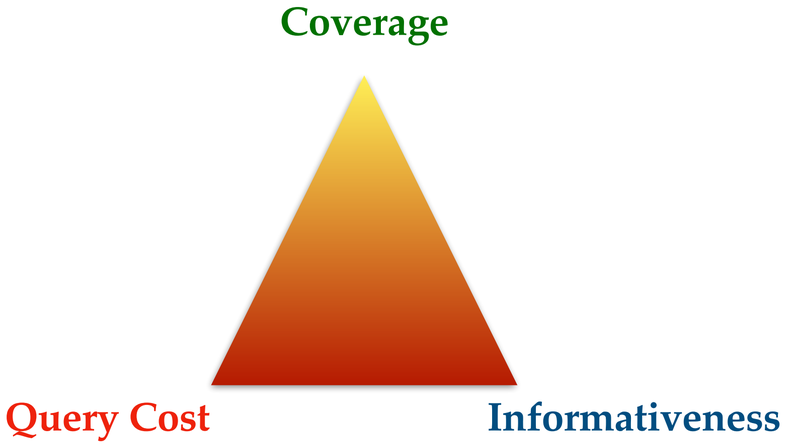
$$C(x) \subset Z_t(x) \cup \{\text{EE}\}$$

Trade-offs

Finite number of queries to explore the output space (budget)

Avoid EE as much as possible (Informativeness)

Guarantee Coverage (coverage validity)



The Main Optimization Problem (In the Population Regime)

The Main Optimization Problem (In the Population Regime)

Consider two main components:

The Main Optimization Problem (In the Population Regime)

Consider two main components:

Query policy

$$T : \mathcal{X} \to \mathbb{N} \cup \{0\}$$

$$Z(T; x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

The Main Optimization Problem (In the Population Regime)

Consider two main components:

Query policy

$$T : \mathcal{X} \to \mathbb{N} \cup \{0\}$$

$$Z(T; x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

Set map

$$f : \mathcal{X} \times 2^{\mathcal{Y}} \to 2^{\mathcal{Y} \cup EE}$$

$$C(x) = f(x, Z(T; x))$$

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \ \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\mathrm{EE} \in C(X)\} + \sum_{y \neq \mathrm{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

$$\mathbb{E}_X[T(X)] \leq B$$

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \ \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\mathrm{EE} \in C(X)\} + \sum_{y \neq \mathrm{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha \quad \text{Coverage}$$

$$\mathbb{E}_X[T(X)] \leq B$$

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \ \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\mathrm{EE} \in C(X)\} + \sum_{y \neq \mathrm{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha \quad \text{Coverage}$$

$$\mathbb{E}_X[T(X)] \leq B \quad \text{Query budget}$$

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \; \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\mathrm{EE} \in C(X)\} + \sum_{y \neq \mathrm{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

Informativeness

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

Coverage

$$\mathbb{E}_X[T(X)] \leq B$$

Query budget

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \; \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\text{EE} \in C(X)\} + \sum_{y \neq \text{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

Informativeness

EE

set size

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

Coverage

$$\mathbb{E}_X[T(X)] \leq B$$

Query budget

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \ \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\mathrm{EE} \in C(X)\} + \sum_{y \neq \mathrm{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

$$\mathbb{E}_X[T(X)] \leq B$$

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \ \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\text{EE} \in C(X)\} + \sum_{y \neq \text{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

$$\mathbb{E}_X[T(X)] \leq B$$

1st step: Solve in the population regime (assume access to $p(y \mid x)$)

The Main Optimization Problem (In the Population Regime)

$$\min_{f,T} \ \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\mathrm{EE} \in C(X)\} + \sum_{y \neq \mathrm{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

$$\mathbb{E}_X[T(X)] \leq B$$

1st step: Solve in the population regime (assume access to $p(y \mid x)$)

2nd step: mimic optimal solution in finite sample regime
(assume access to oracle $\pi(y \mid x)$)
(assume access to finite calibration data)

1st step: Solve in the population regime (assume access to $p(y \mid x)$)

$$\min_{f,T} \ \mathbb{E}_X \left[\lambda \, \mathbf{1}\{\text{EE} \in C(X)\} + \sum_{y \neq \text{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

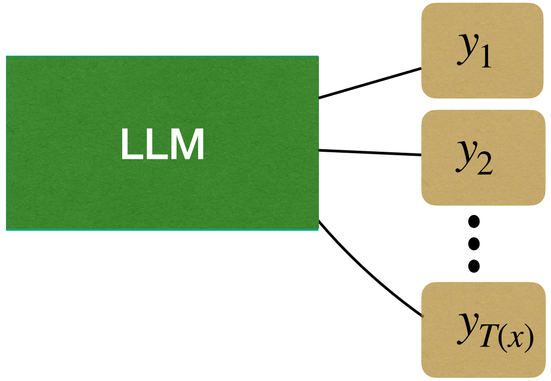
$$\boxed{\mathbb{E}_X[T(X)] \leq B}$$

Let's ask:

What is the role of the query policy?

What is the role of the query policy?

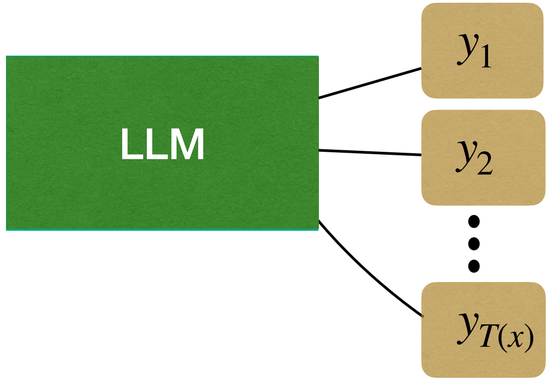
$$x \to T(x) \to$$



$$\longrightarrow \quad Z(T;x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

What is the role of the query policy?

$$x \to T(x) \to$$

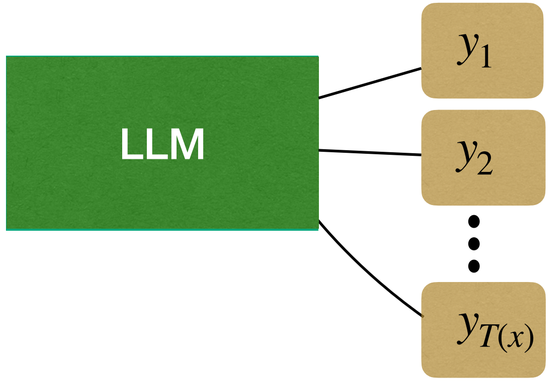


$$\longrightarrow \quad Z(T;x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

Need to sample enough to not miss the correct label !

What is the role of the query policy?

$$x \to T(x) \to$$



$$\longrightarrow \quad Z(T;x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

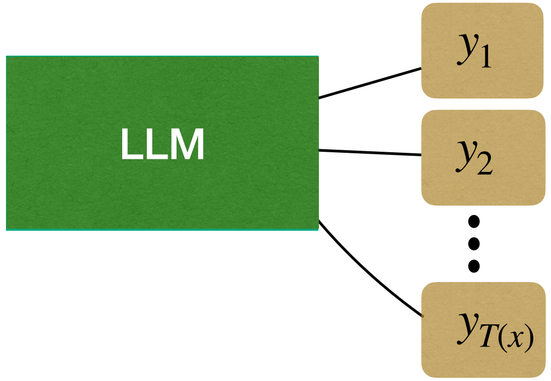
Need to sample enough to not miss the correct label !

But …

Budget constraint: $\mathbb{E}_X[T(X)] \leq B$

What is the role of the query policy?

$$x \to T(x) \to$$



$$\longrightarrow \quad Z(T;x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

Need to sample enough to not miss the correct label !

$$\mathbb{E}_X[T(X)] \leq B$$

Intuition: query policy should be designed to minimize the chance of missing the correct label

A decoupled analysis - two principles

A decoupled analysis - two principles

Question 1: What is the optimal query policy to minimize the change of missing the correct label?

Query policy

$$T : \mathcal{X} \to \mathbb{N} \cup \{0\}$$

$$Z(T; x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

A decoupled analysis - two principles

Question 1: What is the optimal query policy to minimize the change of missing the correct label?

Fix $T(\,.\,)$

A decoupled analysis - two principles

Question 1: What is the optimal query policy to minimize the change of missing the correct label?

Fix $T(\,.\,)$

Question 2: What is the optimal set map for constructing valid, informative sets $C(\,.\,)$?

Set map

$$f : \mathcal{X} \times 2^{\mathcal{Y}} \rightarrow 2^{\mathcal{Y} \cup EE}$$

$$C(x) = f(x, Z(T; x))$$

A decoupled analysis - two principles

Question 1: What is the optimal query policy to minimize the change of missing the correct label?

Fix $T(\,.\,)$

Question 2: What is the optimal set map for constructing valid, informative sets $C(\,.\,)$?

Question 1: What is the optimal query policy to minimize the change of missing the correct label?

Principle 1: Optimal Query Policy

Goal: minimize the chance of missing the true label, under finite budget

$$Z_t(x) = \{y_1^x, \ldots, y_t^x\}$$

Principle 1: Optimal Query Policy

Goal: minimize the chance of missing the true label, under finite budget

$$Z_t(x) = \{y_1^x, \ldots, y_t^x\}$$

$$\theta(x, t) = \mathbb{P}_{Y, Z_t(x)}[Y \notin Z_t(x) \mid X = x]$$

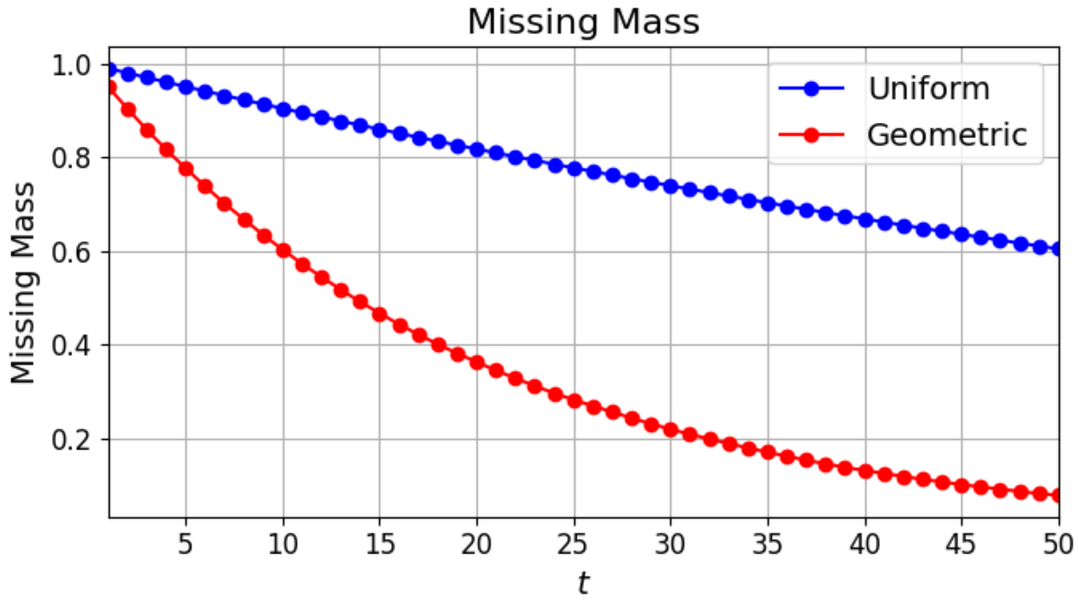
Missing Mass

Principle 1: Optimal Query Policy

Missing Mass : a natural objective

Missing Mass

$$\theta(x,t) = \theta(t) = \mathbb{P}_{Y,Z_t(x)}[Y \notin Z_t(x) \mid X = x]$$

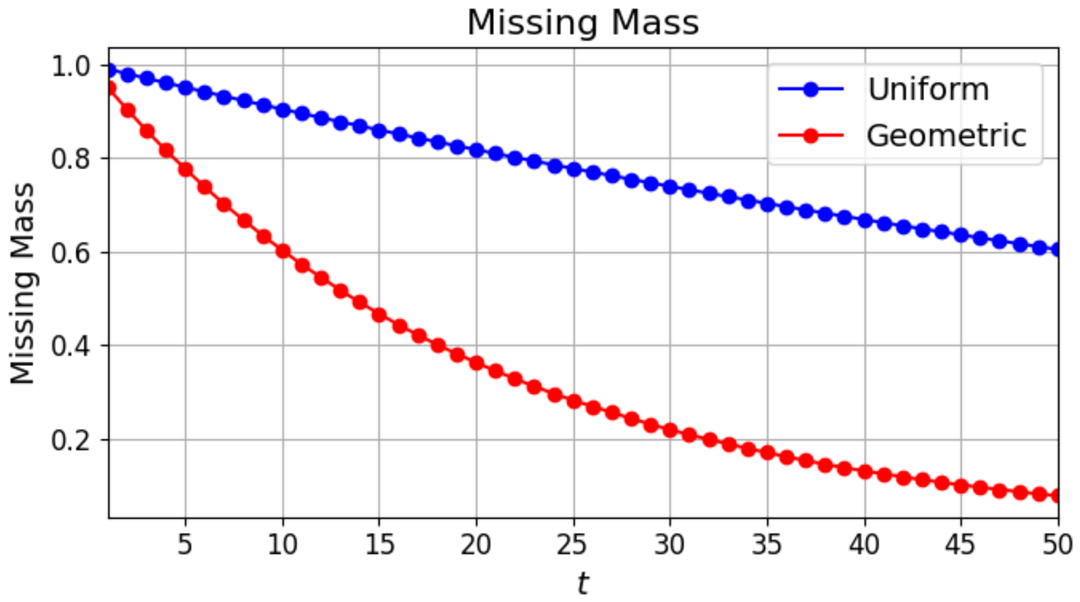


Principle 1: Optimal Query Policy

Missing Mass : a natural objective

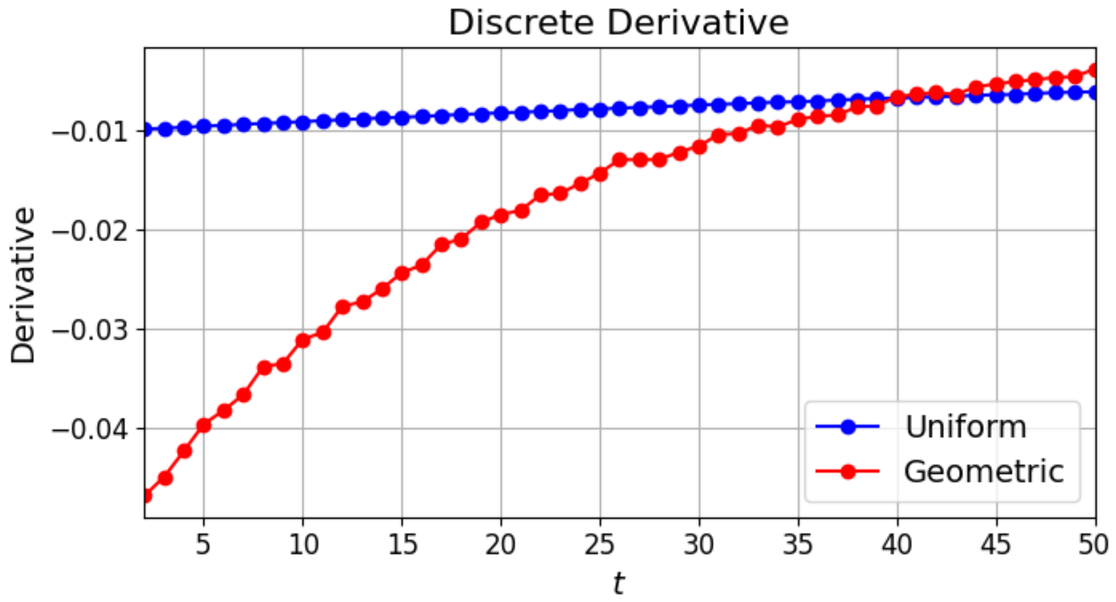
Missing Mass

$$\theta(x,t) = \theta(t) = \mathbb{P}_{Y,Z_t(x)}[Y \notin Z_t(x) \mid X = x]$$



Discrete Derivative

$$\Delta(x,t) := \theta(x,t+1) - \theta(x,t)$$



Principle 1: Optimal Query Policy

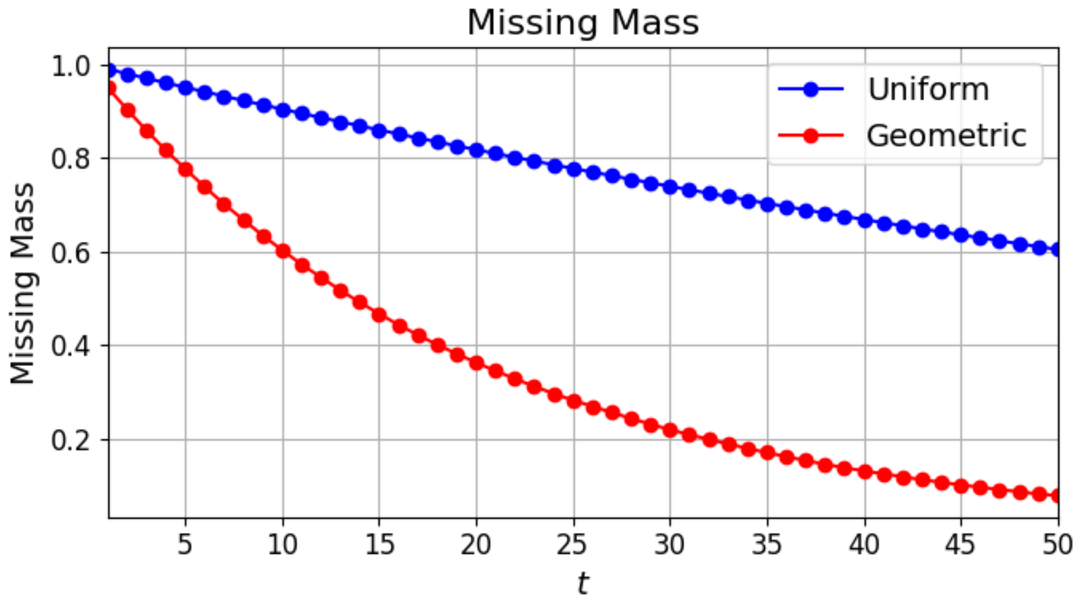
Missing Mass : a natural objective

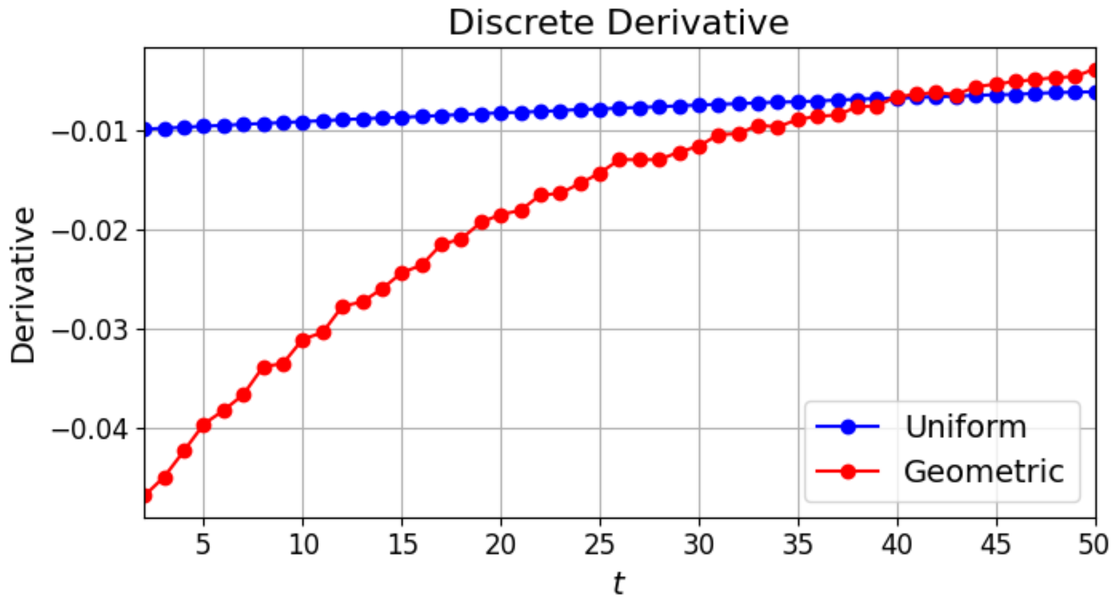
Missing Mass

$$\theta(x,t) = \theta(t) = \mathbb{P}_{Y,Z_t(x)}[Y \notin Z_t(x) \mid X = x]$$

Discrete Derivative

$$\Delta(x,t) := \theta(x,t+1) - \theta(x,t)$$





Missing mass decreases in t, with diminishing returns

Principle 1: Optimal Query Policy

$$\min_{T(.):\mathcal{X}\to\mathbb{N}\cup\{0\}} \quad \mathbb{E}_X\left[\theta(X, T(X))\right]$$

$$s.t \quad \mathbb{E}_X\left[T(X)\right] \leq B$$

Principle 1: Optimal Query Policy

$$\min_{T(.):\mathcal{X}\to\mathbb{N}\cup\{0\}} \quad \mathbb{E}_X\left[\theta(X, T(X))\right]$$

$$s.t \quad \mathbb{E}_X\left[T(X)\right] \leq B$$

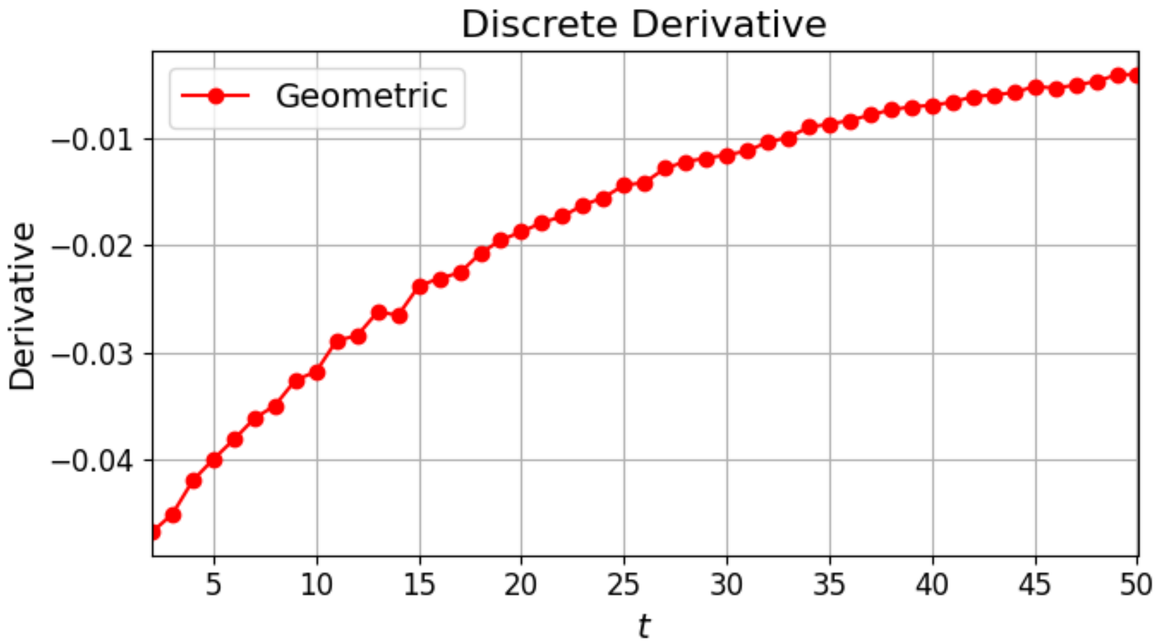
Theorem: Let $T^*(x)$ be the optimal solution. There exists $\beta^* \in \mathbb{R}$ such that for all x

$$\Delta(x, T^*(x)) = \beta^*$$

Principle 1: Optimal Query Policy

For each $x \in \mathcal{X}$, consider the derivative of the missing mass:

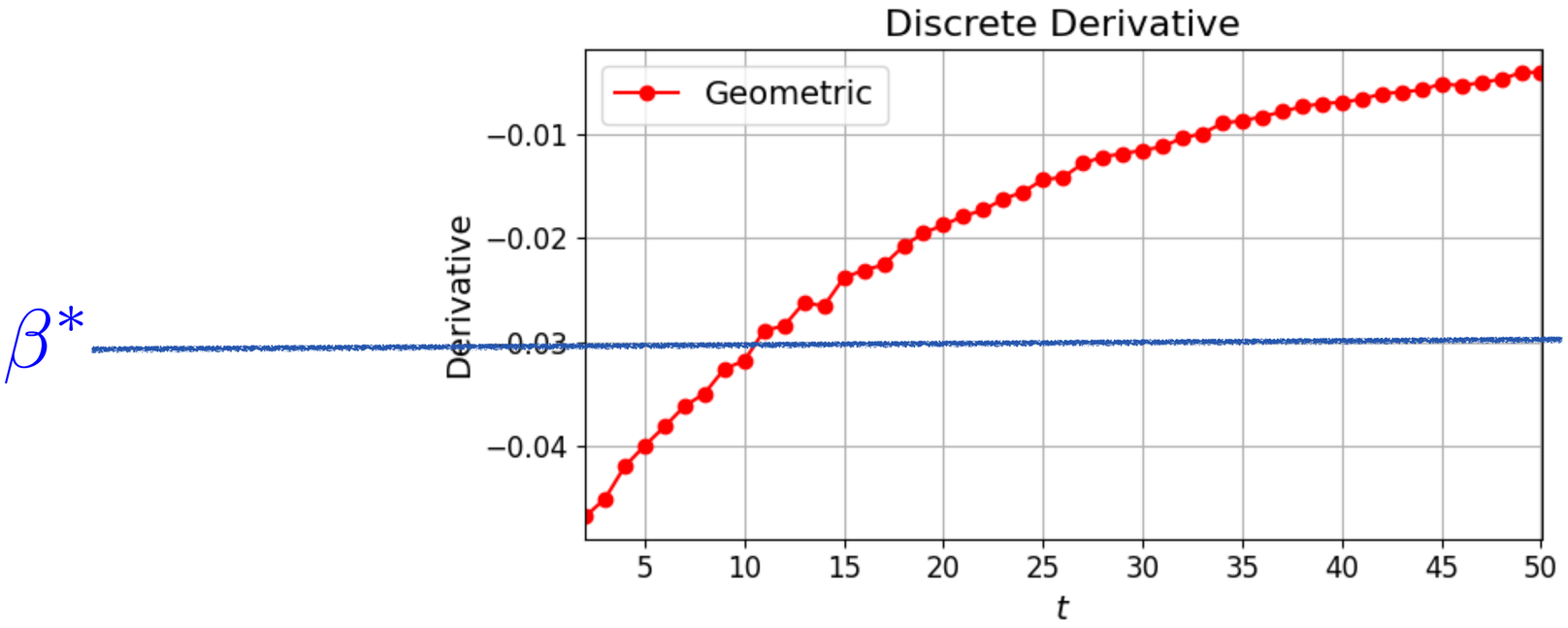
$$\Delta(x,t) := \theta(x,t+1) - \theta(x,t)$$



Principle 1: Optimal Query Policy

For each $x \in \mathcal{X}$, consider the derivative of the missing mass:

