Conformal Prediction: From Images to Agents

CVIS 2025

Jesse Cresswell

Dec. 15 2025

layer 6
AI at TD