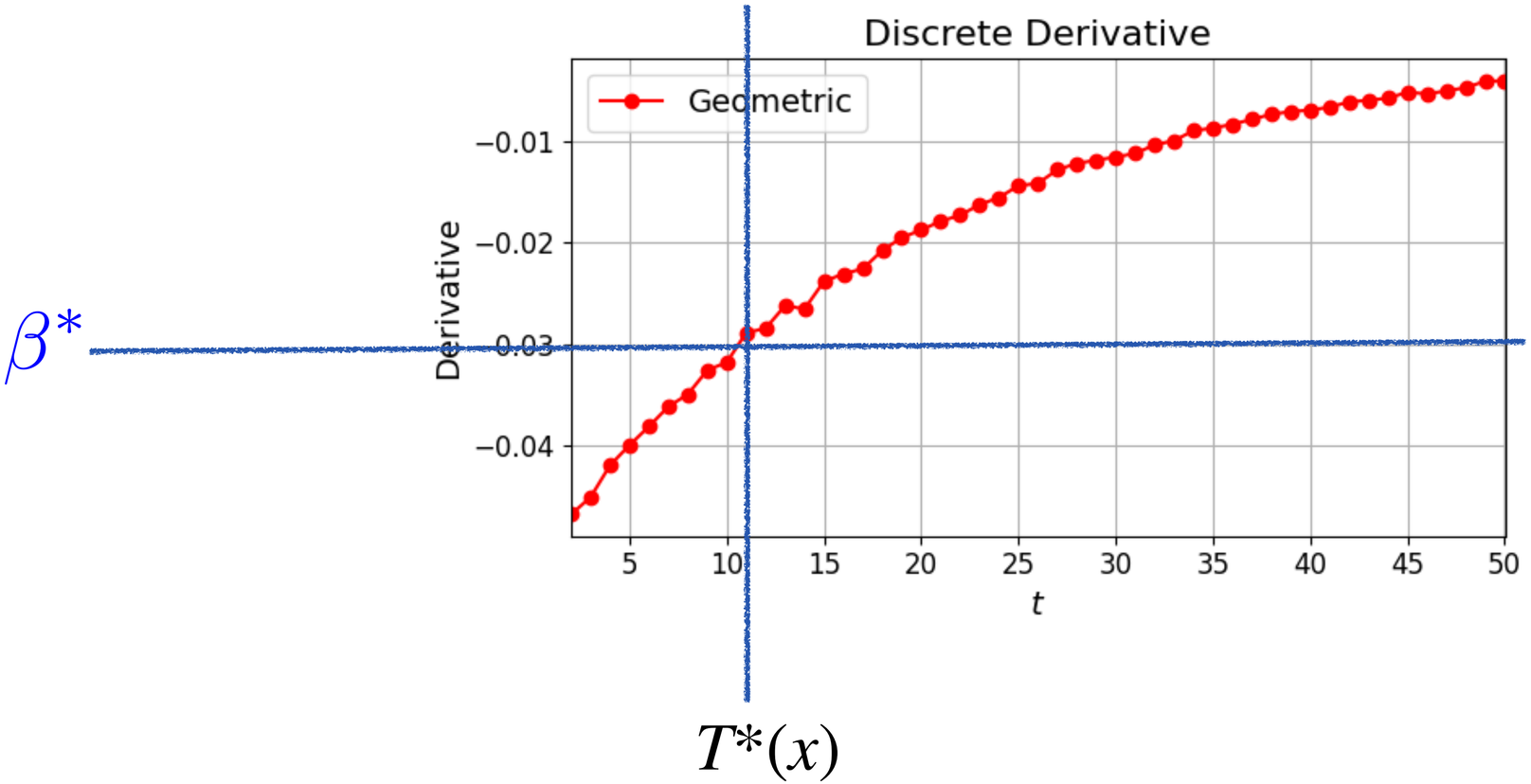
$$\Delta(x,t) := \theta(x,t+1) - \theta(x,t)$$



Principle 1: Optimal Query Policy

For each $x \in \mathcal{X}$, consider the derivative of the missing mass:

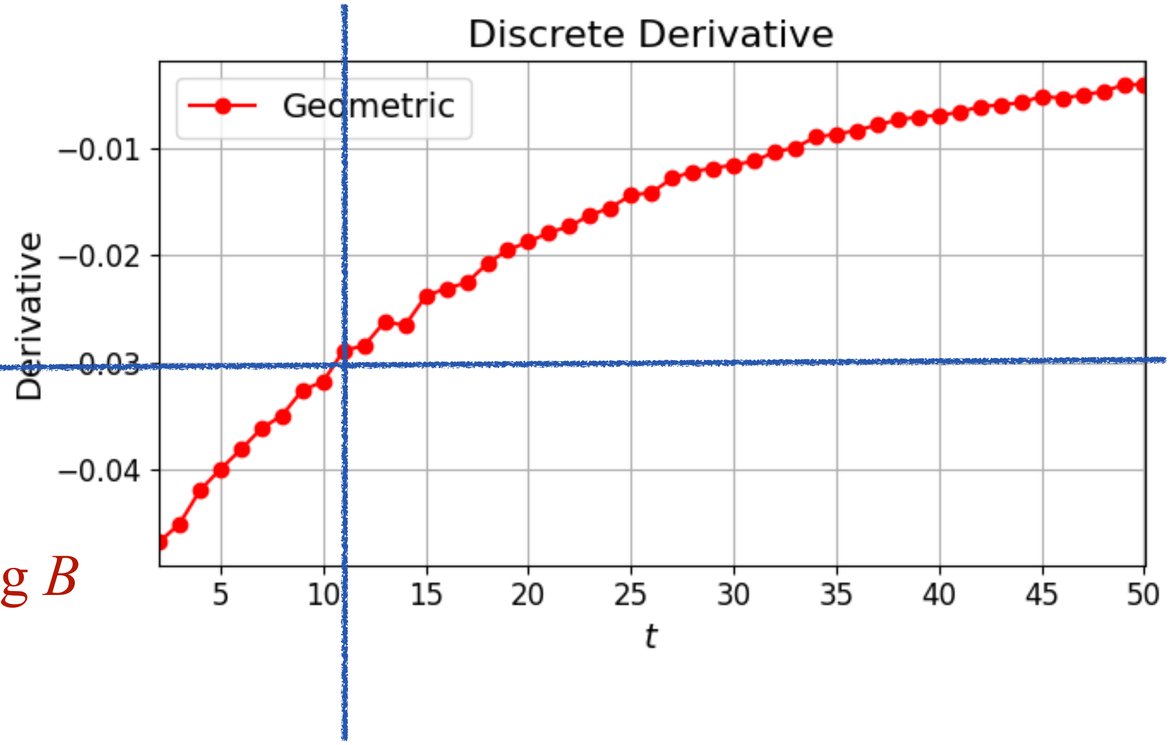
$$\Delta(x,t) := \theta(x,t+1) - \theta(x,t)$$



Principle 1: Optimal Query Policy

For each $x \in \mathcal{X}$, consider the derivative of the missing mass:

$$\Delta(x,t) := \theta(x,t+1) - \theta(x,t)$$



β^*

- single scalar
- found by matching B

$T^*(x)$

A decoupled analysis - two principles

Fix $T(\,.\,)$

Question 2: **What is the optimal set map for constructing valid, informative sets $C(\,.\,)$?**

Question 2: **What is the optimal set map for constructing valid, informative sets $C(\,.\,)$?**

Principle 2: Optimal Set map

Queried samples

$$Z(x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

Set map

$$f : \mathcal{X} \times 2^{\mathcal{Y}} \to 2^{\mathcal{Y} \cup EE}$$

Principle 2: Optimal Set map

Queried samples

$$Z(x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

Set map

$$f : \mathcal{X} \times 2^{\mathcal{Y}} \rightarrow 2^{\mathcal{Y} \cup EE}$$

$$\min_{f(.)} \quad \mathbb{E}_X \left[\lambda \mathbf{1}\{\text{EE} \in C(X)\} + \sum_{y \neq \text{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y} \left[Y \in C(X) \right] \geq 1 - \alpha$$

Principle 2: Optimal Set map

$$\min_{f(.)} \quad \mathbb{E}_X \left[\lambda \mathbf{1}\{\text{EE} \in C(X)\} + \sum_{y \neq \text{EE}} \mathbf{1}\{y \in C(X)\} \right]$$

$$s.t \qquad \Pr_{X,Y}\left[Y \in C(X)\right] \geq 1 - \alpha$$

Theorem: Let f_λ^* be the optimal solution. There exists a scalar threshold $q^* \in \mathbb{R}^+$ s.t for every x

$$C(x) = \{y \in Z(x) \cup EE : S(x, y) \leq q^*\}$$

where with $p(EE \mid x) = Pr[Y \notin Z(x) \mid X = x]$

$$S(x, y) = \begin{cases} 1 - p(y \mid x), & if \quad y \neq EE \\ 2 - p(y \mid x), & if \quad y = EE \end{cases}$$

In summary ~ over population … :

Question 1: What is the optimal query policy to minimize the change of missing the correct label?

Fix $T(\,.\,)$

Question 2: What is the optimal set map for constructing valid, informative sets $C(\,.\,)$?

In summary ~ over population … :

Assume access to $p(y \mid x)$

Answer 1: Find optimal set query by $\Delta(x, T^*(x)) = \beta^*$

In summary ~ over population … :

Assume access to $p(y \mid x)$

Answer 1: Find optimal set query by $\Delta(x, T^*(x)) = \beta^*$

$$Z(x) = \{y_1^x, \ldots, y_{T^*(x)}^x\}$$

In summary ~ over population … :

Assume access to $p(y \mid x)$

Answer 1: Find optimal set query by $\Delta(x, T^*(x)) = \beta^*$

$$Z(x) = \{y_1^x, \ldots, y_{T^*(x)}^x\}$$

Answer 2: Build optimal set $C(x) = \{y \in Z(x) \cup EE : S(x, y) \leq q^*\}$

where: $S(x, y) = \begin{cases} 1 - p(y \mid x), & if \quad y \neq EE \\ 2 - p(y \mid x), & if \quad y = EE \end{cases}$

In summary ~~– over population … :~~ **In finite sample regime**

Assume access to $p(y \mid x)$

Answer 1: Find optimal set query by $\Delta(x, T^*(x)) = \beta^*$

$$Z(x) = \{y_1^x, \ldots, y_{T^*(x)}^x\}$$

Answer 2: Build optimal set $C(x) = \{y \in Z(x) \cup EE : S(x, y) \leq q^*\}$

where: $S(x, y) = \begin{cases} 1 - p(y \mid x), & if \quad y \neq EE \\ 2 - p(y \mid x), & if \quad y = EE \end{cases}$

In summary ~~– over population … :~~ **In finite sample regime**

~~– Assume access to $p(y|x)$ –~~ **Need to estimate**

Answer 1: Find optimal set query by $\Delta(x, T^*(x)) = \beta^*$

$$Z(x) = \{y_1^x, \ldots, y_{T^*(x)}^x\}$$

Answer 2: Build optimal set $C(x) = \{y \in Z(x) \cup EE : S(x,y) \leq q^*\}$

where: $$S(x,y) = \begin{cases} 1 - p(y|x) & if \quad y \neq EE \\ 2 - p(y|x) & if \quad y = EE \end{cases}$$

Our Main Algorithm (In the finite-sample regime)

Our Main Algorithm (In the finite-sample regime)

We don’t have the true $p(y \mid x)$

Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Need to estimate $\theta(x, t)$ and $\Delta(x, t)$

Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Need to estimate $\theta(x, t)$ and $\Delta(x, t)$

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$

Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Need to estimate $\theta(x, t)$ and $\Delta(x, t)$

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$

Calibrate the scalar thresholds β^*, q^*

Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Need to estimate $\theta(x, t)$ and $\Delta(x, t)$

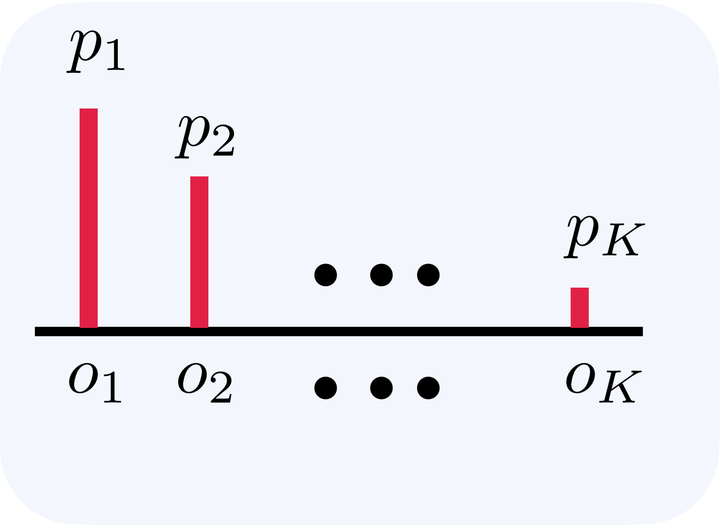
Find the right alternative for $S(x, y) = 1 - p(y \mid x)$

Calibrate the scalar thresholds β^*, q^*

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

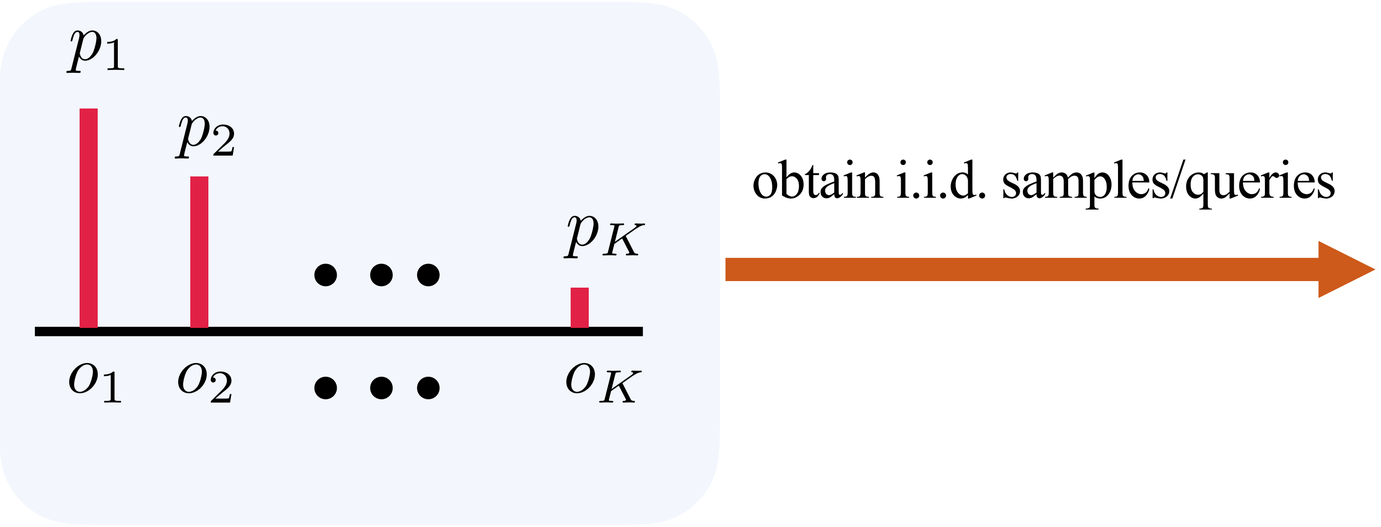
Good Turing Estimator for missing mass



Distribution over objects

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

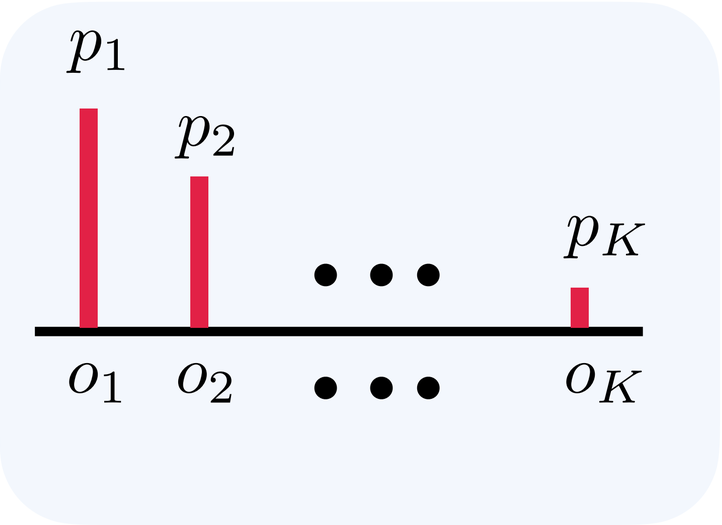
Good Turing Estimator for missing mass



Distribution over objects

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass



Distribution over objects

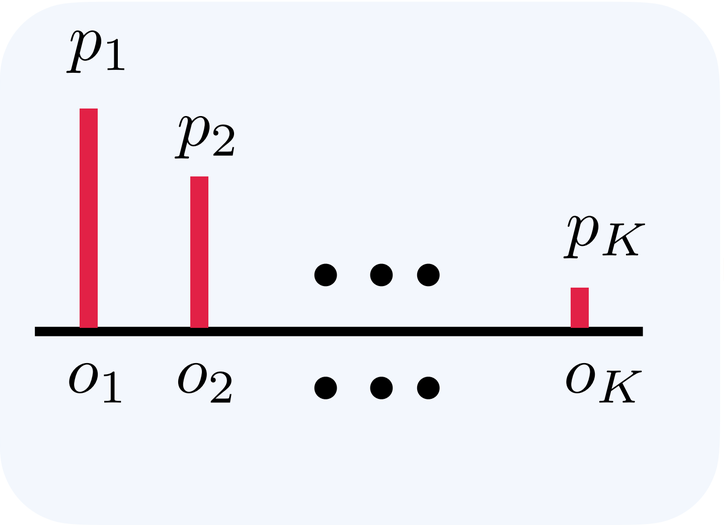
obtain i.i.d. samples/queries

$\{o_1, o_1, o_1, o_3, o_4, o_4, \cdots, o_{K-1}\}$

T samples

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass



Distribution over objects

obtain i.i.d. samples/queries

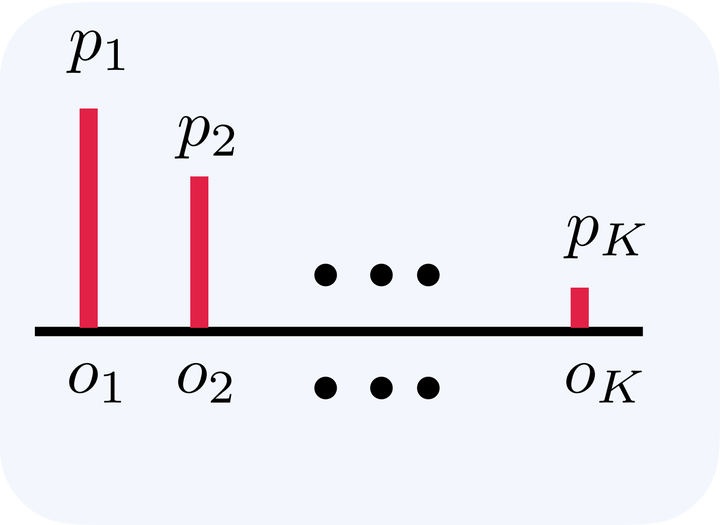
$$\{o_1, o_1, o_1, o_3, o_4, o_4, \cdots, o_{K-1}\}$$

T samples

Some objects missing (e.g o_2)

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass



Distribution over objects

obtain i.i.d. samples/queries

$$\{o_1, o_1, o_1, o_3, o_4, o_4, \cdots, o_{K-1}\}$$

T samples

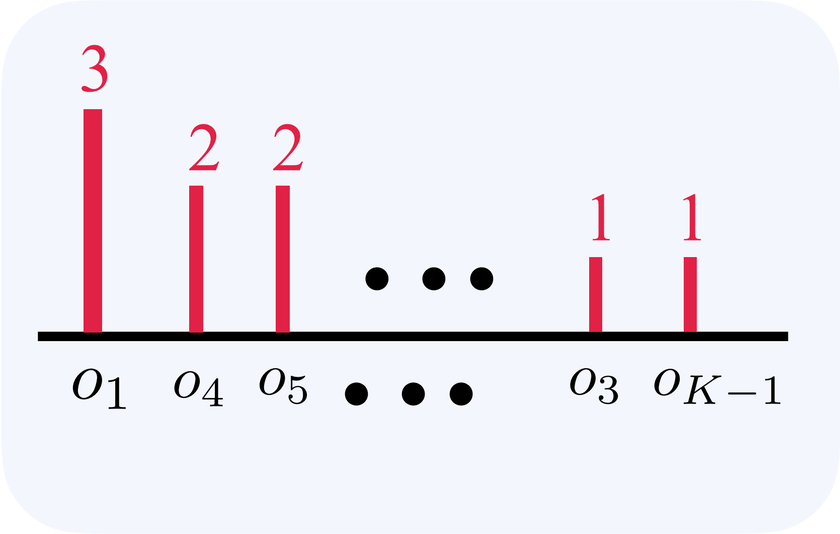
Some objects missing (e.g o_2)

Need to estimate the probability of missing objects

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass

$$\{o_1, o_1, o_1, o_3, o_4, o_4, \cdots, o_{K-1}\}$$



Frequency of observed objects

We have $Z_t(x) = \{y_1^x, \ldots, y_t^x\} \sim \pi(y \mid x)$

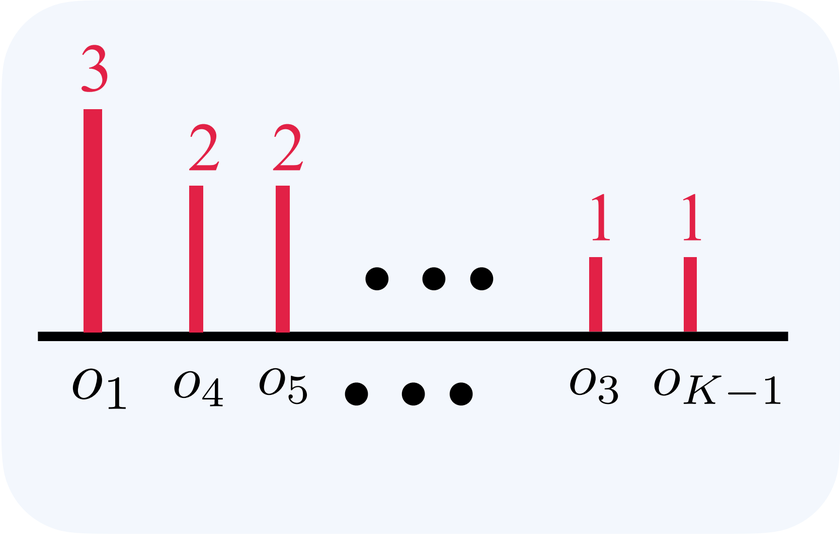
Want to estimate $\theta(x,t) = \mathbb{P}_{Y, Z_t(x)}[Y \notin Z_t(x) \mid X = x]$

N_r : # objects that appeared r times

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass

$$\{o_1, o_1, o_1, o_3, o_4, o_4, \cdots, o_{K-1}\}$$



Frequency of observed objects

We have $Z_t(x) = \{y_1^x, \ldots, y_t^x\} \sim \pi(y \mid x)$

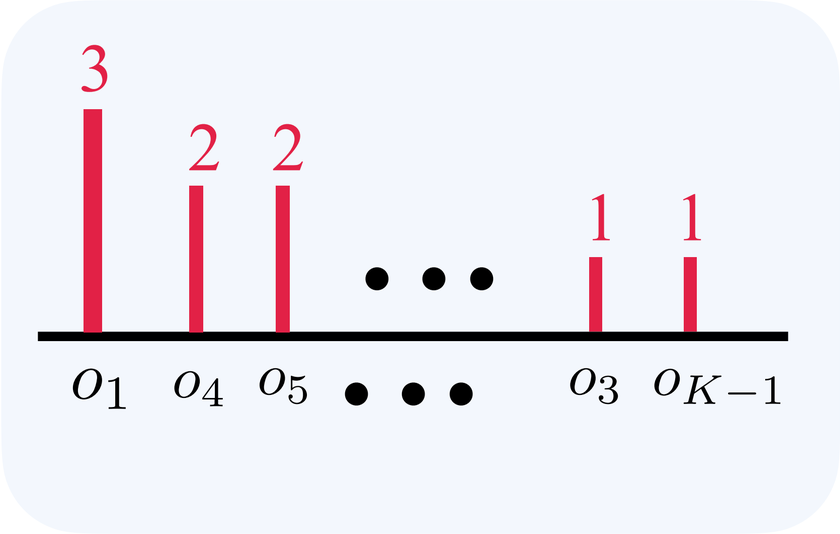
Want to estimate $\theta(x,t) = \mathbb{P}_{Y, Z_t(x)}[Y \notin Z_t(x) \mid X = x]$

$N_r(x,t) = |\{y \in Z_t(x) : \#(y) = r\}|$

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass

$$\{o_1, o_1, o_1, o_3, o_4, o_4, \cdots, o_{K-1}\}$$



Frequency of observed objects

We have $Z_t(x) = \{y_1^x, \ldots, y_t^x\} \sim \pi(y \mid x)$

Want to estimate $\theta(x,t) = \mathbb{P}_{Y, Z_t(x)}[Y \notin Z_t(x) \mid X = x]$

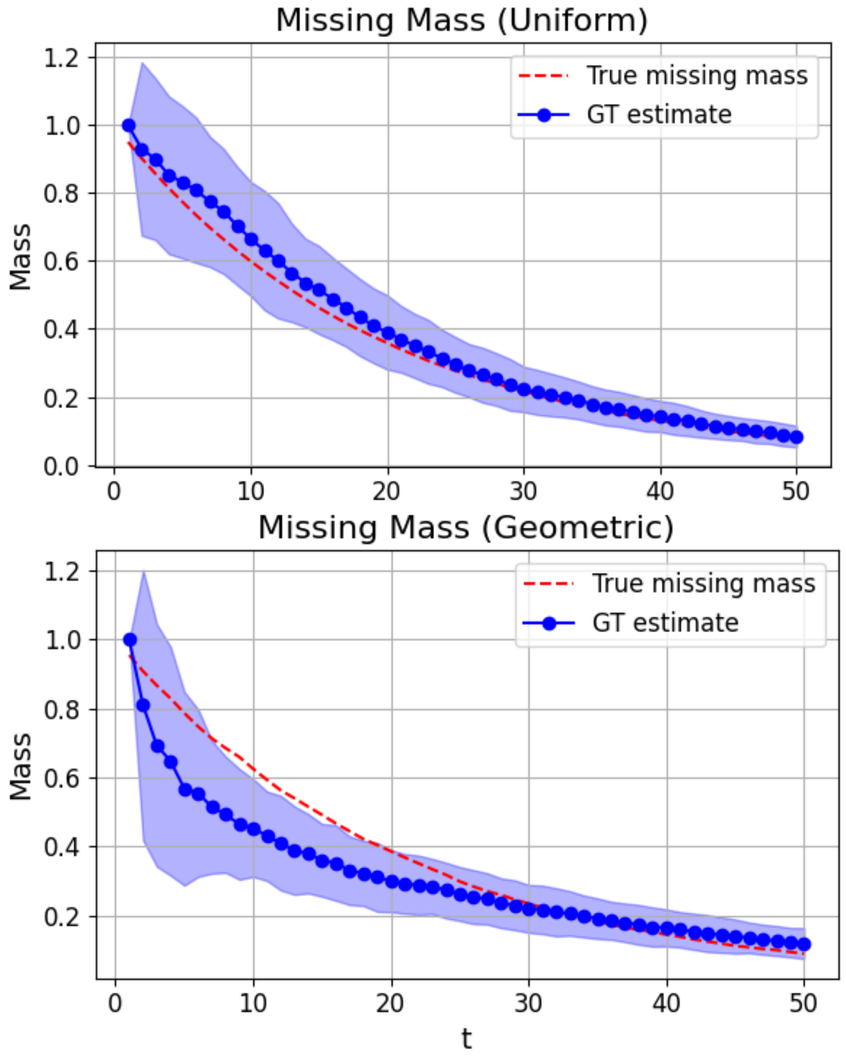
$$N_r(x,t) = |\{y \in Z_t(x) : \#(y) = r\}|$$

Good-Turing estimate for missing mass

$$\hat{\theta}(x,t) = \frac{N_1}{T}$$

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass



We have $Z_t(x) = \{y_1^x, \ldots, y_t^x\} \sim \pi(y \mid x)$

Want to estimate $\theta(x,t) = \mathbb{P}_{Y,Z_t(x)}[Y \notin Z_t(x) \mid X = x]$

$N_r(x,t) = |\{y \in Z_t(x) : \#(y) = r\}|$

Good-Turing estimate for missing mass

$$\hat{\theta}(x,t) = \frac{N_1}{T}$$

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass

$$N_r(x,t) = |\{y \in Z_t(x) : \#(y) = r\}|$$

Good-Turing estimate for missing mass

$$\theta(x,t) = \mathbb{P}_{Y,Z_t(x)}[Y \notin Z_t(x) \mid X = x]$$

$$\hat{\theta}(x,t) = \frac{N_1}{T}$$

Need to estimate $\theta(x,t)$ and $\Delta(x,t)$

Good Turing Estimator for missing mass

$$N_r(x,t) = |\{y \in Z_t(x) : \#(y) = r\}|$$

Good-Turing estimate for missing mass

$$\theta(x,t) = \mathbb{P}_{Y,Z_t(x)}[Y \notin Z_t(x) \mid X = x]$$

$$\hat{\theta}(x,t) = \frac{N_1}{T}$$

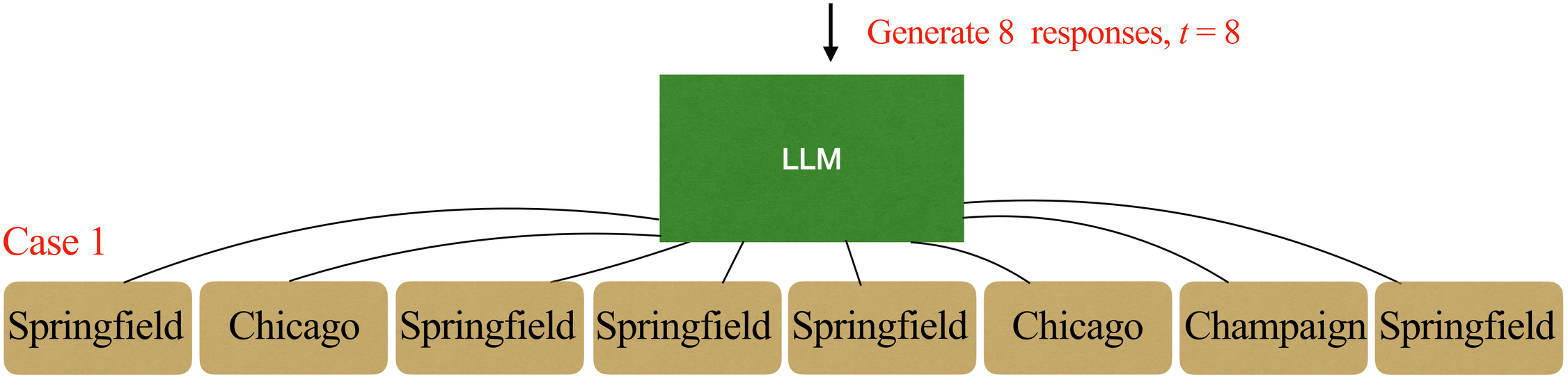
Our estimate for missing mass derivative

$$\Delta(x,t) := \theta(x,t+1) - \theta(x,t)$$

$$\hat{\Delta}(x,t) = \frac{-2N_2}{T^2}$$

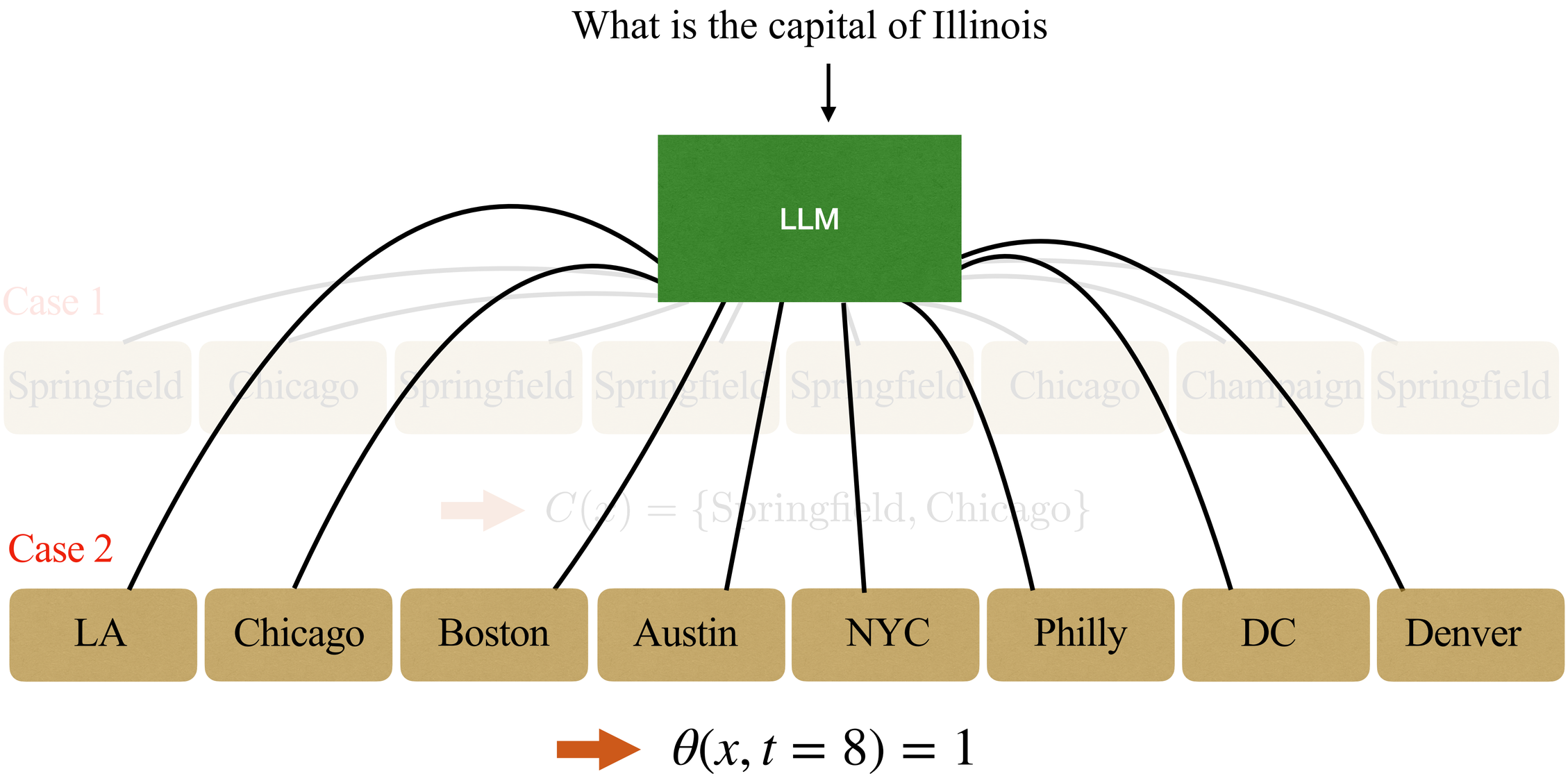
“Toy” Running Example

What is the capital of Illinois



$$\theta(x, t = 8) = \frac{1}{8}$$

“Toy” Running Example



Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Need to estimate $\theta(x, t)$ and $\Delta(x, t)$ ✓

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$

Calibrate the scalar thresholds β^*

Calibrate the scalar thresholds q^*

Calibrate the scalar thresholds β^*

Calibrate the scalar thresholds β^*

Finite sample version of principle 1

$$\Delta(x, T^*(x)) = \beta^*$$

Calibrate the scalar thresholds β^*

Finite sample version of principle 1

$$\Delta(x, T^*(x)) = \beta^*$$

For each $x \in D_{cal}$

Sample $y_{1:T(x)} \sim \pi(y \mid x)$ until $\hat{\Delta}(x, T(x)) \leq \beta^*$

Calibrate the scalar thresholds β^*

Finite sample version of principle 1

$$\Delta(x, T^*(x)) = \beta^*$$

For each $x \in D_{cal}$

Sample $y_{1:T(x)} \sim \pi(y \mid x)$ until $\hat{\Delta}(x, T(x)) \leq \beta^*$

Obtain $Z(x) = \{y_1, \ldots, y_{T(x)}\}$

Calibrate the scalar thresholds β^*

Finite sample version of principle 1

$$\Delta(x, T^*(x)) = \beta^*$$

For each $x \in D_{cal}$

Sample $y_{1:T(x)} \sim \pi(y \mid x)$ until $\hat{\Delta}(x, T(x)) \leq \beta^*$

Obtain $Z(x) = \{y_1, \ldots, y_{T(x)}\}$

Grid search

Make sure $\frac{1}{|D_{cal}|} \sum_x T(x) \leq B$

Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Need to estimate $\theta(x, t)$ and $\Delta(x, t)$ ✓

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$

Calibrate the scalar thresholds β^* ✓

Calibrate the scalar thresholds q^*

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$ and $\Delta(x, t)$

Finite sample version of principle 2

$$S(x, y) = \begin{cases} 1 - p(y \mid x), & if \quad y \neq EE \\ 2 - p(y \mid x), & if \quad y = EE \end{cases}$$

$$Z(x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$ and $\Delta(x, t)$

Finite sample version of principle 2

$$S(x, y) = \begin{cases} 1 - p(y \mid x), & if \quad y \neq EE \\ 2 - p(y \mid x), & if \quad y = EE \end{cases}$$

$$Z(x) = \{y_1^x, \ldots, y_{T(x)}^x\}$$

Using the Good Turing estimator we can estimate:

$$\hat{p}(y \mid x)$$

$$\hat{p}(EE \mid x) = \theta(x, T(x))$$

Our Main Algorithm (In the finite-sample regime)

We don't have the true $p(y \mid x)$

But we can query $\pi(y \mid x)$

Need to estimate $\theta(x, t)$ and $\Delta(x, t)$ ✓

Find the right alternative for $S(x, y) = 1 - p(y \mid x)$ ✓

Calibrate the scalar thresholds β^* ✓

Calibrate the scalar thresholds q^*

Finite Sample Algorithm

Algorithm 1 Conformal Prediction with Query Oracle (CPQ)

Input: Query oracle $\pi(y \mid x)$, conformity score $\hat{S}(x, y)$, calibration data $\mathcal{D}_{\text{cal}_2}$, test point x_{test}, miscoverage α, query budget B, missing-mass estimator $\hat{\Delta}(x, t)$, threshold β^*

Query Module → Principle 1

- For each $x \in \mathcal{D}_{\text{cal}_2} \cup \{x_{\text{test}}\}$:
  - Sample $y_{1:T(x)} \sim \pi(y \mid x)$ until $\hat{\Delta}(x, T(x)) \leq \beta^*$. Let $Z(x) = \{y_1, \ldots, y_{T(x)}\}$.

Calibration Module → Principle 2

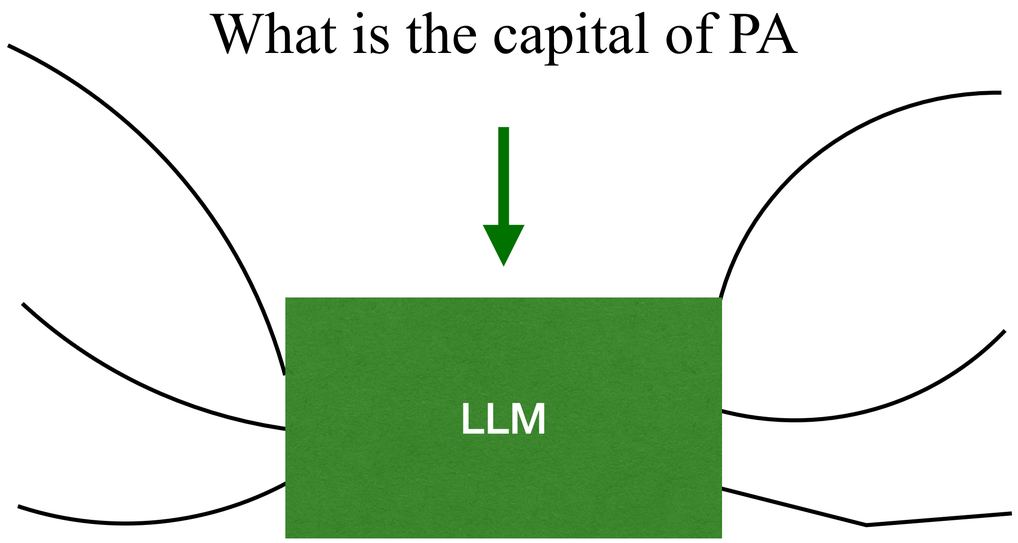
- For each $(x_i, y_i) \in \mathcal{D}_{\text{cal}_2}$ compute $s_i = \hat{S}(x_i, y_i)$.
- Set $q^* = \text{Quantile}_{1-\alpha}\big(s_1, \ldots, s_{|\mathcal{D}_{\text{cal}_2}|}, \infty\big)$.

Output: $C(x_{\text{test}}) = \big\{\, y \in Z(x_{\text{test}}) \cup \{\texttt{EE}\} : \hat{S}(x_{\text{test}}, y) \leq q^* \,\big\}$.

Let’s apply our algorithm to LLMs

Let’s apply our algorithm to LLMs

Clustering outputs



Clusters

Fine grained component wise comparison

We consider three methods:

Vanilla - fixed non-adaptive querying, valid but sub-optimal calibration

P1 - adds adaptive querying

P1 + P2 - optimal querying and optimal calibration -> CPQ

Fine grained component wise comparison

We consider three methods:

Vanilla - fixed non-adaptive querying, valid but sub-optimal calibration

P1 - adds adaptive querying

P1 + P2 - optimal querying and optimal calibration -> CPQ

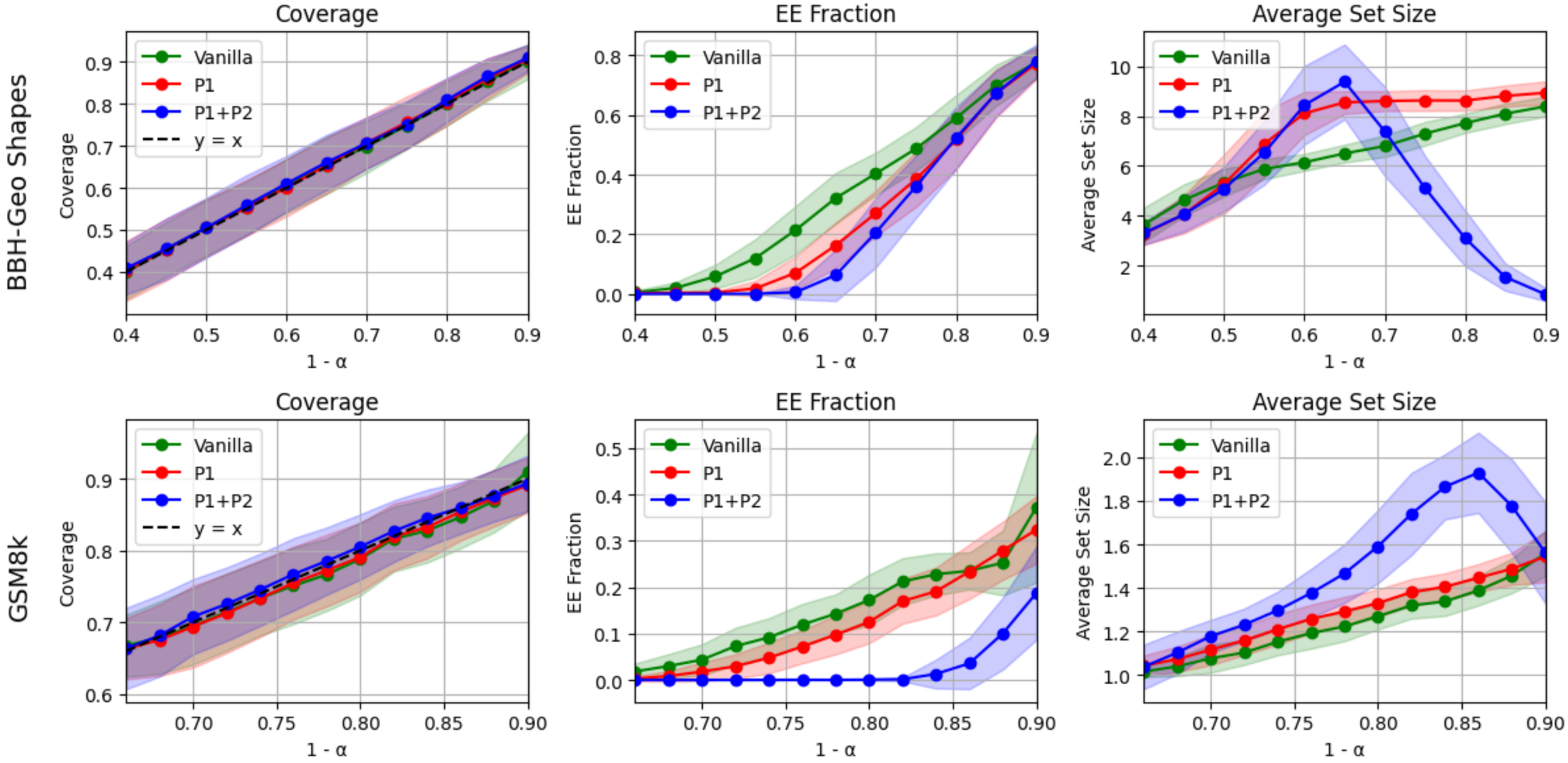
And three evaluation metrics:

Empirical Coverage

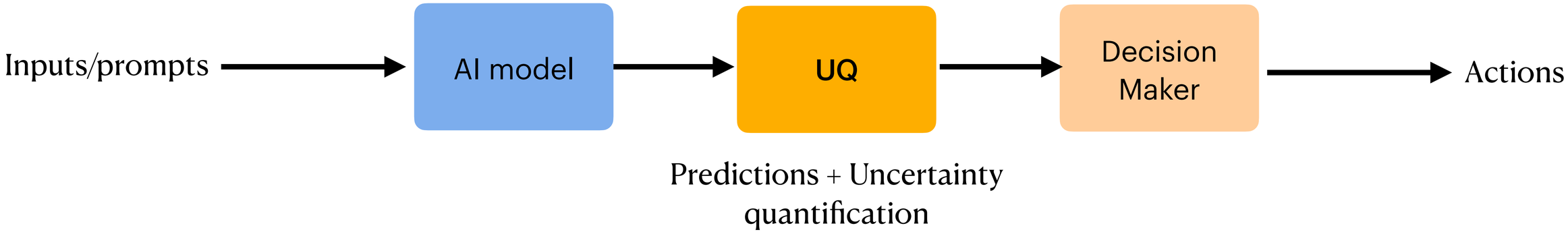
EE fraction

Prediction set size

Fine grained component wise comparison



AI-powered Decision making pipeline



Part I: CP for Generative Models

Conformal Prediction Beyond the Seen: A Missing Mass Perspective for Uncertainty Quantification in Generative Models

Sima Noorani[*1], Shayan Kiyani[*1], George Pappas[1], and Hamed Hassani[1]

[1]University of Pennsylvania

Part II: Collaborative CP

Human AI Collaborative Uncertainty Quantification

Sima Noorani[*1], Shayan Kiyani[*1], George Pappas[1], and Hamed Hassani[1]

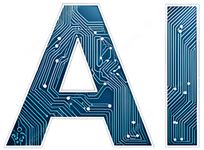
[1]University of Pennsylvania

Arxiv - Oct '25

UQ when humans and AI are jointly in the loop?

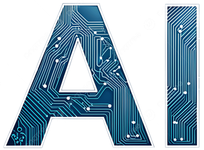
Human and AI should coexist in the decision making pipeline

Human and AI should coexist in the decision making pipeline



Handle large unstructured data

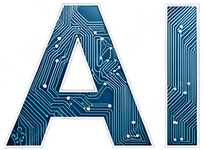
Human and AI should coexist in the decision making pipeline



Handle large unstructured data

Excel at patten extraction

Human and AI should coexist in the decision making pipeline



Handle large unstructured data

Excel at patten extraction

Offer statistical accuracy

Human and AI should coexist in the decision making pipeline

AI

Handle large unstructured data

Excel at patten extraction

Offer statistical accuracy

Domain Knowledge

Human and AI should coexist in the decision making pipeline

AI

- Handle large unstructured data
- Excel at patten extraction
- Offer statistical accuracy

- Domain Knowledge
- Persistent Memory

Human and AI should coexist in the decision making pipeline

AI

- Handle large unstructured data
- Excel at patten extraction
- Offer statistical accuracy

- Domain Knowledge
- Persistent Memory
- Reason and act in the physical world

Human and AI should coexist in the decision making pipeline

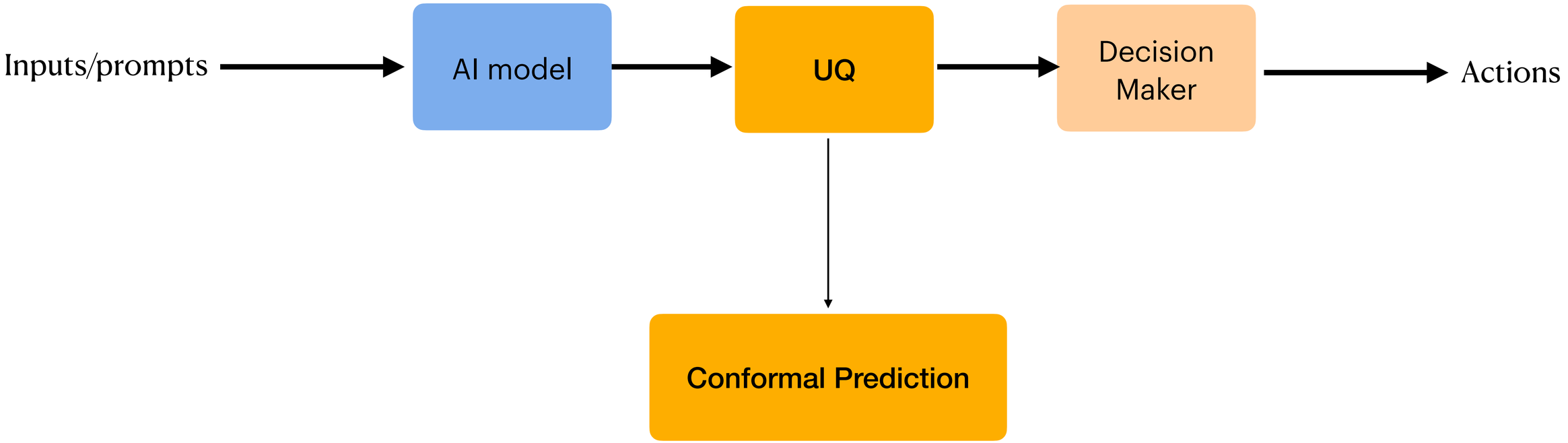
AI

- Handle large unstructured data
- Excel at patten extraction
- Offer statistical accuracy

- Domain Knowledge
- Persistent Memory
- Reason and act in the physical world

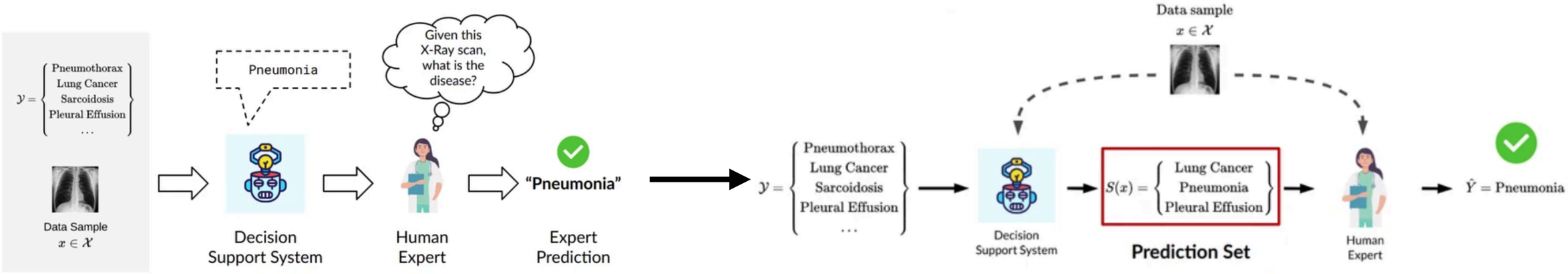
Human and AI should jointly exist in the decision making process!

AI-powered Decision making pipeline



Prediction sets are useful for decision makers

Prediction sets are useful for decision makers

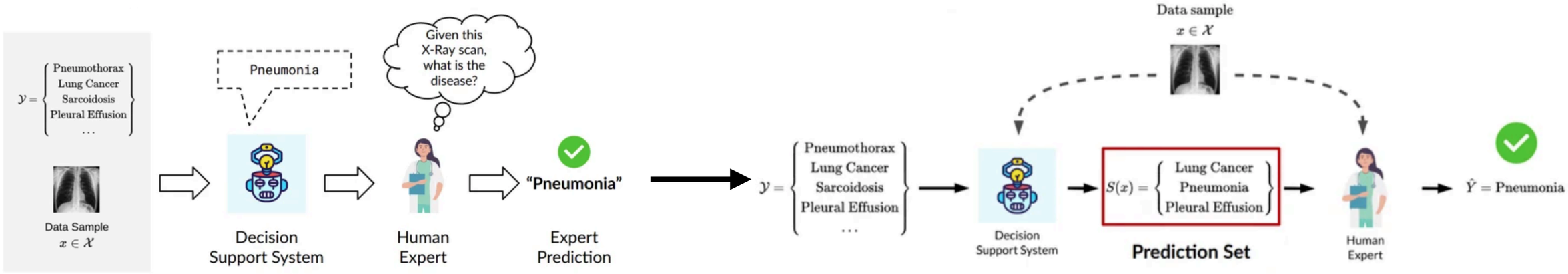


Straitouri et al., Improving Expert Predictions with Conformal Prediction, ICML, 2023.

E. Straitouri & M. Gomez-Rodriguez, *Designing Decision Support Systems using Counterfactual Prediction Sets*, ICML, 2024.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. Towards human-AI complementarity with prediction sets. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

Prediction sets are useful for decision makers



Straitouri et al., Improving Expert Predictions with Conformal Prediction, ICML, 2023.

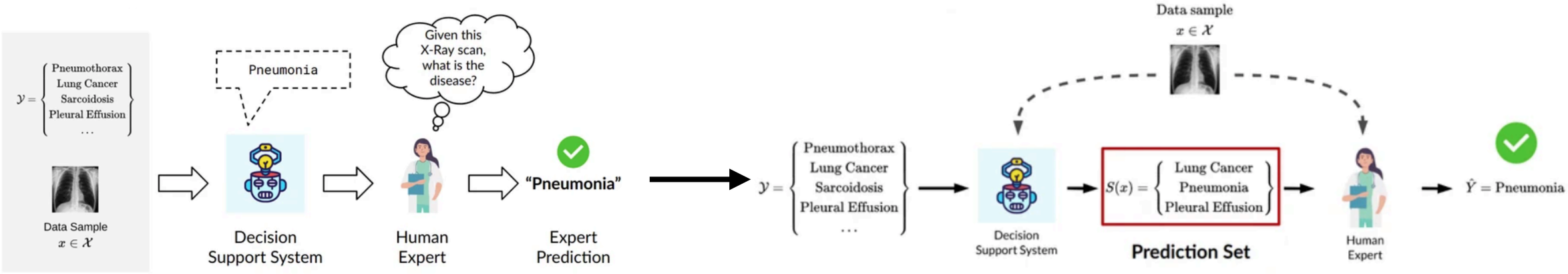
E. Straitouri & M. Gomez-Rodriguez, *Designing Decision Support Systems using Counterfactual Prediction Sets*, ICML, 2024.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. Towards human-AI complementarity with prediction sets. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

"Decision-theoretic framework for evaluating predictive uncertainty as informative signals"

Hullman et. al.., *Conformal Prediction and Human Decision Making*, 2025.

Prediction sets are useful for decision makers



Straitouri et al., Improving Expert Predictions with Conformal Prediction, ICML, 2023.

E. Straitouri & M. Gomez-Rodriguez, *Designing Decision Support Systems using Counterfactual Prediction Sets*, ICML, 2024.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. Towards human-AI complementarity with prediction sets. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

"Decision-theoretic framework for evaluating predictive uncertainty as informative signals"

Hullman et. al.., *Conformal Prediction and Human Decision Making*, 2025.

"Well designed" prediction sets are a sufficient statistic for risk averse decision making

S. Kiyani et Al., *Decision Theoretic Foundations for Conformal Prediction: Optimal Uncertainty Quantification for Risk-Averse Agents*, ICML, 2025.

Thus far …

Thus far …

Human and AI should coexist in the decision making pipeline

Thus far …

✓ **Human and AI should coexist in the decision making pipeline**

✓ **UQ is essential in the decision making pipeline**

UQ

Thus far …

- ✓ **Human and AI should coexist in the decision making pipeline**

- ✓ **UQ is essential in the decision making pipeline**

UQ

- ✓ **CP is a promising tool for UQ of AI decision support systems**

Thus far …

- ✓ **Human and AI should coexist in the decision making pipeline**

- ✓ **UQ is essential in the decision making pipeline**

UQ

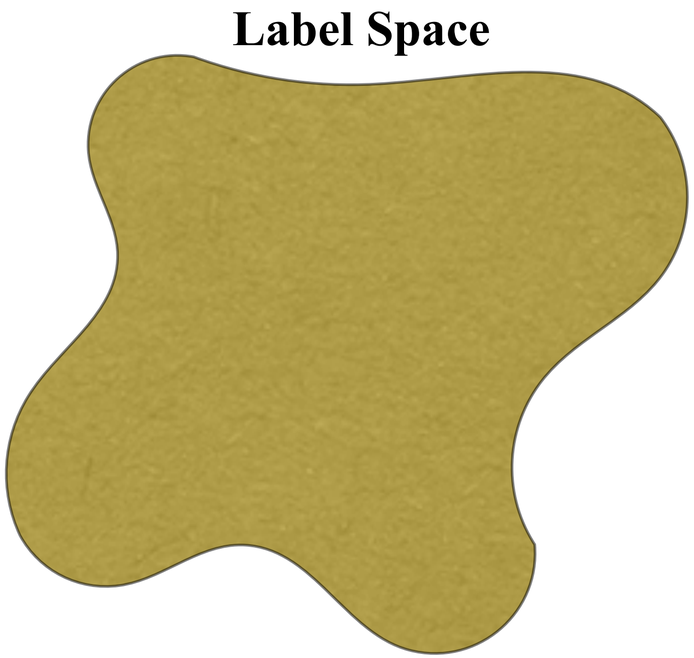
- ✓ **CP is a promising tool for UQ of AI decision support systems**

What should be the principles of UQ when Human and AI are jointly in the loop?

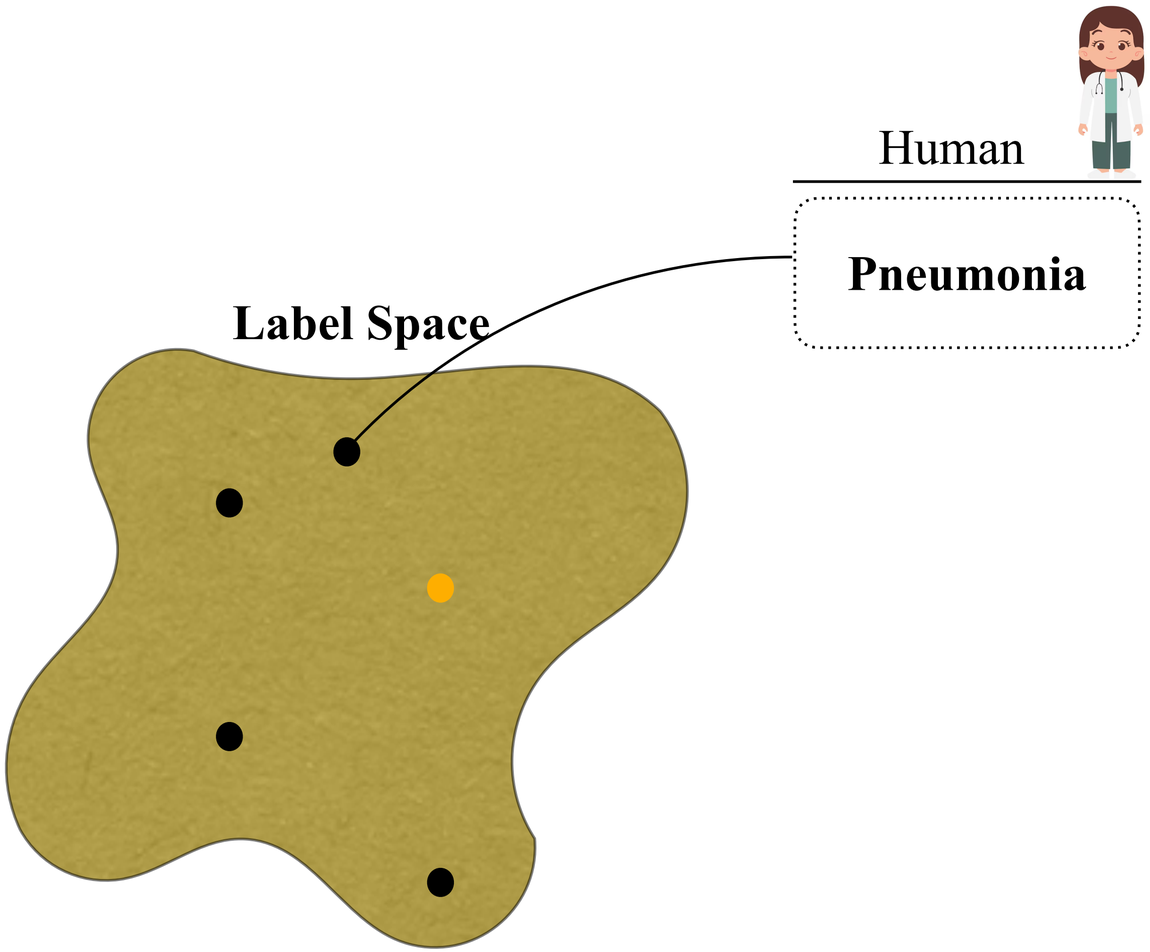
Question: what constitutes a good collaboration?

Question: what constitutes a good collaboration?

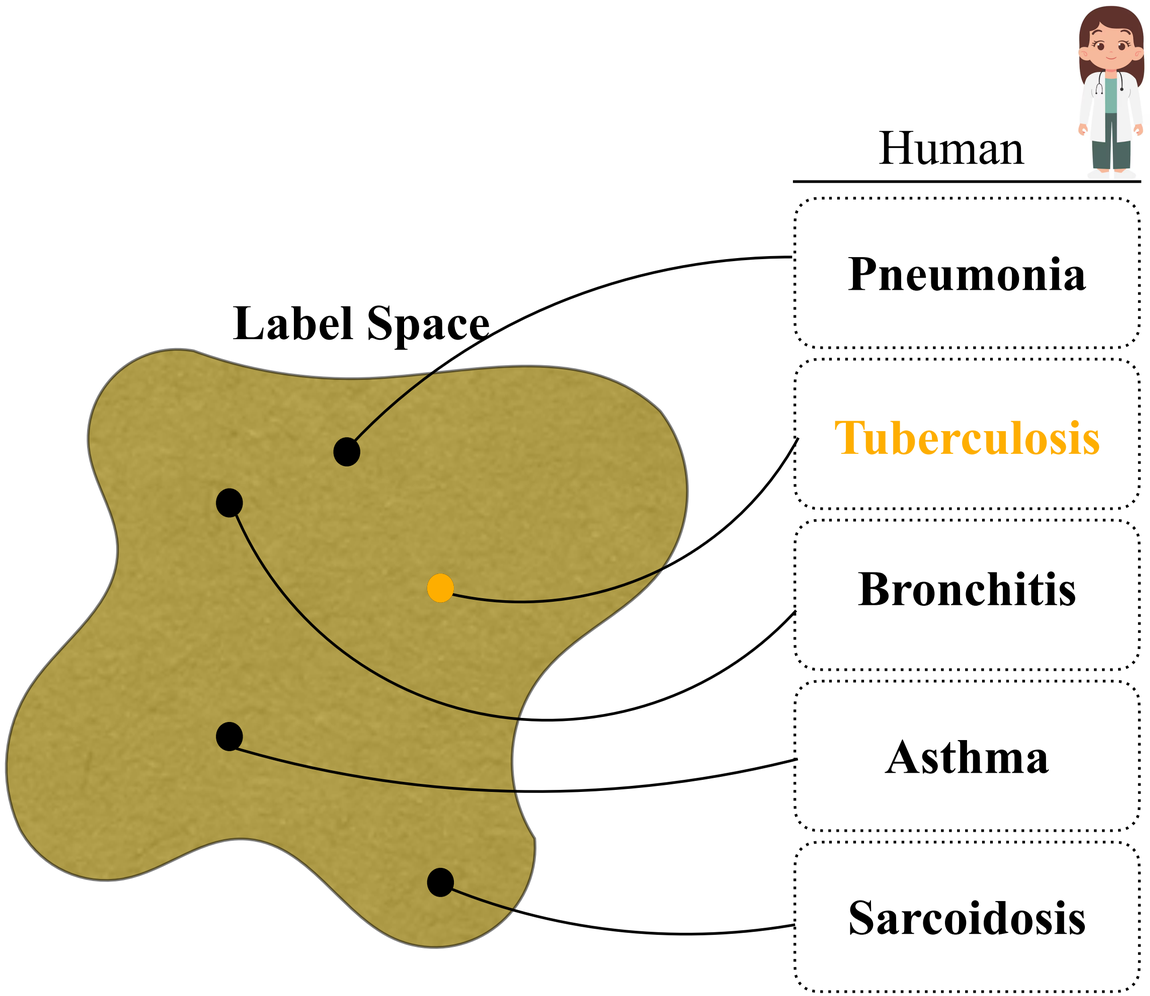
Question: what constitutes a good collaboration?



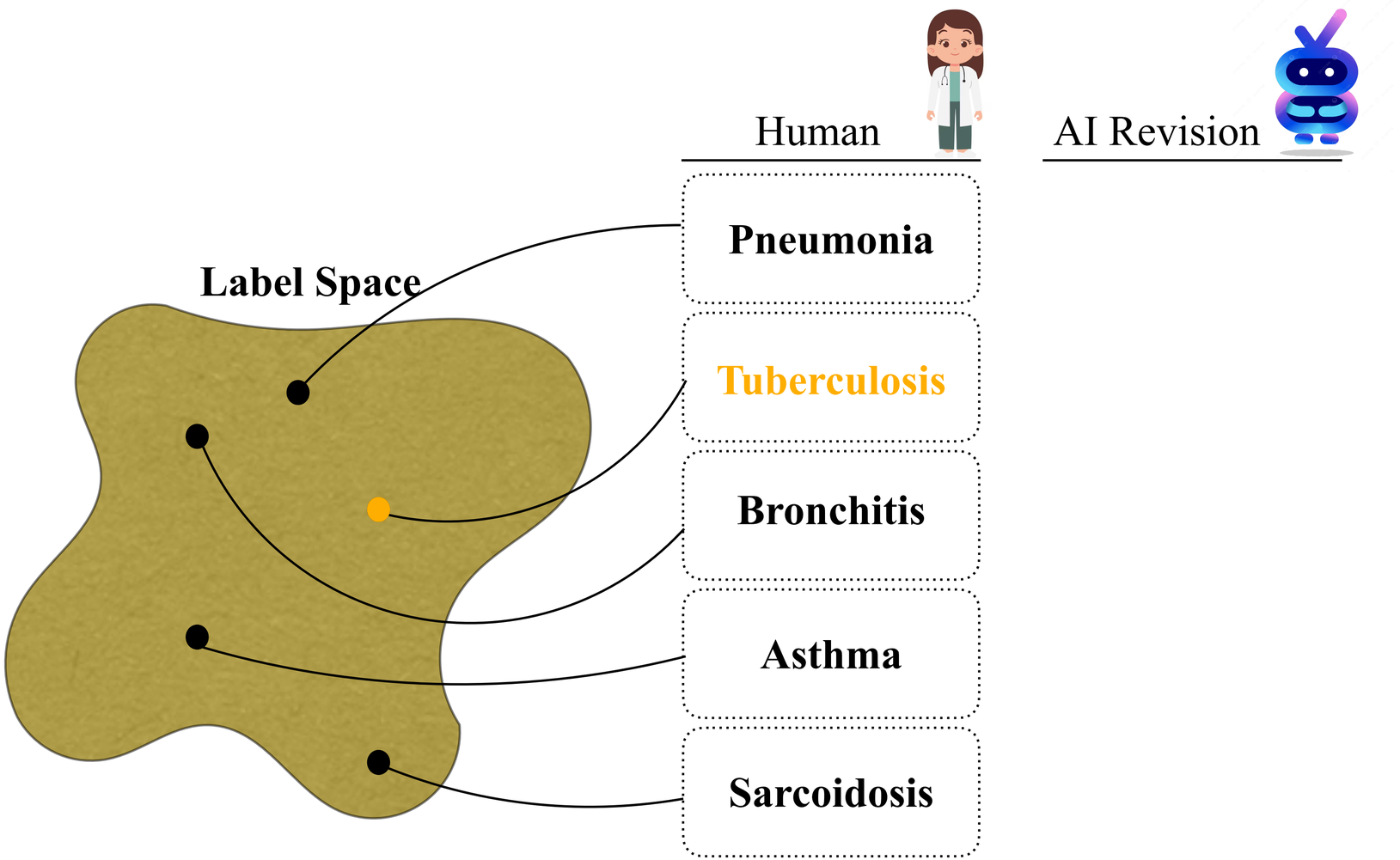
Question: what constitutes a good collaboration?



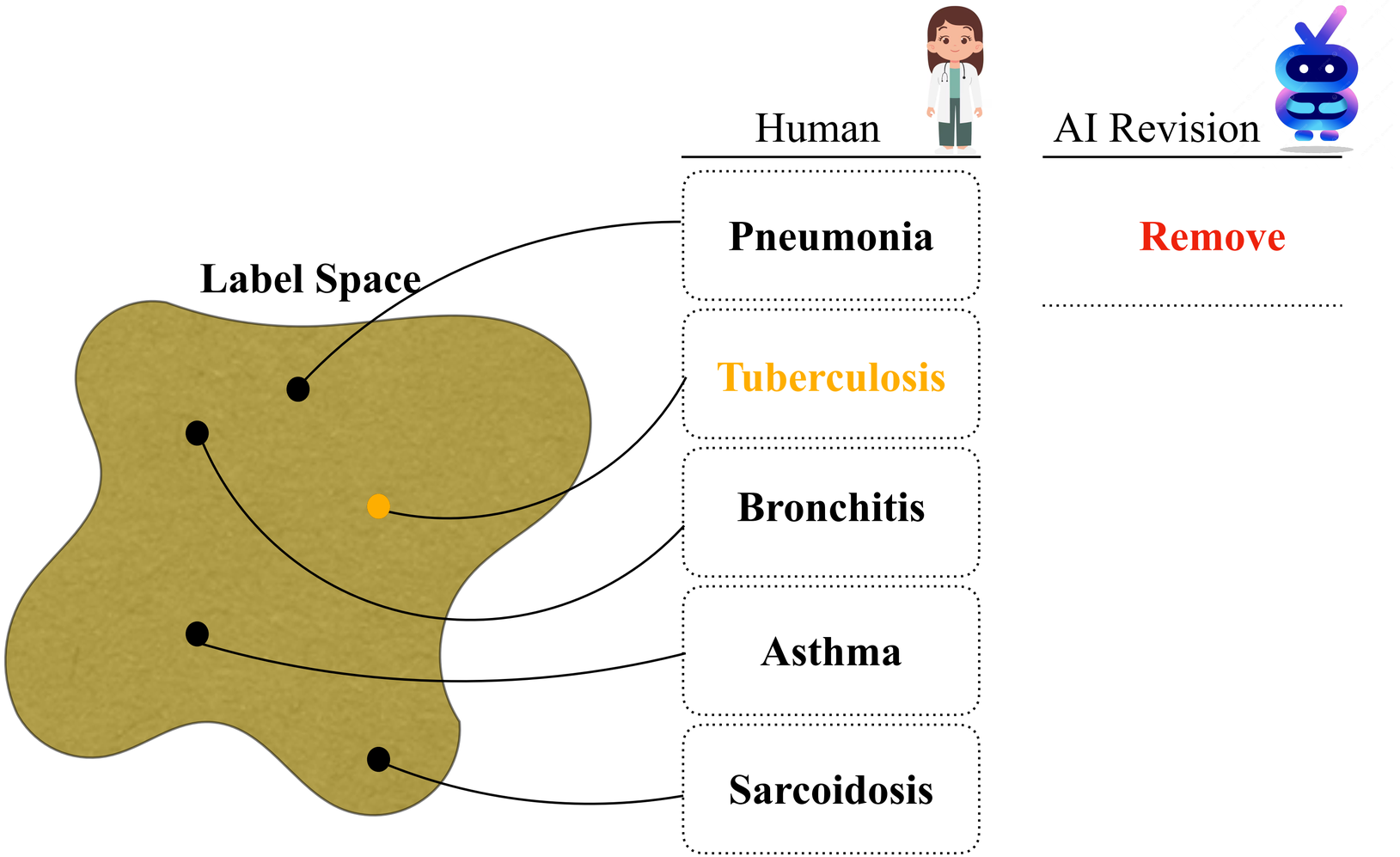
Question: what constitutes a good collaboration?



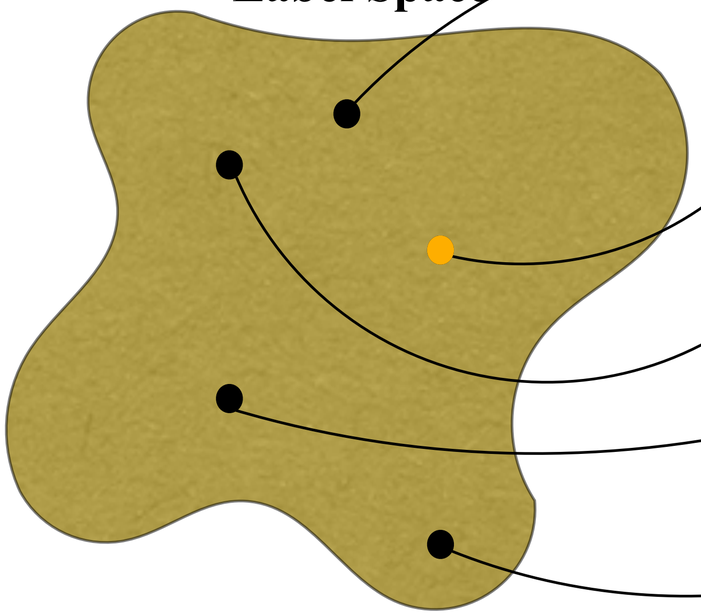
Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?

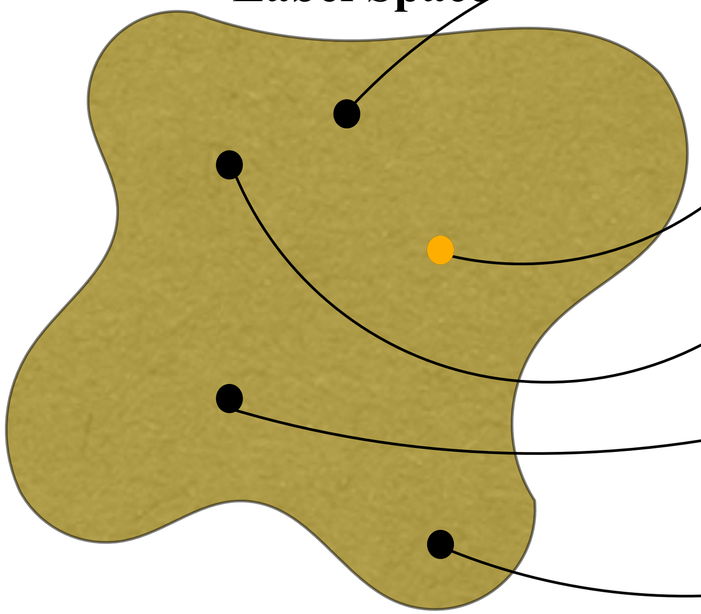


Question: what constitutes a good collaboration?



Human	AI Revision
Pneumonia	Remove
Tuberculosis	Keep
Bronchitis	Remove
Asthma	Keep
Sarcoidosis	Remove

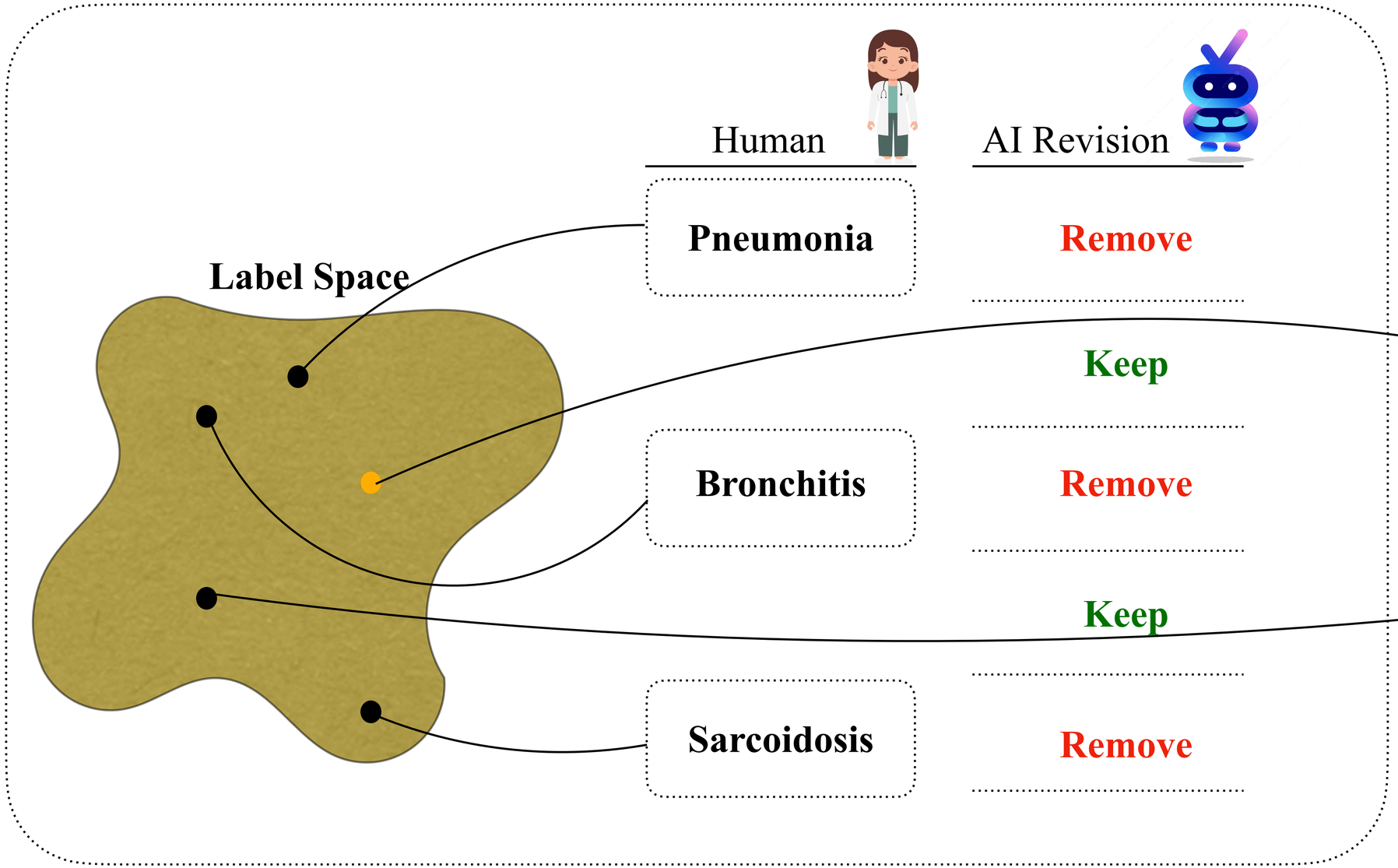
Question: what constitutes a good collaboration?



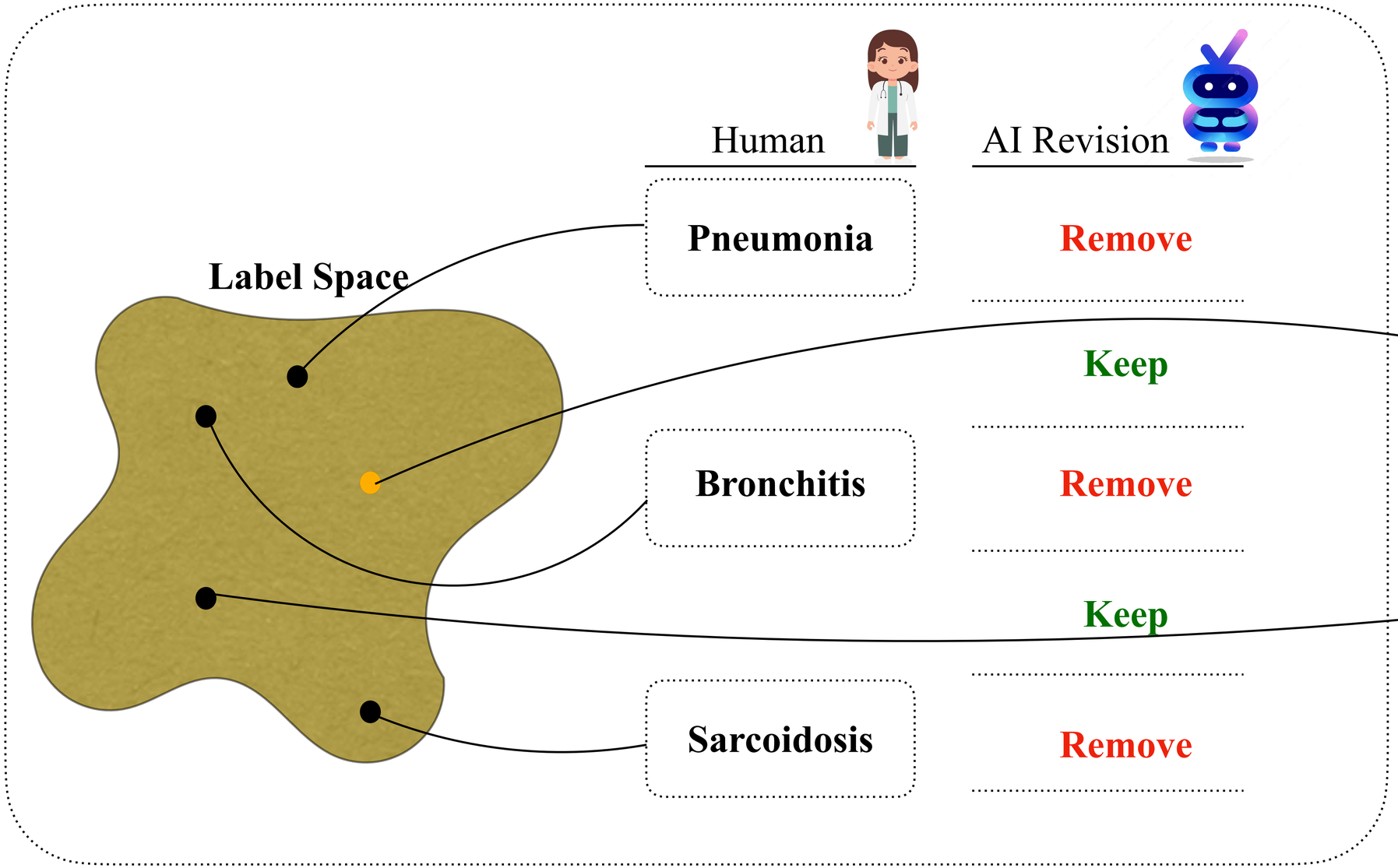
Human	AI Revision
Pneumonia	Remove
Tuberculosis	Keep
Bronchitis	Remove
Asthma	Keep
Sarcoidosis	Remove

Collaborative Proposal

Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?

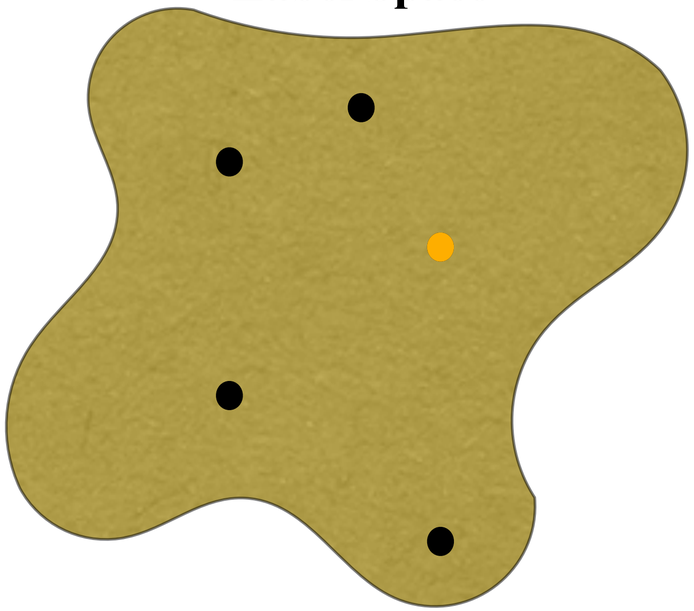


First: dont cause harm!

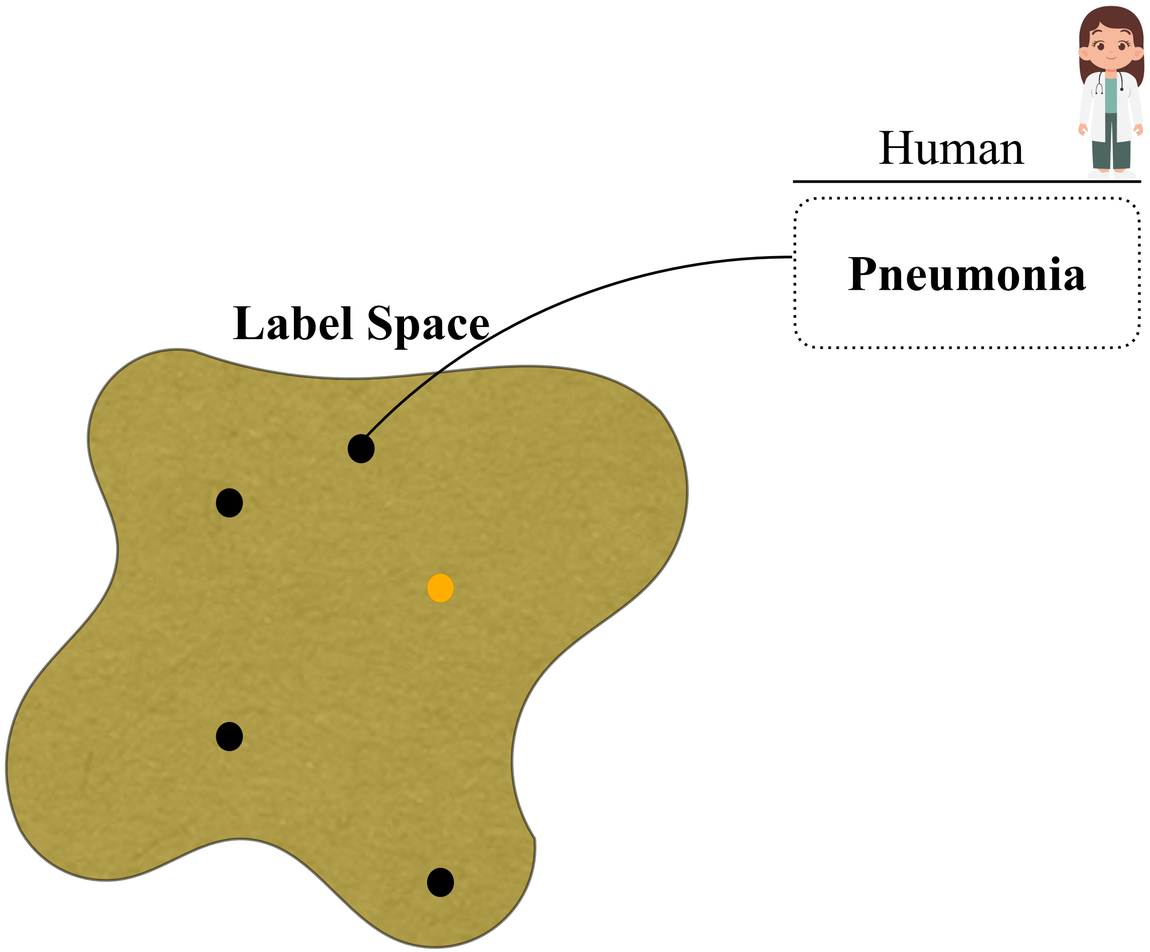
Question: what constitutes a good collaboration?

Human 

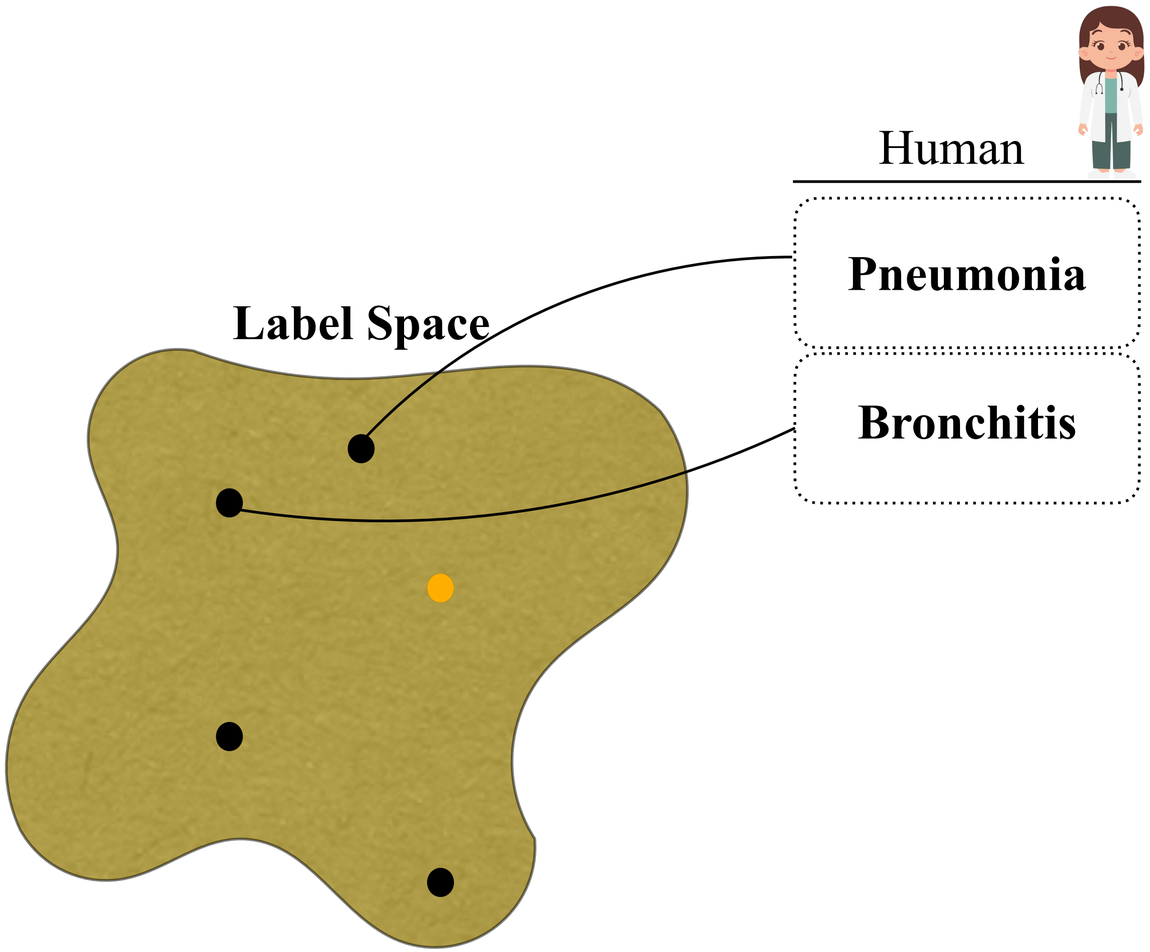
Label Space



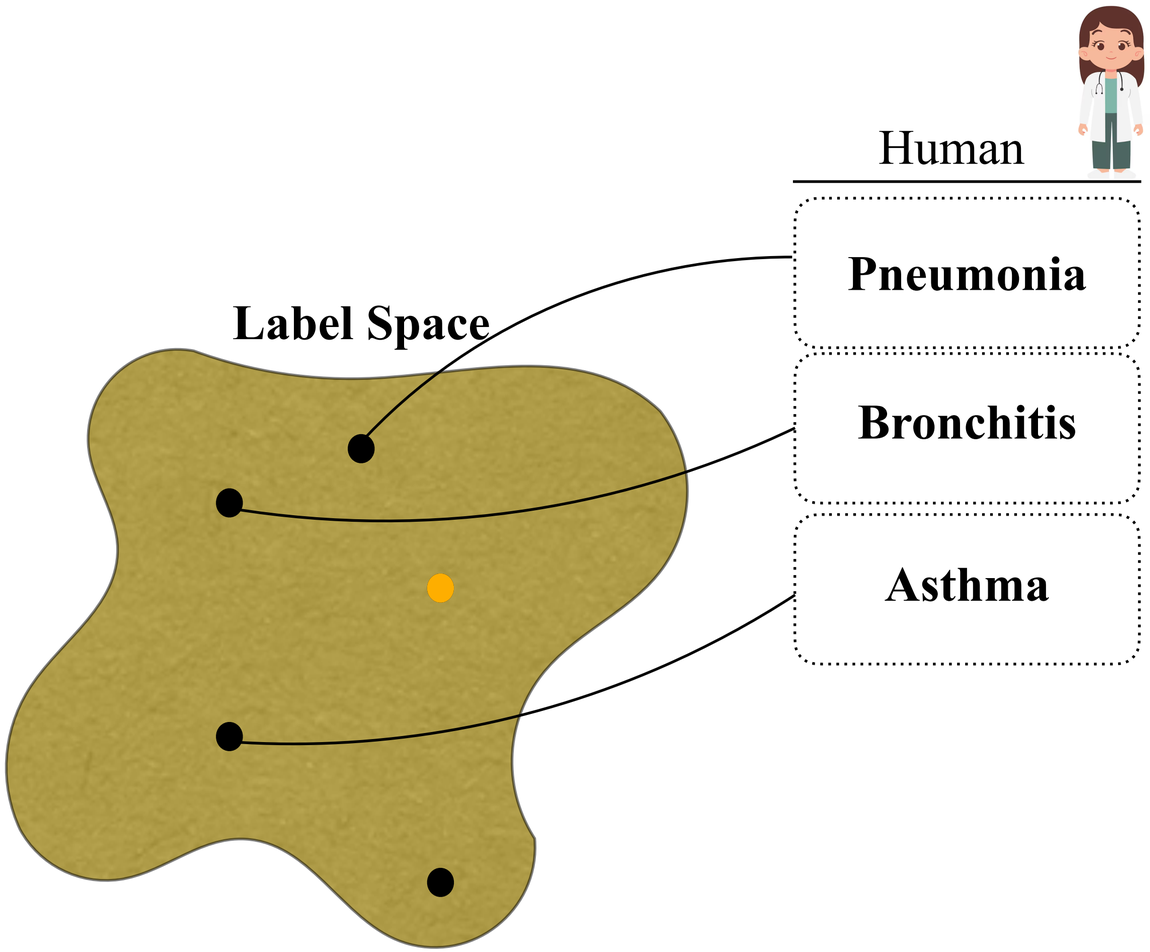
Question: what constitutes a good collaboration?



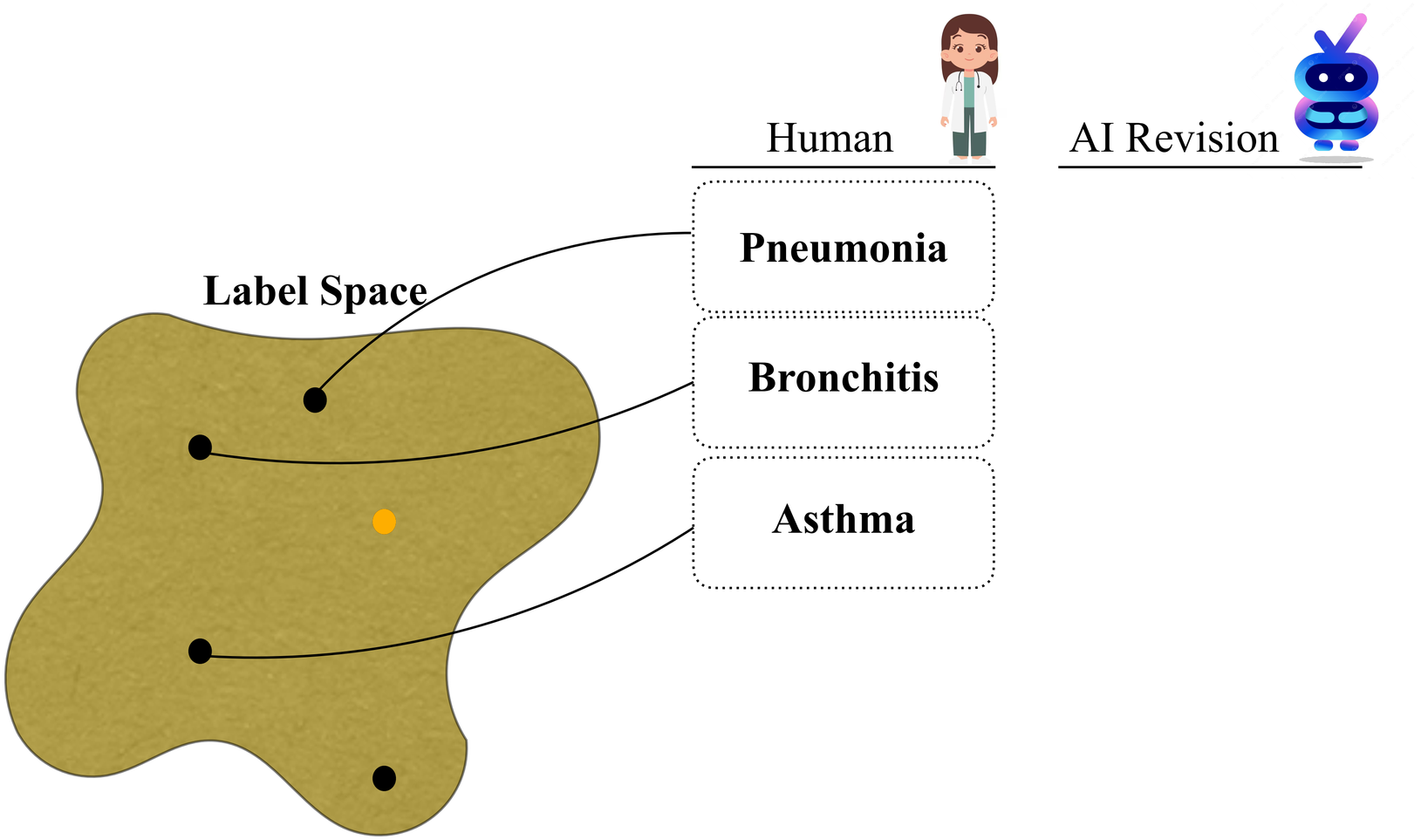
Question: what constitutes a good collaboration?



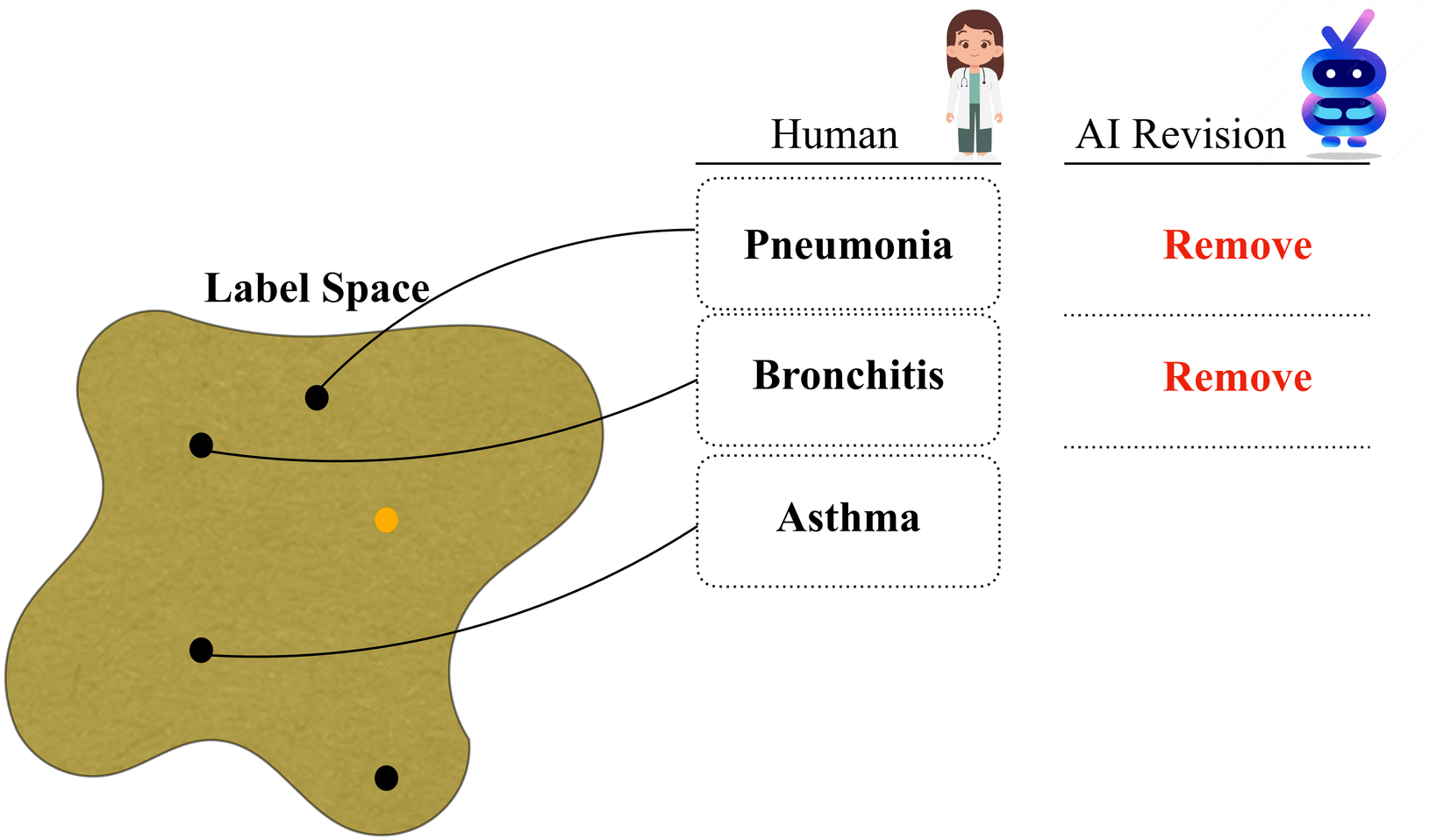
Question: what constitutes a good collaboration?



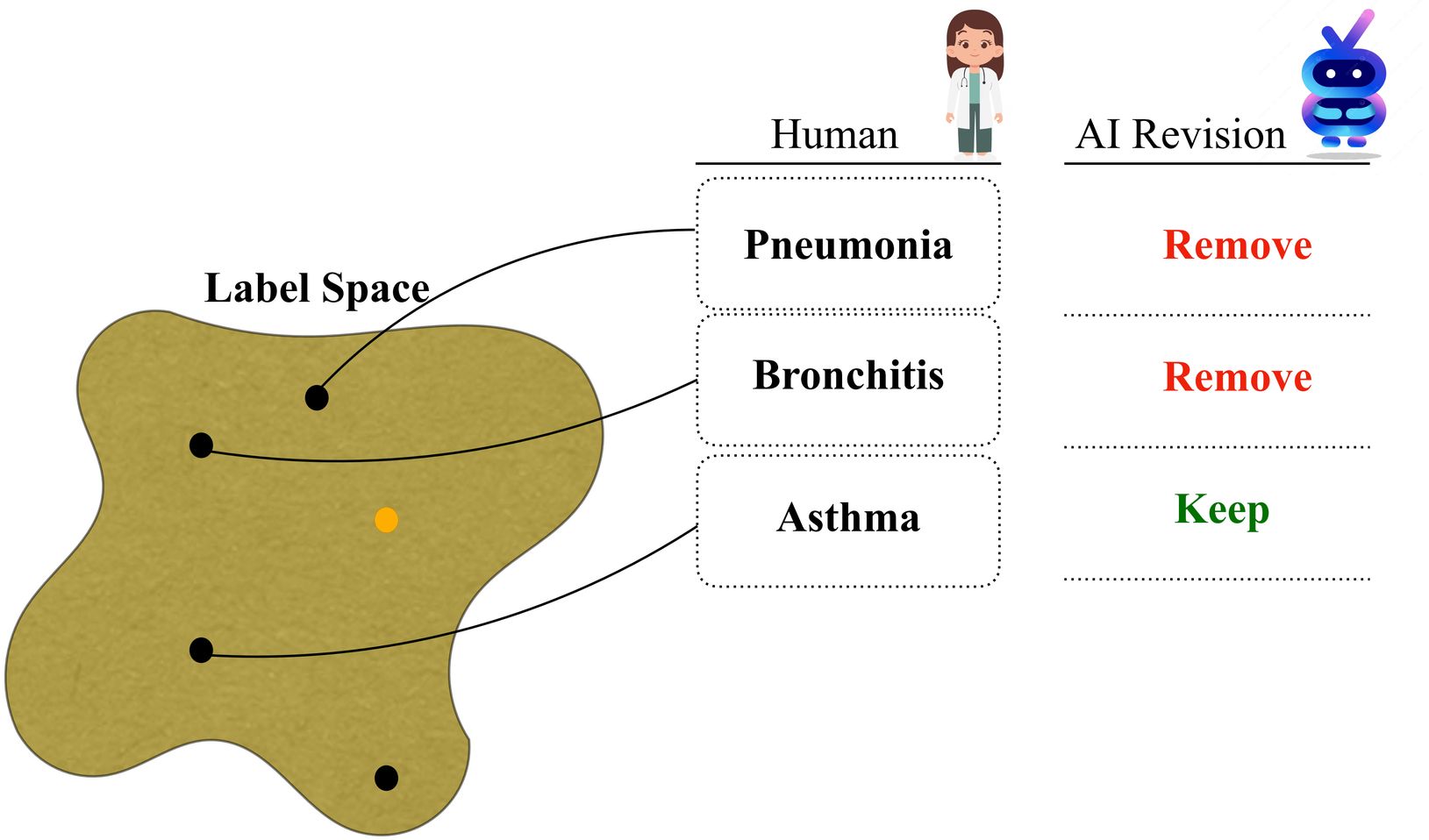
Question: what constitutes a good collaboration?



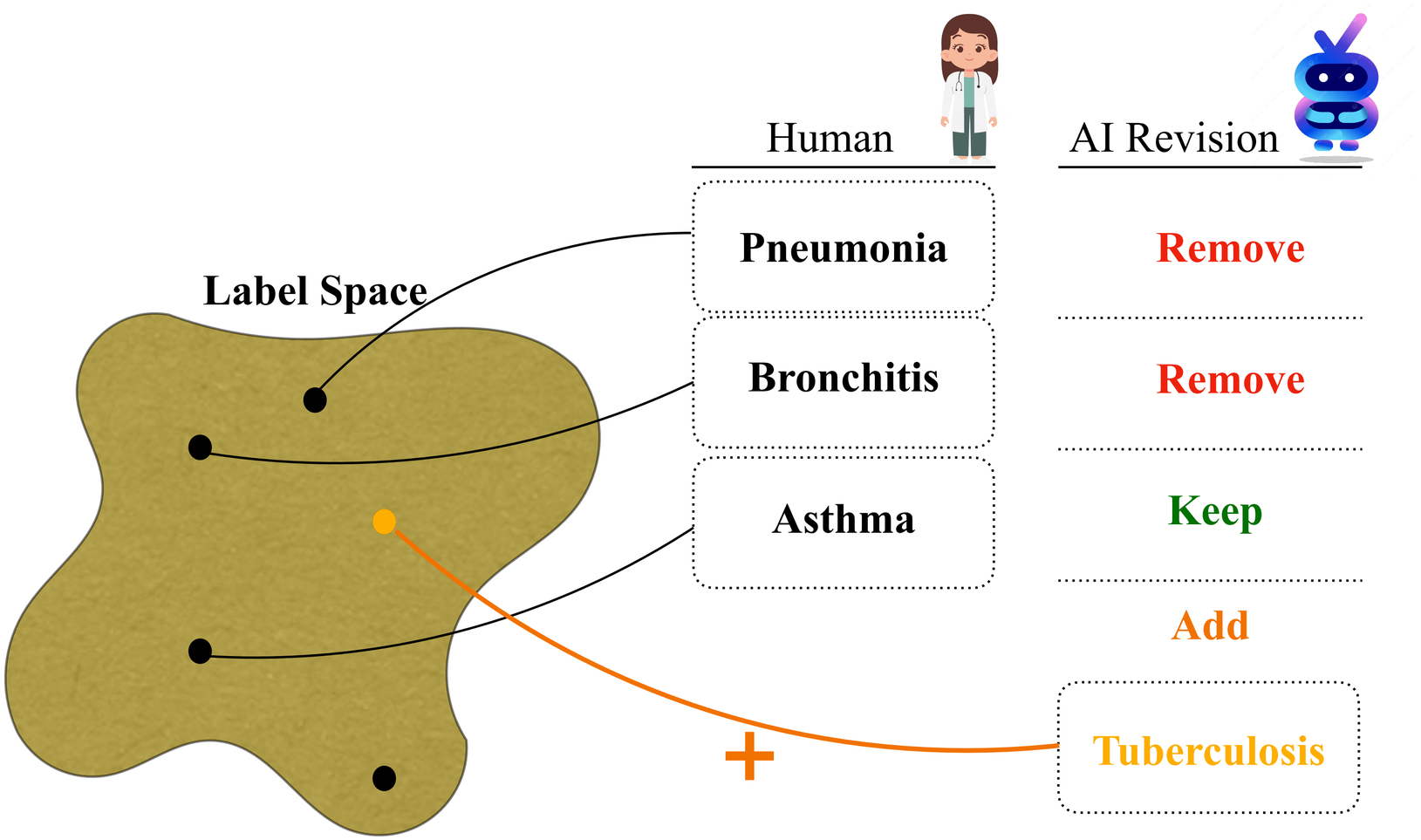
Question: what constitutes a good collaboration?



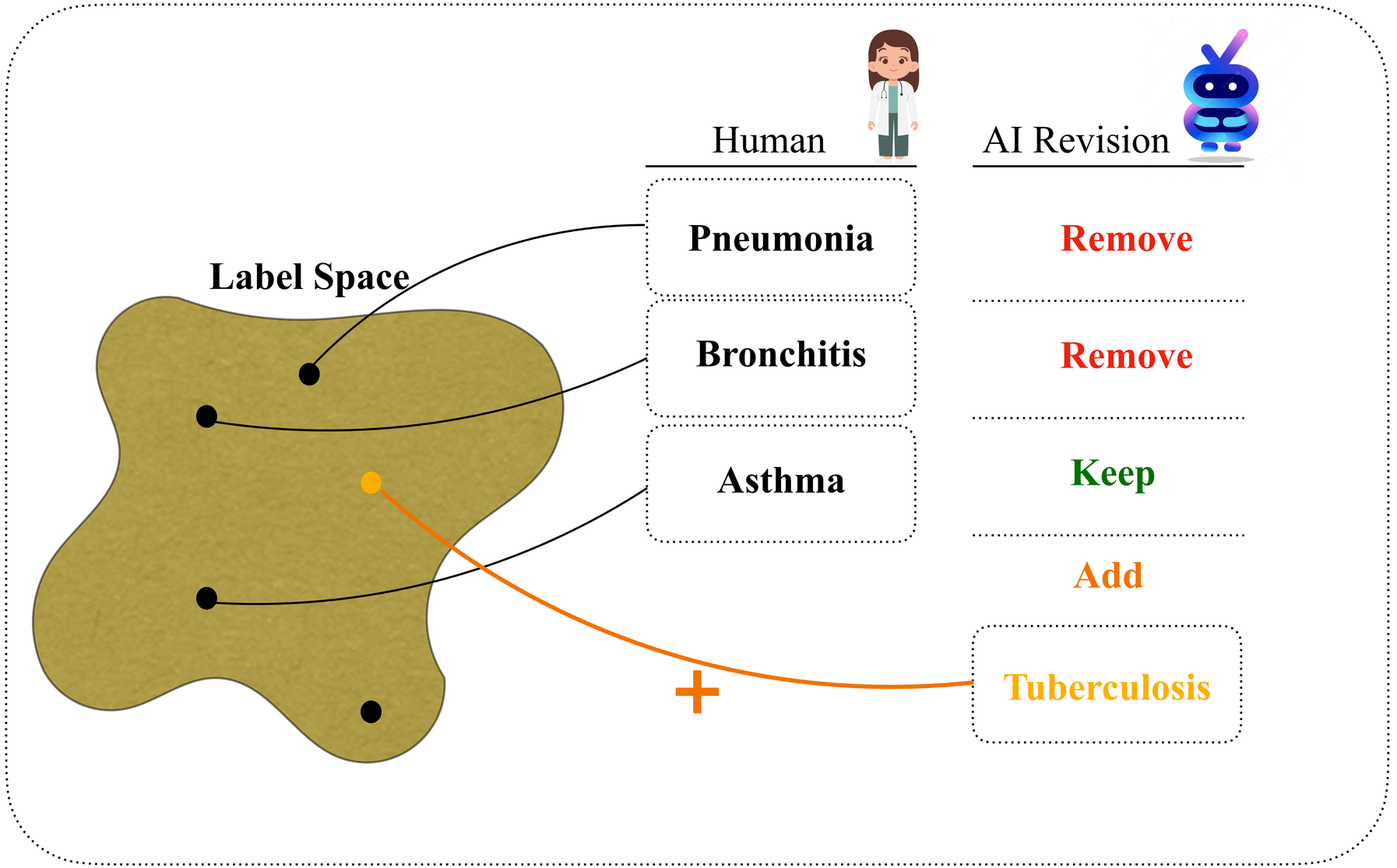
Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?

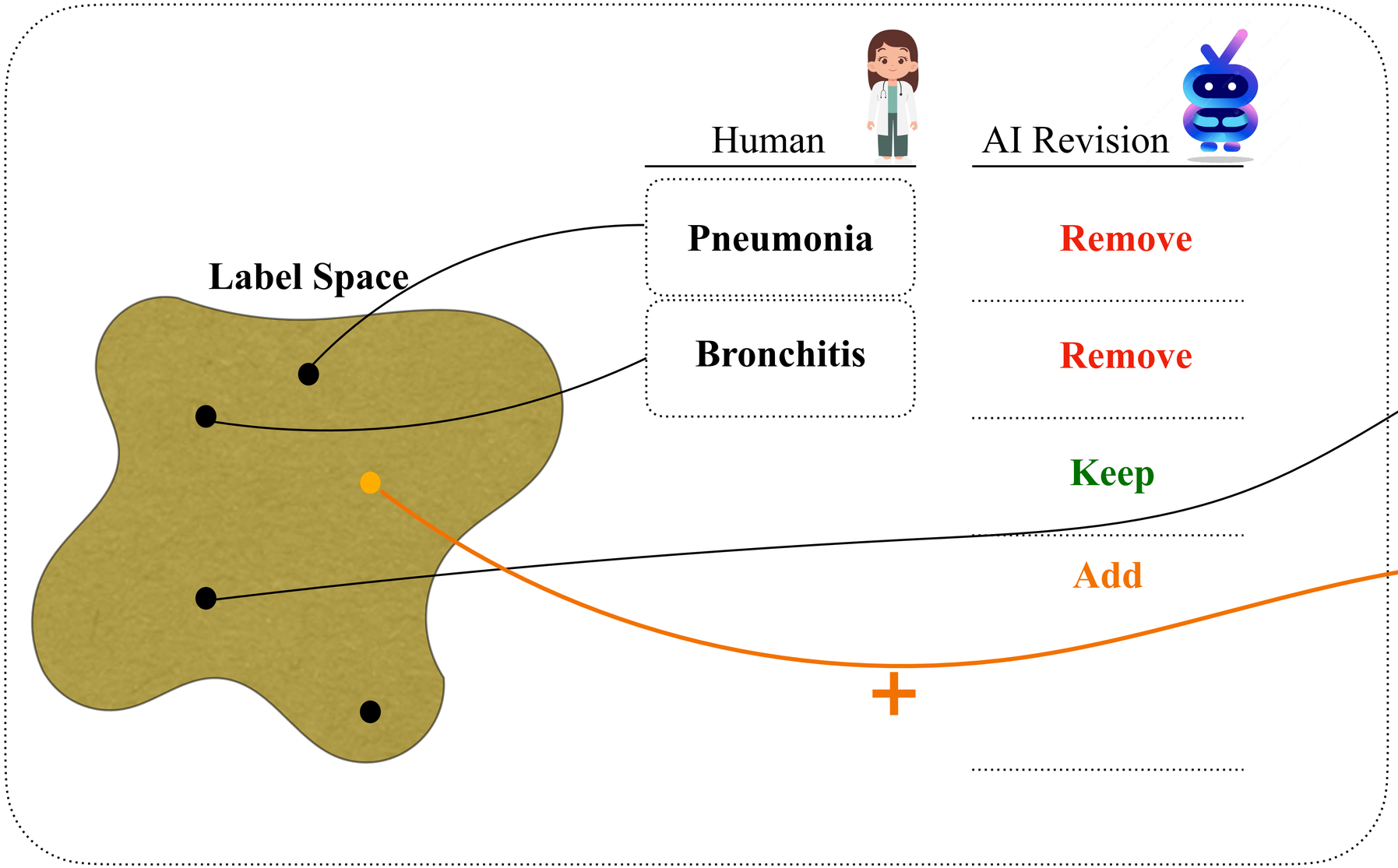


Question: what constitutes a good collaboration?

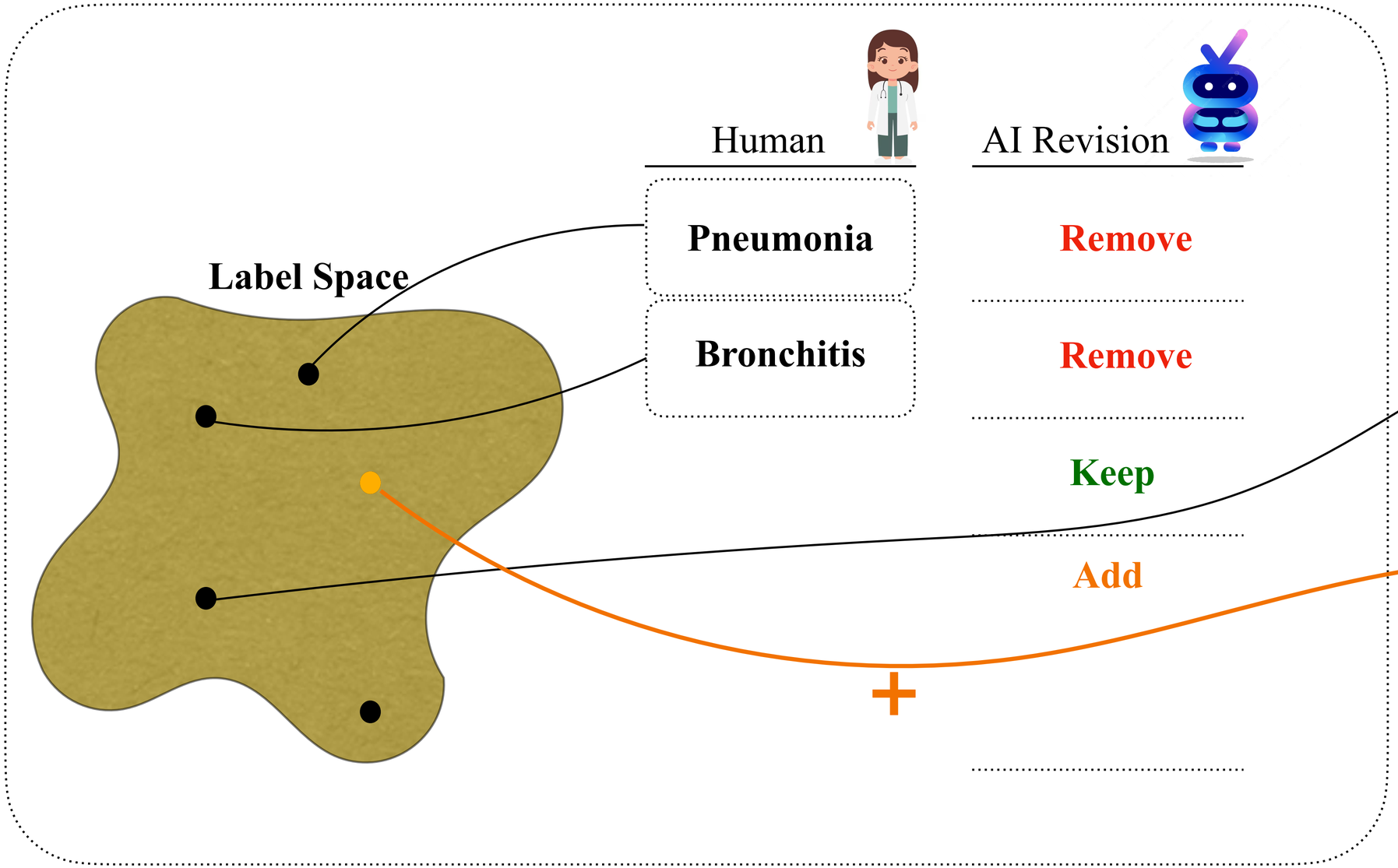


Collaborative Proposal

Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?



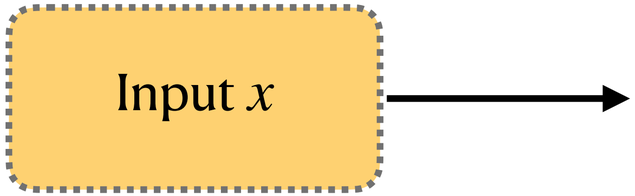
Second: Add value!

Two fundamentals of collaboration

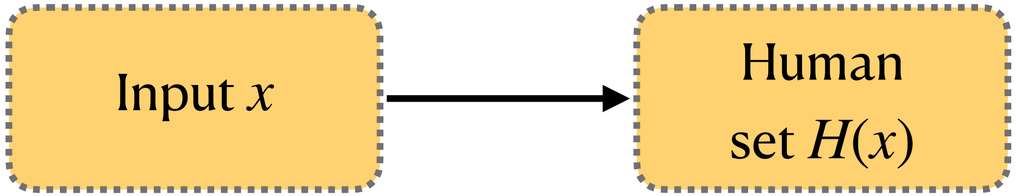
Two fundamentals of collaboration

Input x

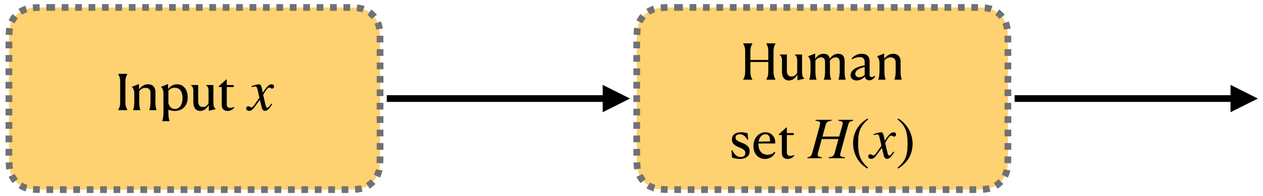
Two fundamentals of collaboration



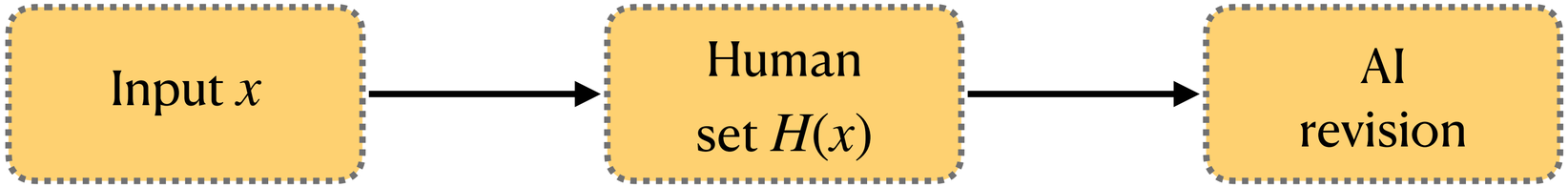
Two fundamentals of collaboration



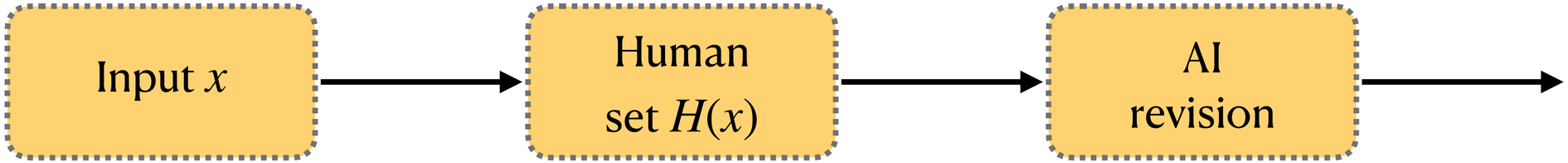
Two fundamentals of collaboration



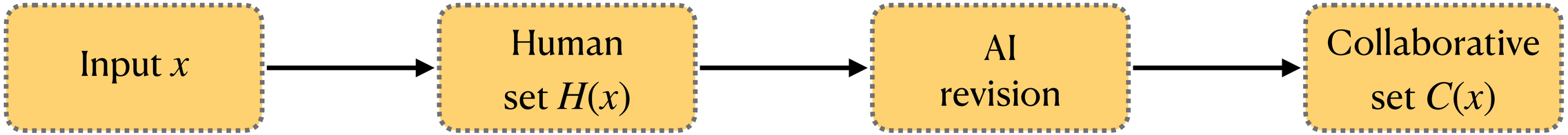
Two fundamentals of collaboration



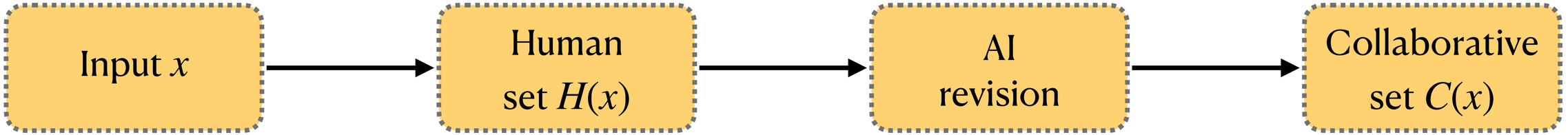
Two fundamentals of collaboration



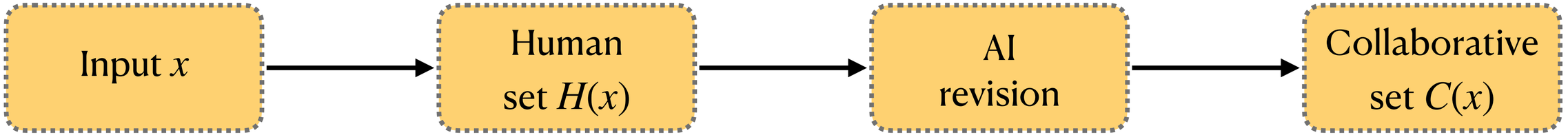
Two fundamentals of collaboration

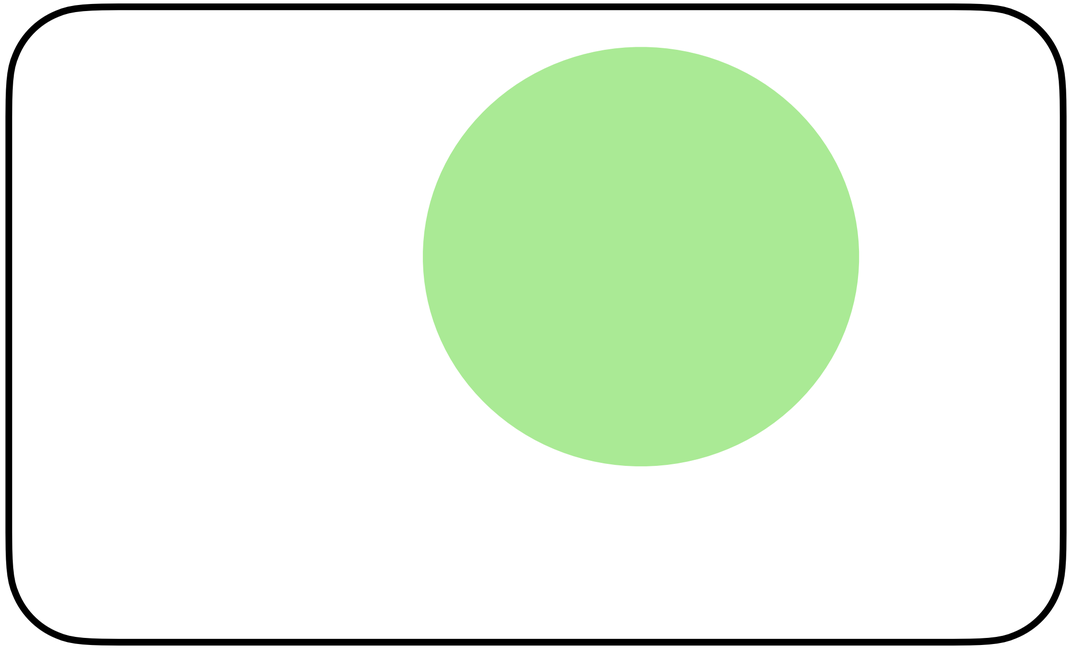


Two fundamentals of collaboration

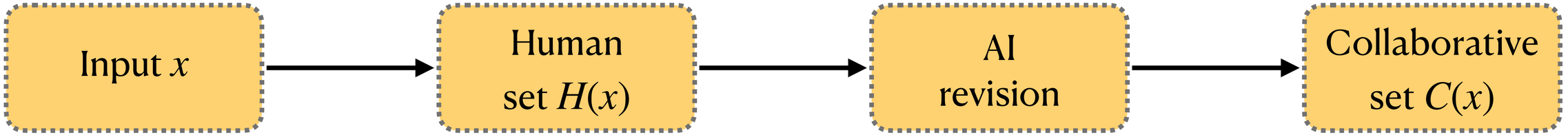


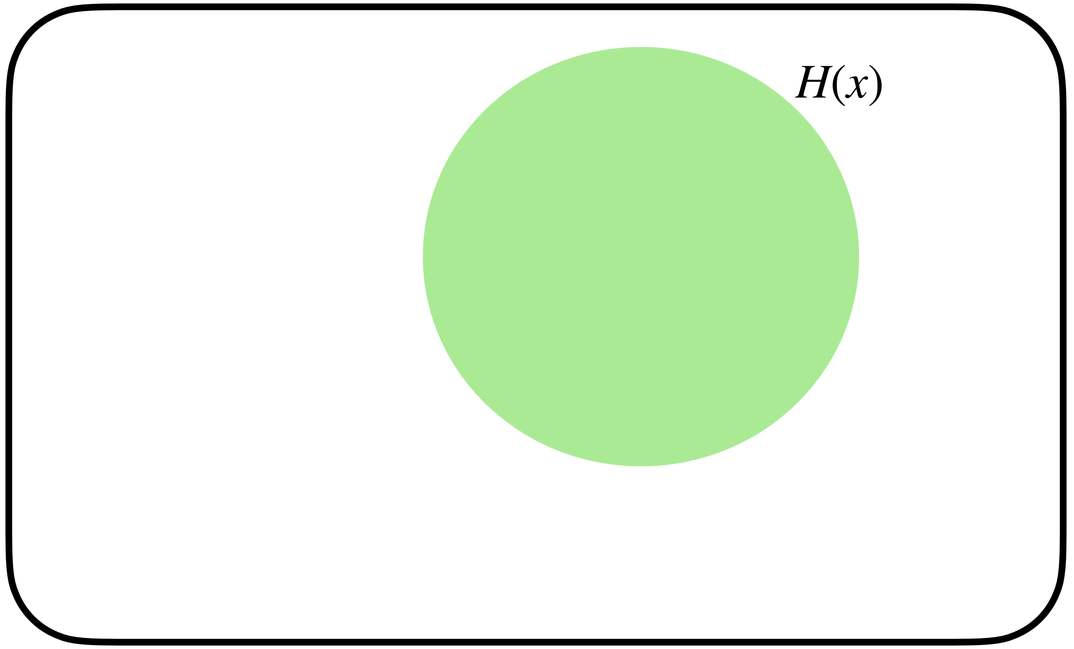
Two fundamentals of collaboration



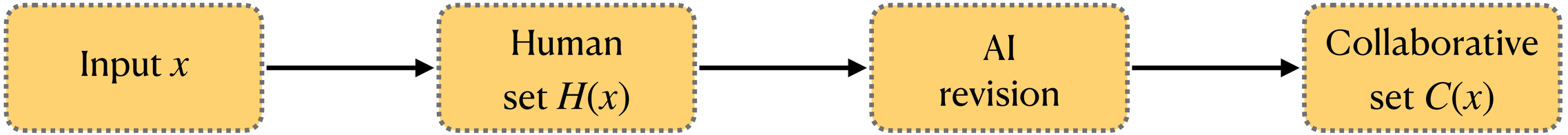


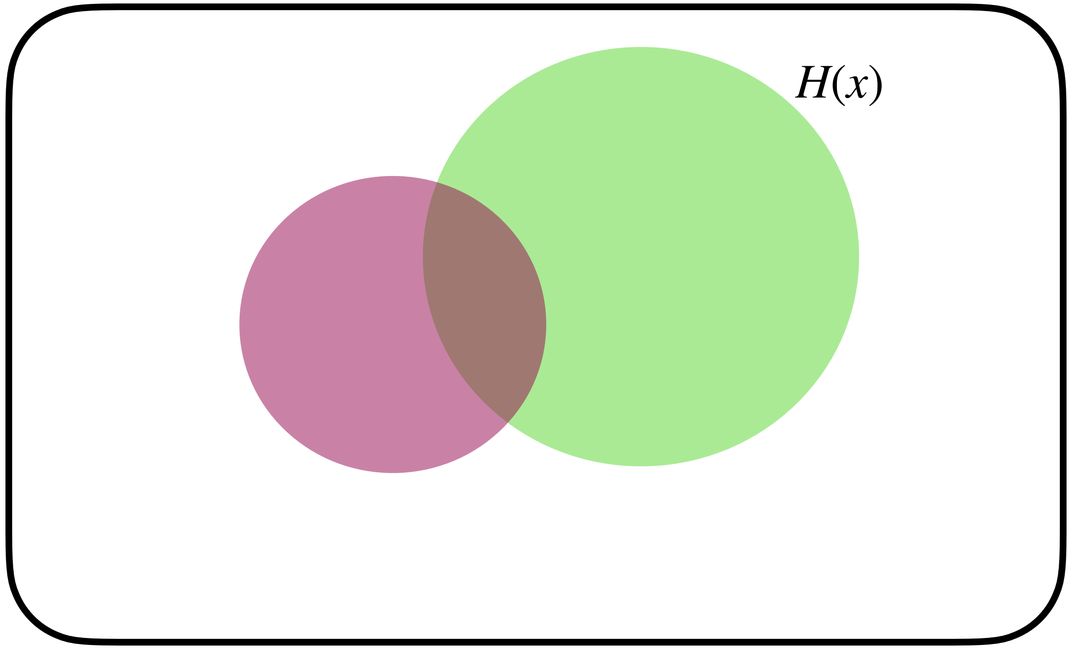
Two fundamentals of collaboration



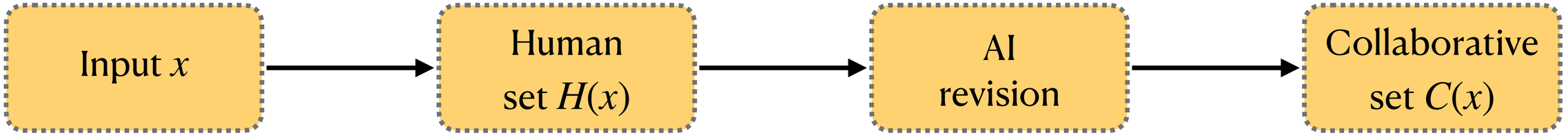


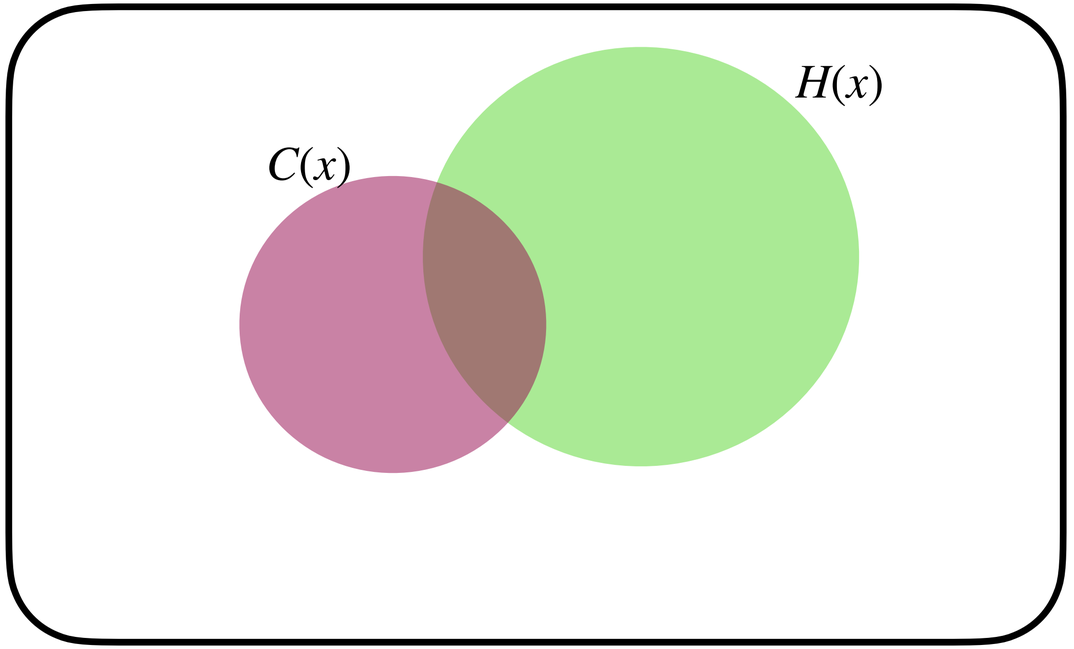
Two fundamentals of collaboration



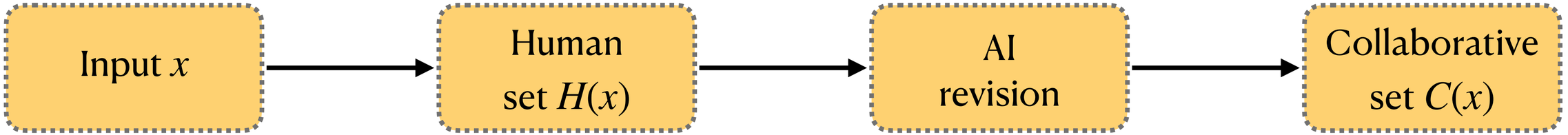


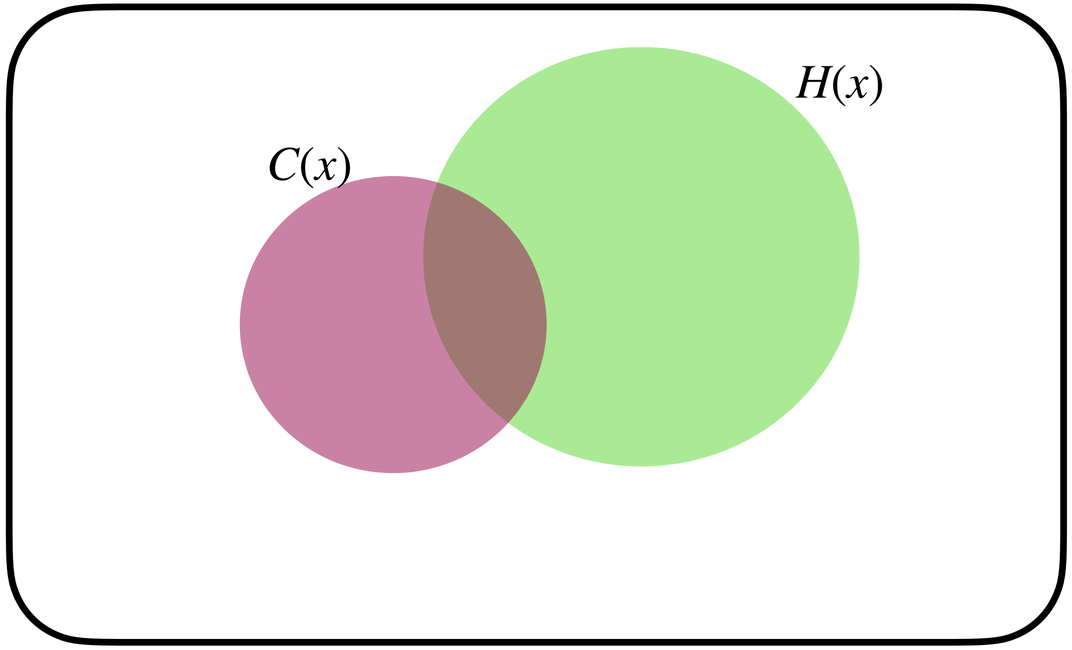
Two fundamentals of collaboration





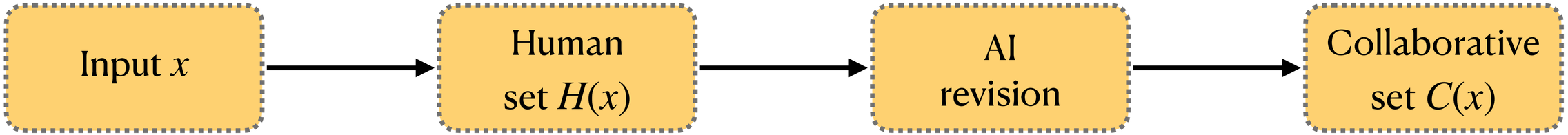
Two fundamentals of collaboration

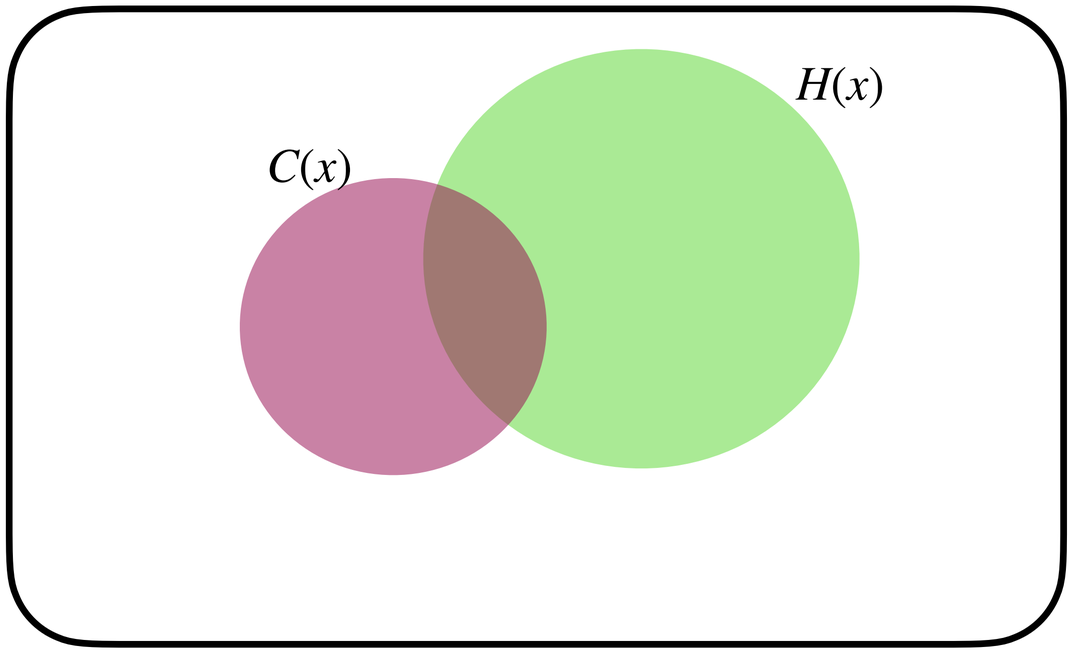




Counterfactual Harm

Two fundamentals of collaboration

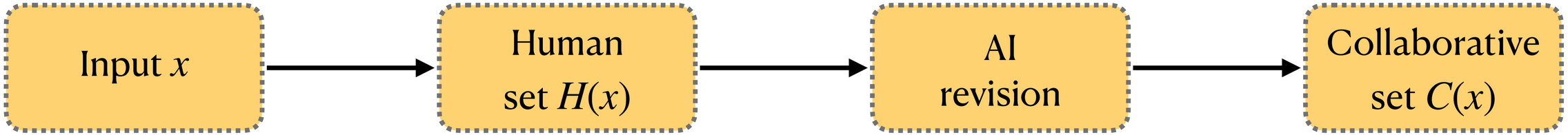


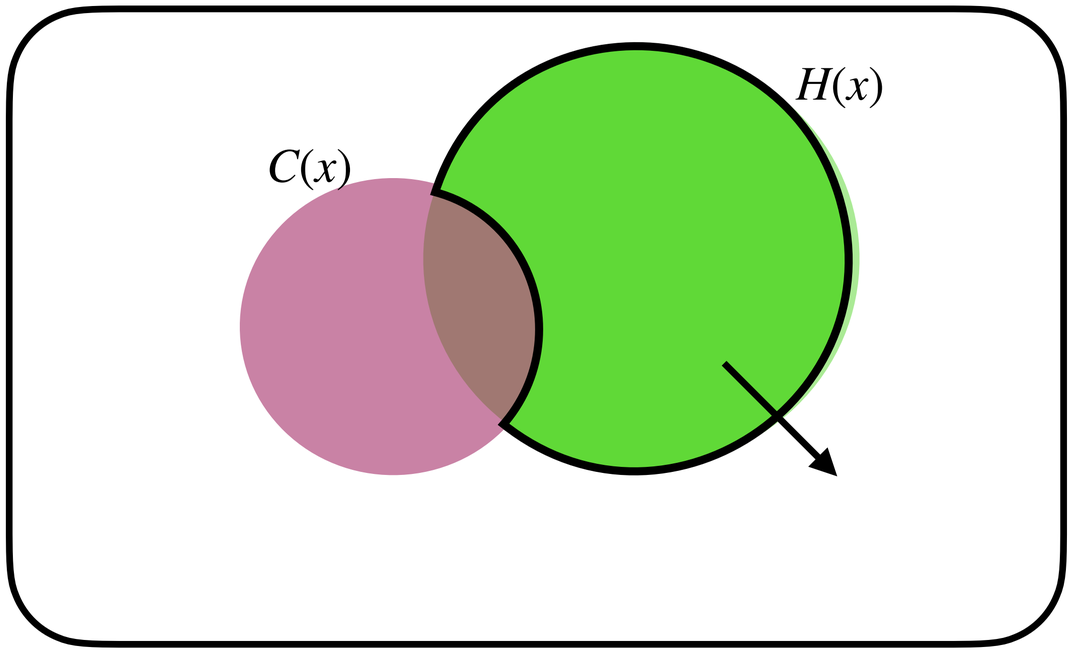


Counterfactual Harm

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

Two fundamentals of collaboration

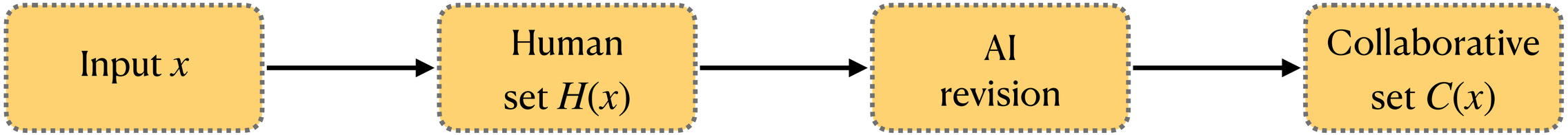


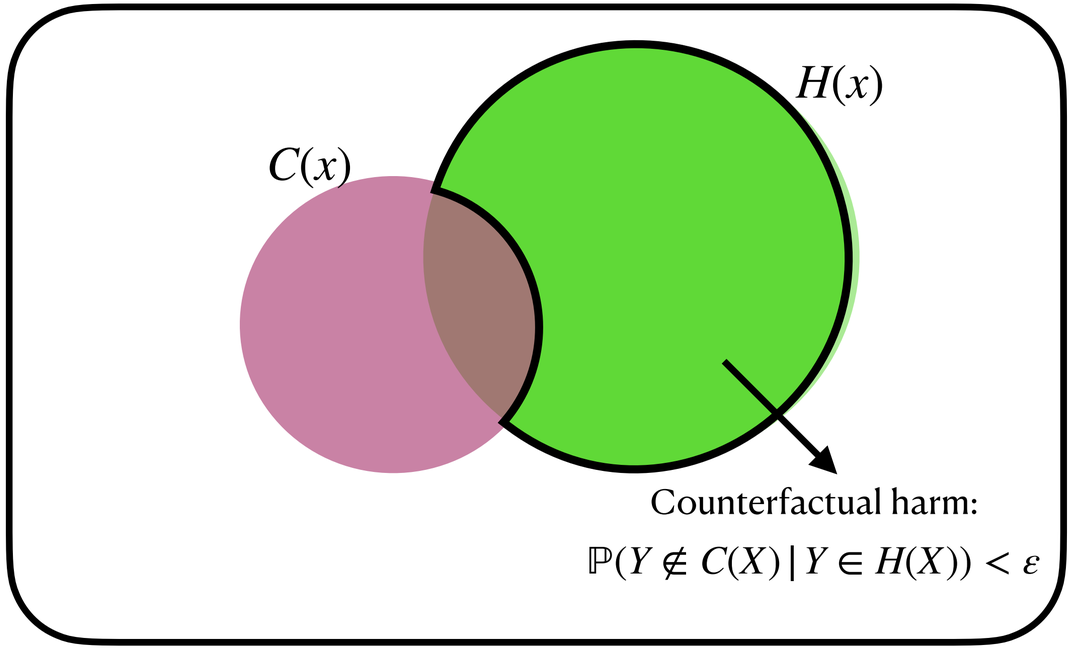


Counterfactual Harm

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

Two fundamentals of collaboration

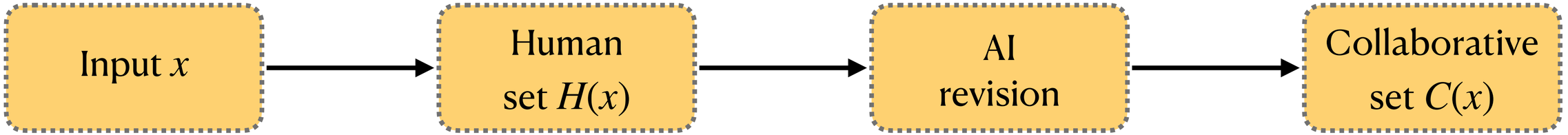


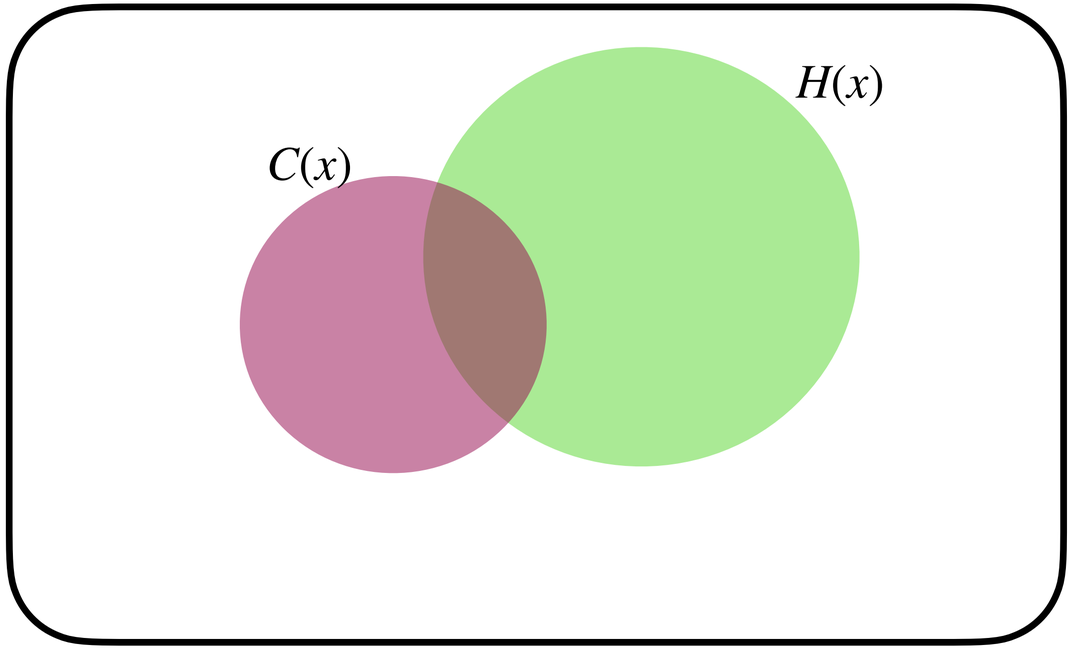


Counterfactual Harm

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

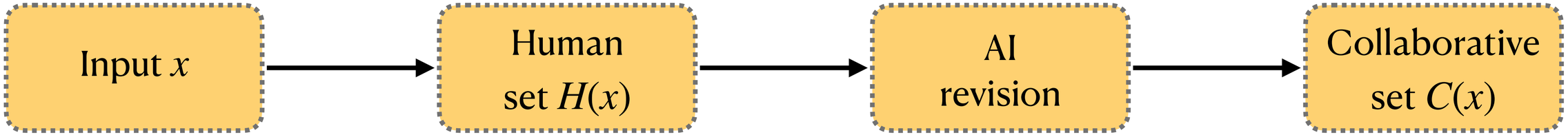
Two fundamentals of collaboration

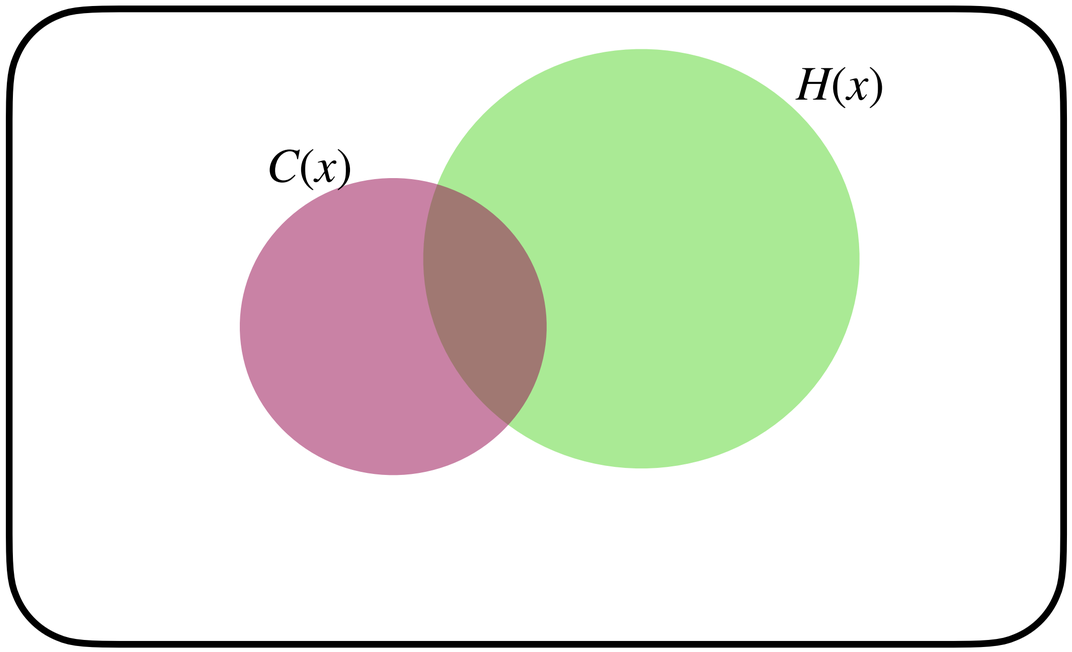




Complementarity

Two fundamentals of collaboration

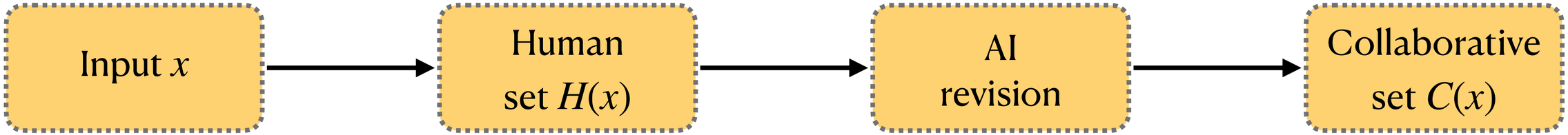


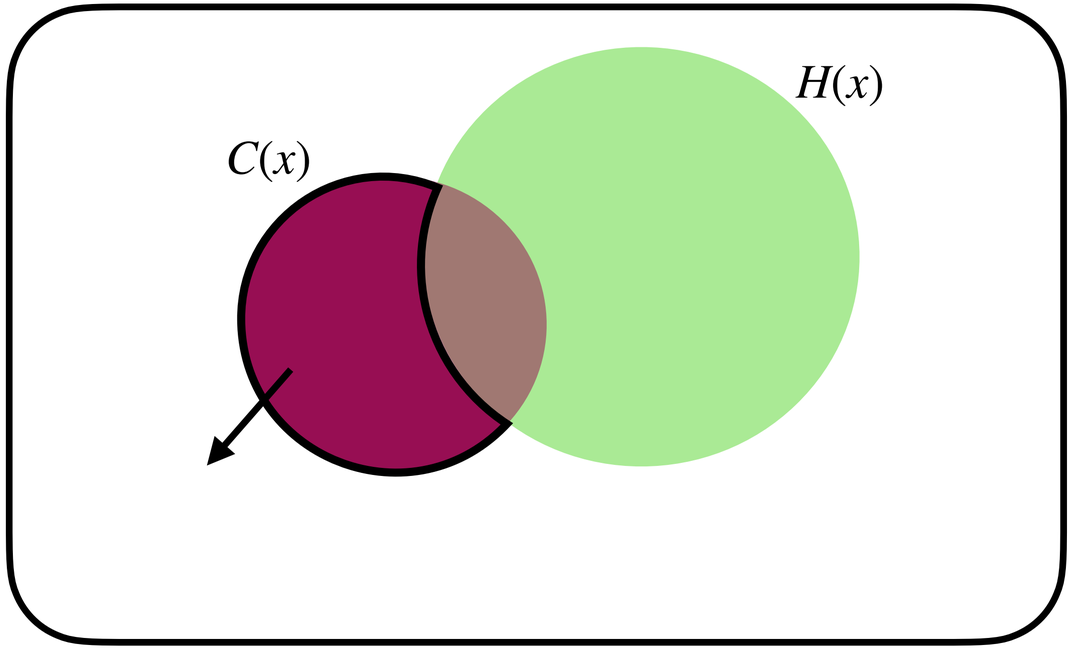


Complementarity

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

Two fundamentals of collaboration

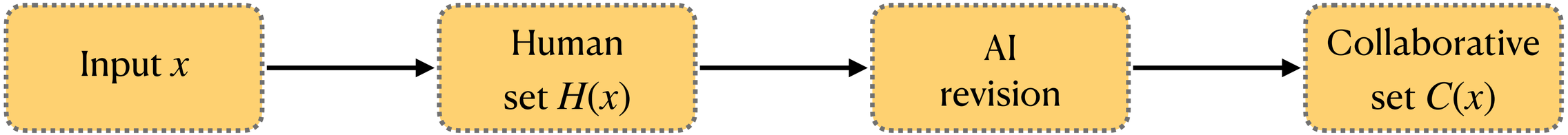


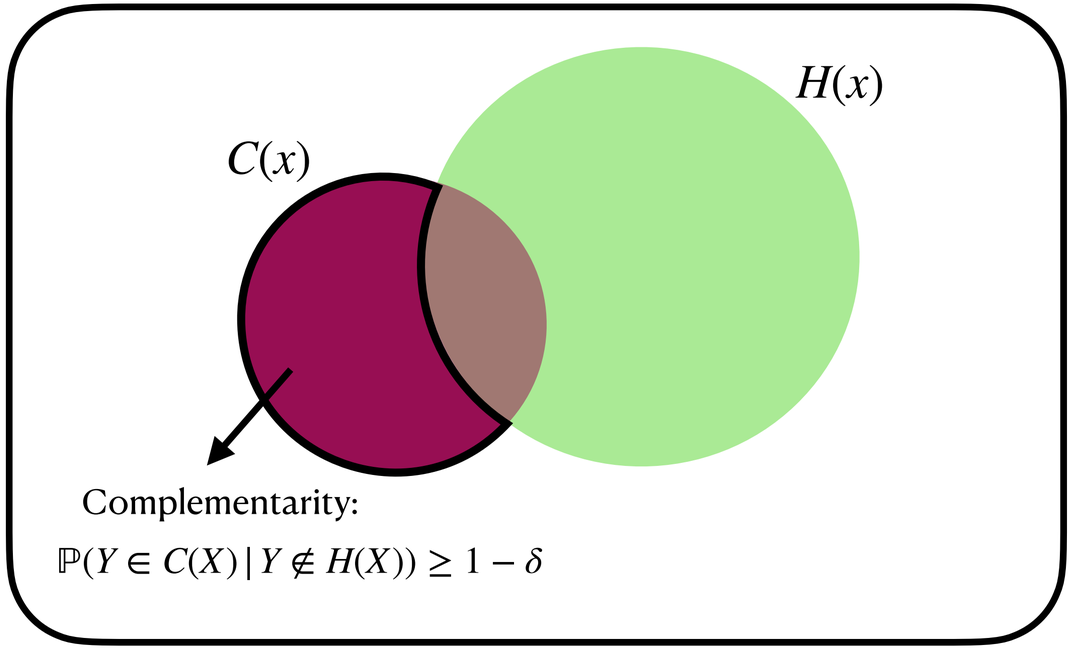


Complementarity

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

Two fundamentals of collaboration

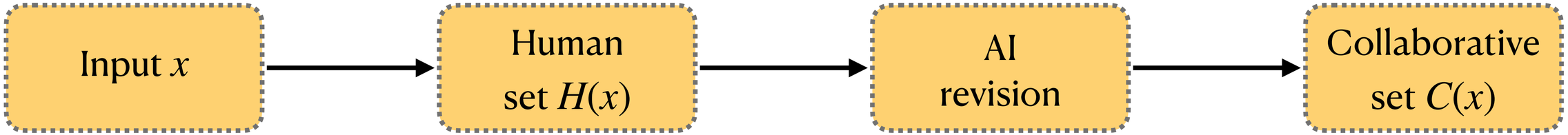


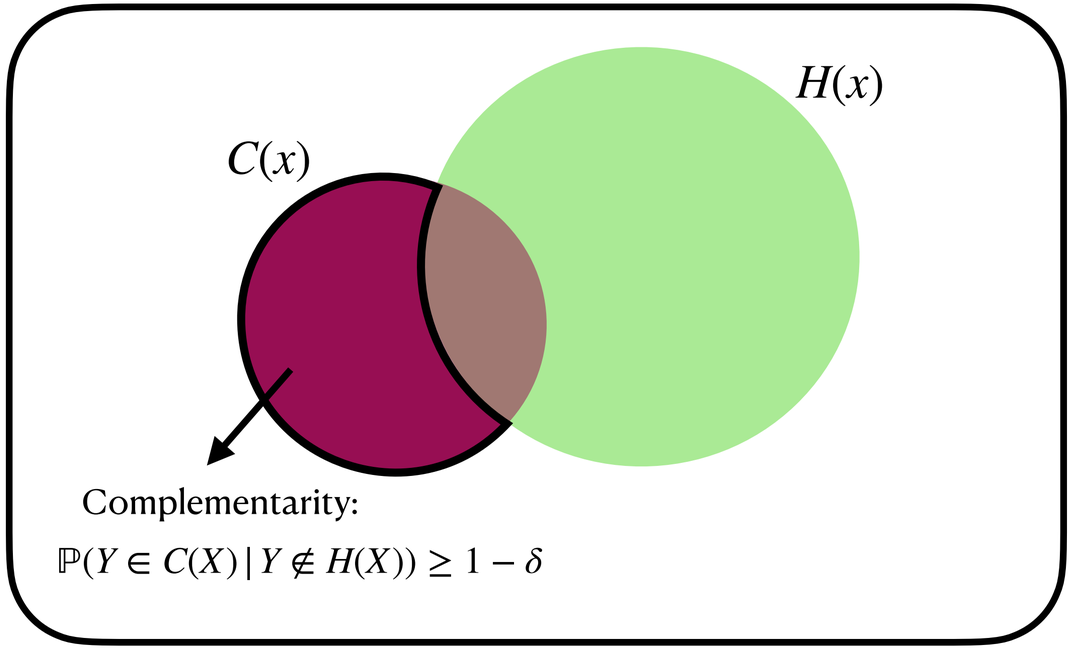


Complementarity

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

Two fundamentals of collaboration

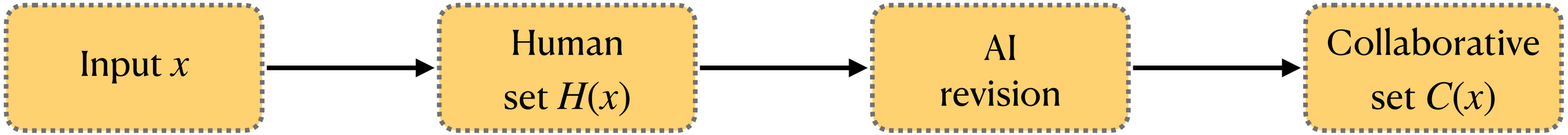


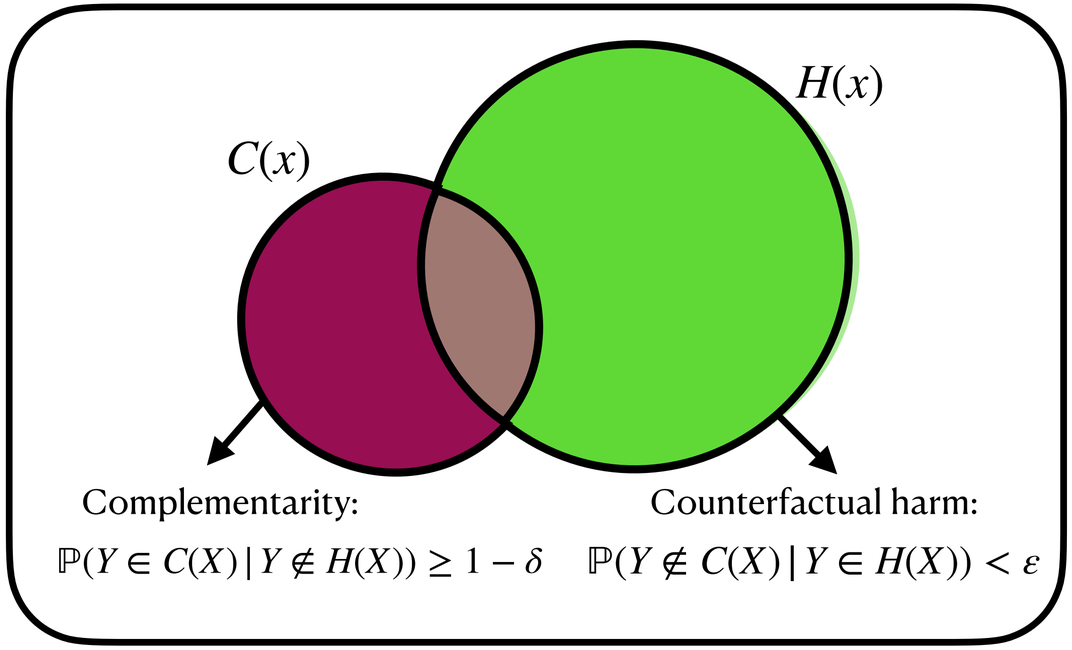


Complementarity

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

Two fundamentals of collaboration





Counterfactual Harm

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

Complementarity

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

Question: what constitutes a good collaboration?

Question: **what constitutes a good collaboration?**

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

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

Question: what constitutes a good collaboration?

$$\min_{C:\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|C(X)|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) < \varepsilon, \qquad \text{Counterfactual Harm}$$

$$\mathbb{P}\Big(Y \in C(X) \,\Big|\, Y \notin H(X)\Big) \geq 1-\delta\,.$$

Question: what constitutes a good collaboration?

$$\min_{C:\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|C(X)|$$

$$\text{s.t.} \quad \mathbb{P}\big(Y \notin C(X) \,\big|\, Y \in H(X)\big) < \varepsilon, \qquad \text{Counterfactual Harm}$$

$$\mathbb{P}\big(Y \in C(X) \,\big|\, Y \notin H(X)\big) \geq 1-\delta. \qquad \text{Complementarity}$$

Human-AI Collaborative Optimization

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

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

HACO

Human-AI Collaborative Optimization

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

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

HACO

Theorem: The optimal solution to HACO is of the form

Human-AI Collaborative Optimization

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

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

HACO

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

VS

Ordinary CP:

$$C(x) = \{ y \mid 1 - p(y \mid x) \leq q^* \}$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

$s(x, y)$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; s(x, y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; s(x, y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

1 **Question:** How to design the score function?

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; s(x,y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

1. **Question:** How to design the score function?

2. **Question:** How to debias the thresholds?

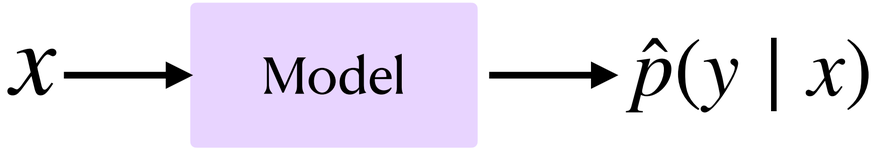
Question: How to design the score function?

Question: How to design the score function?

$$C^*(x) = \left\{ y \;\middle|\; \boxed{1 - p(y \mid x)} \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

$s(x, y)$

Classification



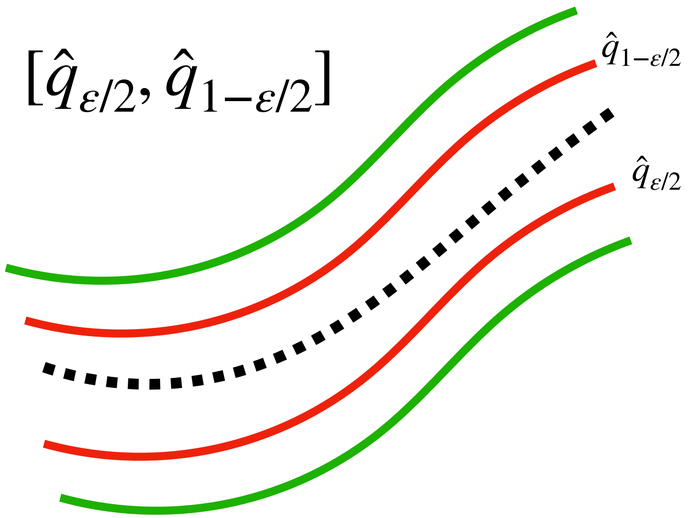
$$\hat{s}(x, y) = 1 - \hat{p}(y \mid x)$$

Question: How to design the score function?

$$C^*(x) = \left\{ y \;\middle|\; \boxed{1 - p(y \mid x)} \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

$s(x, y)$

Conformalized Quantile Regression



$$\hat{s}(x, y) = \max\left\{\hat{q}_{\varepsilon/2}(x) - y,\, y - \hat{q}_{1-\varepsilon/2}(x)\right\}$$

Our two threshold structure

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

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; \hat{s}(x,y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

1. **Question:** How to design the score function?

2. **Question:** How to debias the thresholds?

Question: How to debias the thresholds?

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

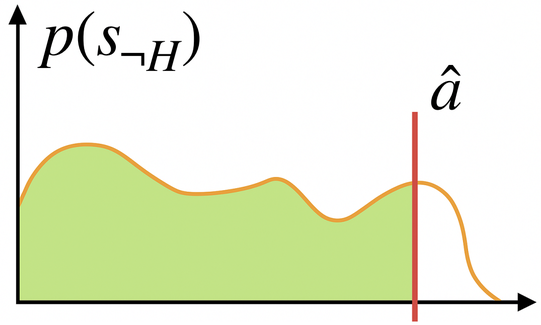
Assume Exchangeability **What should a and b be?**

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability **What should a and b be?**

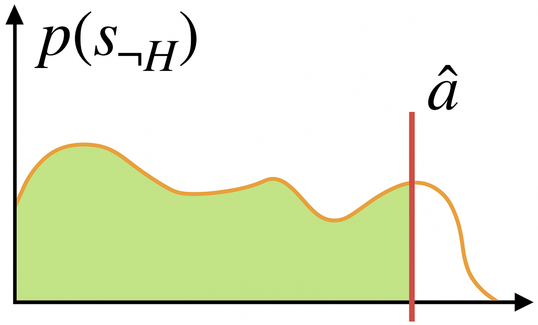


Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability **What should a and b be?**



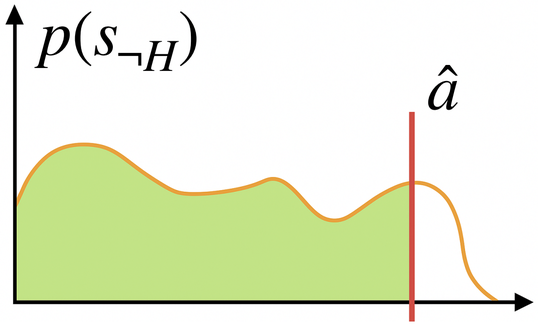
$$\hat{a} = Q_{1-\delta}\big(\{ s_i : \; Y_i \notin H(X_i) \} \cup \{\infty\}\big)$$

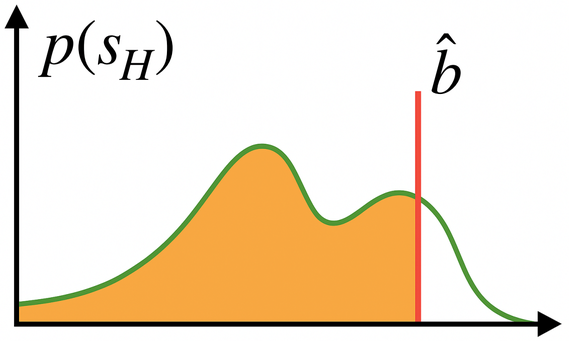
Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability **What should a and b be?**





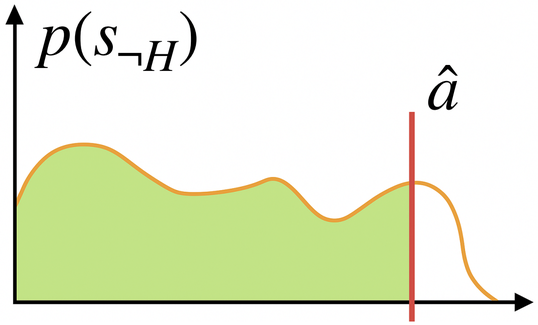
$$\hat{a} = Q_{1-\delta}\big(\{\, s_i : \;\; Y_i \notin H(X_i) \,\} \cup \{\infty\}\big)$$

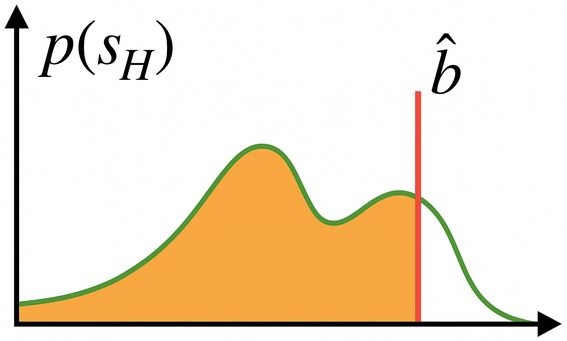
Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability **What should a and b be?**





$$\hat{a} = Q_{1-\delta}\big(\{\, s_i : \;\; Y_i \notin H(X_i) \,\} \cup \{\infty\}\big)$$

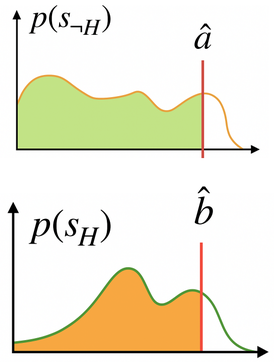
$$\hat{b} = Q_{1-\varepsilon}\big(\{\, s_i : \;\; Y_i \in H(X_i) \,\} \cup \{\infty\}\big)$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability

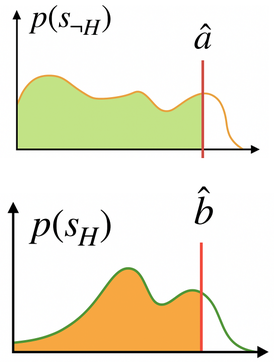


Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



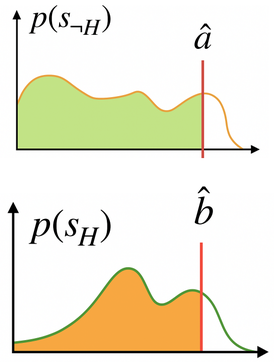
For a new test example (x_{test}, y_{test}) :

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

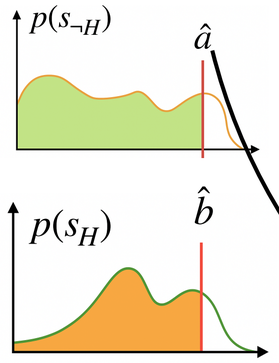
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

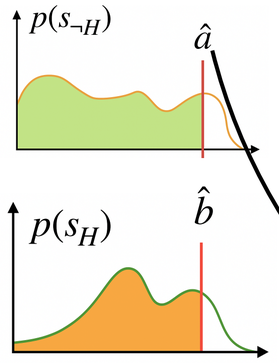
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

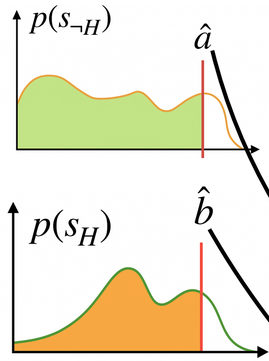
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

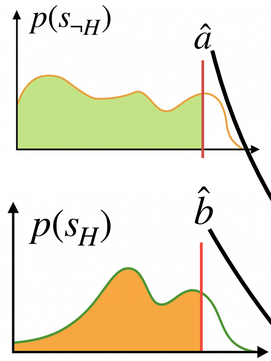
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \hat{b} \mathbf{1}_{\{y \in H(x_{\text{test}})\}} \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability $\hat{a} = Q_{1-\delta}\big(\{\, s_i : \; Y_i \notin H(X_i) \,\} \cup \{\infty\}\big)\,; \quad \hat{b} = Q_{1-\varepsilon}\big(\{\, s_i : \; Y_i \in H(X_i) \,\} \cup \{\infty\}\big)$

$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a}\, \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \hat{b}\, \mathbf{1}_{\{y \in H(x_{\text{test}})\}} \right\}.$$

$$1 - \varepsilon \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \in H(X_{test})\right] \;<\; 1 - \varepsilon + \frac{1}{n_1 + 1}$$

$$1 - \delta \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \notin H(X_{test})\right] \;<\; 1 - \delta + \frac{1}{n_2 + 1}$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability $\hat{a} = Q_{1-\delta}\big(\{ s_i : \; Y_i \notin H(X_i) \} \cup \{\infty\}\big)$; $\hat{b} = Q_{1-\varepsilon}\big(\{ s_i : \; Y_i \in H(X_i) \} \cup \{\infty\}\big)$

$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \hat{b} \mathbf{1}_{\{y \in H(x_{\text{test}})\}} \right\}.$$

Offline Guarantees

$$1 - \varepsilon \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \in H(X_{test})\right] \;<\; 1 - \varepsilon + \frac{1}{n_1 + 1}$$

$$1 - \delta \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \notin H(X_{test})\right] \;<\; 1 - \delta + \frac{1}{n_2 + 1}$$

~~Assume Exchangeability~~

~~Assume Exchangeability~~

Exchangeability is fragile

~~Assume Exchangeability~~

Exchangeability is fragile

Question: How to debias the thresholds in the online setting

Question: How to debias the thresholds in the online setting

Question: How to debias the thresholds in the online setting

At round t

Question: How to debias the thresholds in the online setting

At round t

input x_t

Question: How to debias the thresholds in the online setting

At round t

input x_t

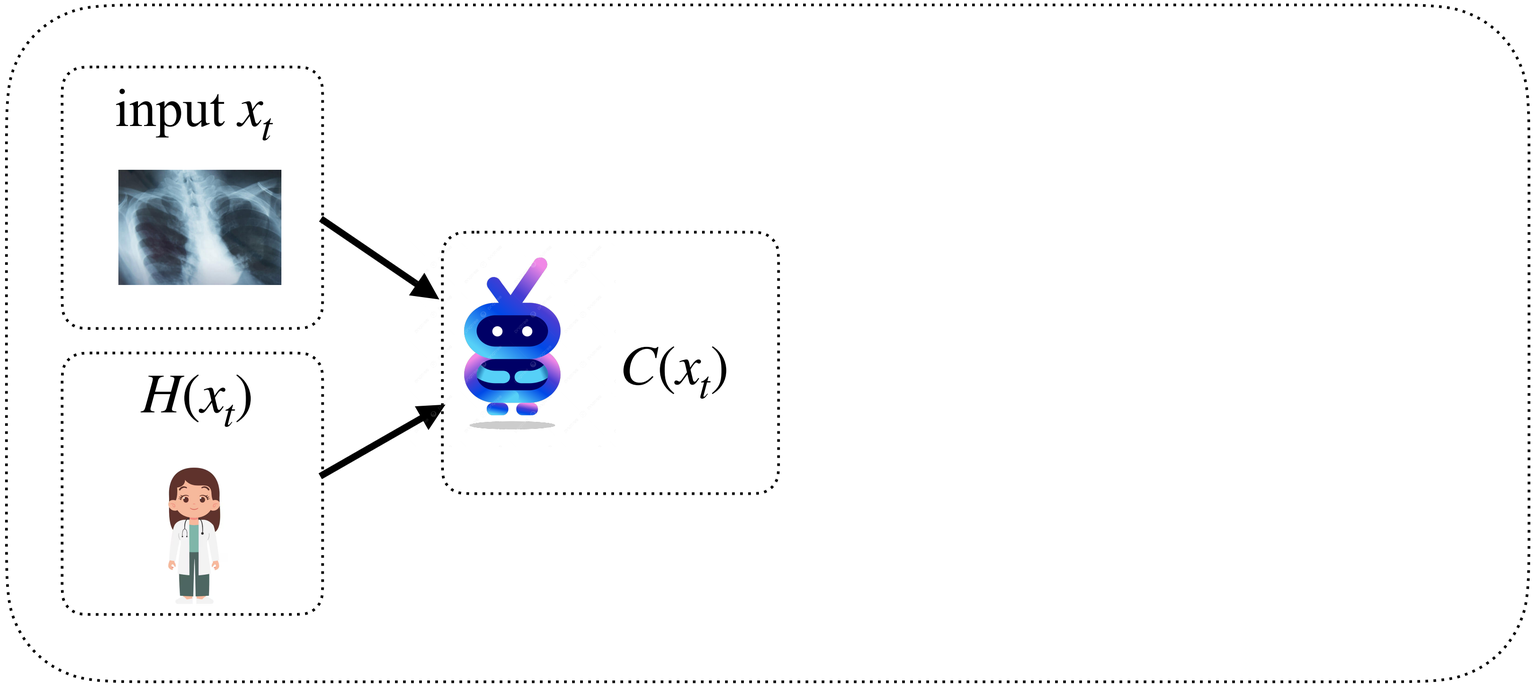
$H(x_t)$

Question: How to debias the thresholds in the online setting

At round t

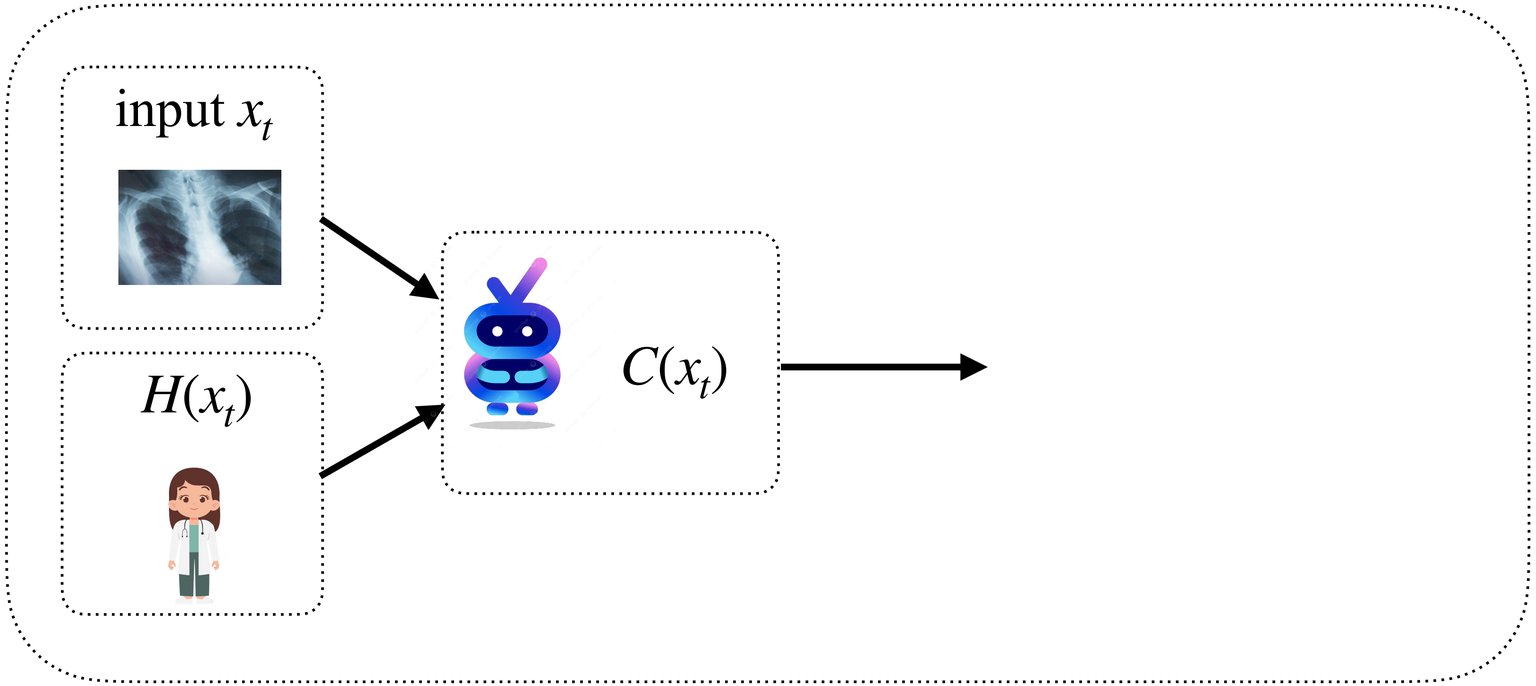
Question: How to debias the thresholds in the online setting

At round t



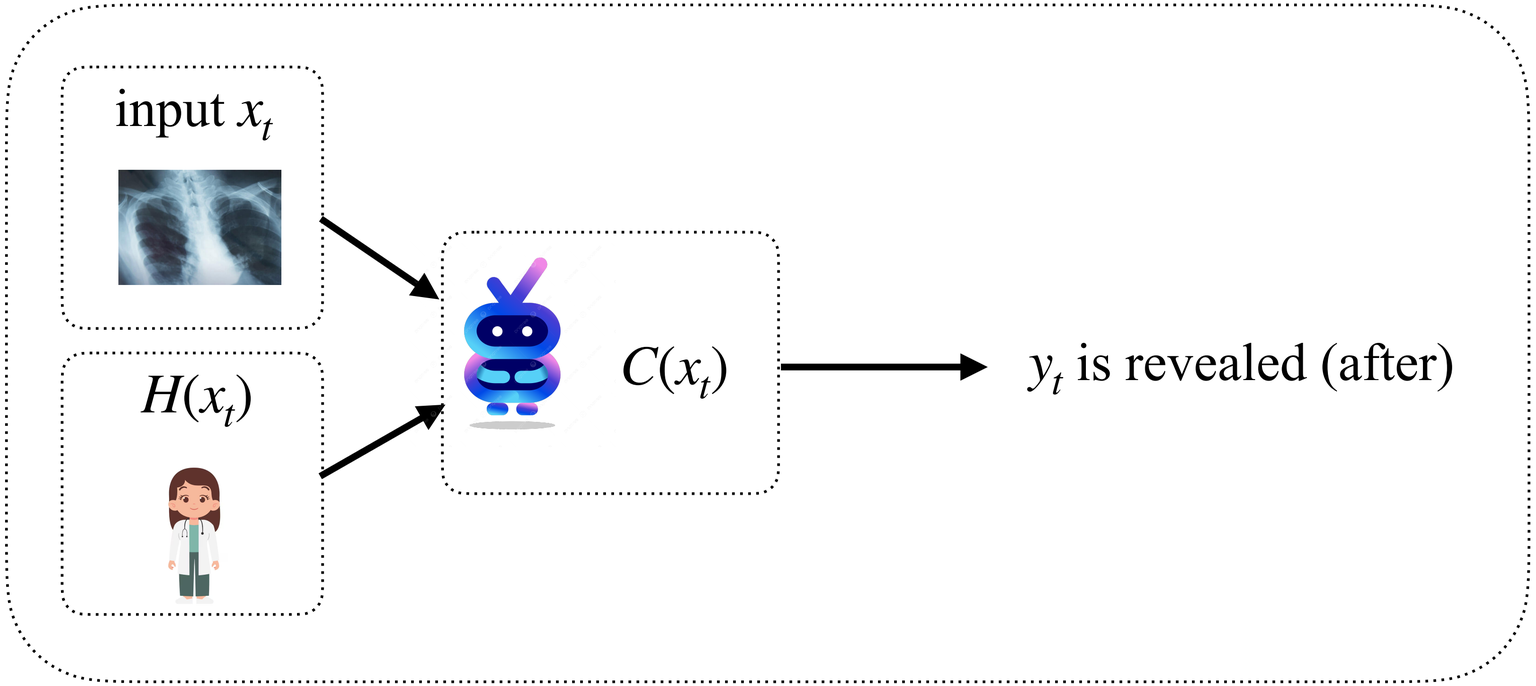
Question: How to debias the thresholds in the online setting

At round t



Question: How to debias the thresholds in the online setting

At round t



Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \middle| \; s(x_t, y) \leq a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \middle| \; s(x_t, y) \leq a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \,\middle|\, s(x_t, y) \le a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

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

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \,\middle|\, s(x_t, y) \le a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

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

Counterfactual Harm Error

Question: How to debias the thresholds in the **online** setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \,\middle|\, s(x_t, y) \le a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

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

Counterfactual Harm Error

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

Complementarity Error

Question: How to debias the thresholds in the online setting

We track two errors

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

Counterfactual Harm Error

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

Complementarity Error

We update the thresholds as follows

$$\textit{If} \quad (\, y_t \in H(x_t) \,) \longrightarrow$$

Question: How to debias the thresholds in the online setting

We track two errors

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

Counterfactual Harm Error | **Complementarity Error**

We update the thresholds as follows

$$\textit{If} \quad (\, y_t \in H(x_t)\,) \longrightarrow \quad b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

Question: How to debias the thresholds in the online setting

We track two errors

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

Counterfactual Harm Error | **Complementarity Error**

We update the thresholds as follows

$$\textit{If} \quad (\, y_t \in H(x_t)\,) \longrightarrow \quad b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Question: How to debias the thresholds in the online setting

We track two errors

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

Counterfactual Harm Error | **Complementarity Error**

We update the thresholds as follows

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

If $(\, y_t \notin H(x_t) \,)$

Question: How to debias the thresholds in the online setting

We track two errors

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

Counterfactual Harm Error

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

Complementarity Error

We update the thresholds as follows

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

Question: How to debias the thresholds in the online setting

We track two errors

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

Counterfactual Harm Error

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

Complementarity Error

We update the thresholds as follows

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

Question: How to debias the thresholds in the online setting

Counterfactual Harm Error

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

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Complementarity Error

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

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

Theorem: Assume the conformity score is bounded, i.e $s(x, y) \in [0,1]$:

Question: How to debias the thresholds in the online setting

Counterfactual Harm Error

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

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Complementarity Error

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

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

Theorem: Assume the conformity score is bounded, i.e $s(x, y) \in [0,1]$:

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

Question: How to debias the thresholds in the online setting

Counterfactual Harm Error

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

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Complementarity Error

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

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

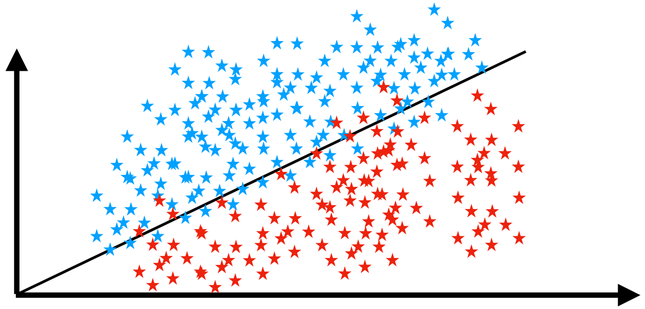
Theorem: Assume the conformity score is bounded, i.e $s(x, y) \in [0,1]$:

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

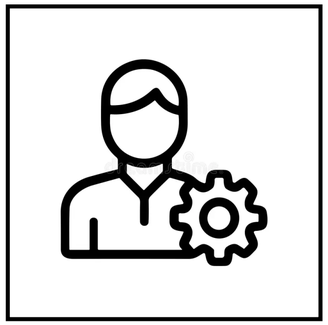
$$\left| \frac{1}{N_2(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{out}} - \delta \right| \leq \frac{1 + \eta \ \max(\delta, 1-\delta)}{\eta\, N_2(T)}$$

Experiments

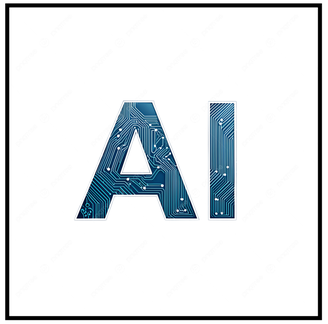
We consider three modalities of data



Baselines

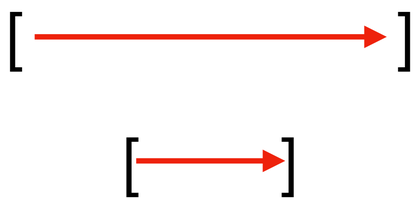


Human Alone

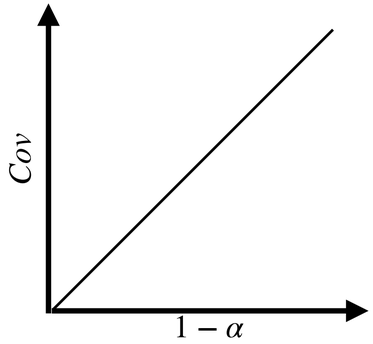


AI Alone

Evaluation Metics

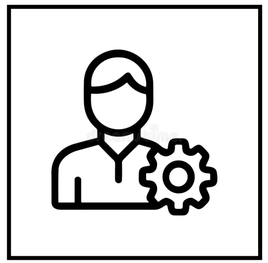


Set Size

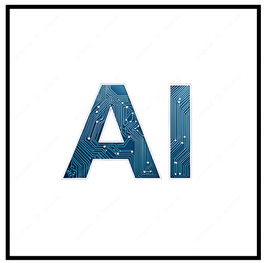


Marginal Coverage

Text Based Medical Diagnosis



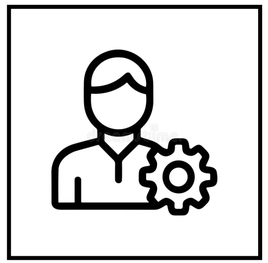
Rule-based diagnostic system



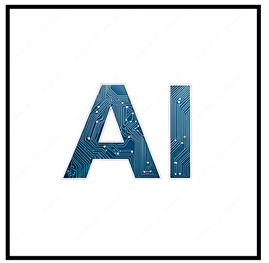
LLMs

Offline Setting

Text Based Medical Diagnosis



Rule-based diagnostic system

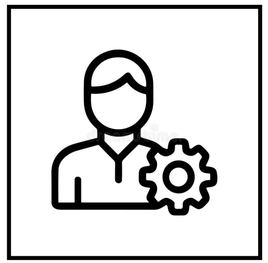


LLMs

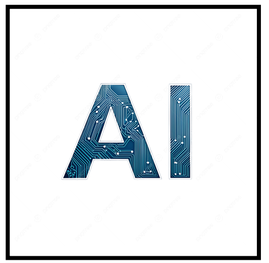
Offline Setting

	Human	GPT-4o			GPT-5		
Strategy	**C/S**	**CUP C/S**	**(ϵ, δ)**	**AI C/S**	**CUP C/S**	**(ϵ, δ)**	**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

Text Based Medical Diagnosis



Rule-based diagnostic system



LLMs

Offline Setting

	Human	GPT-4o			GPT-5		
Strategy	**C/S**	**CUP C/S**	**(ϵ, δ)**	**AI C/S**	**CUP C/S**	**(ϵ, δ)**	**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

AI's quality affects the overall collaboration quality!

Thank You!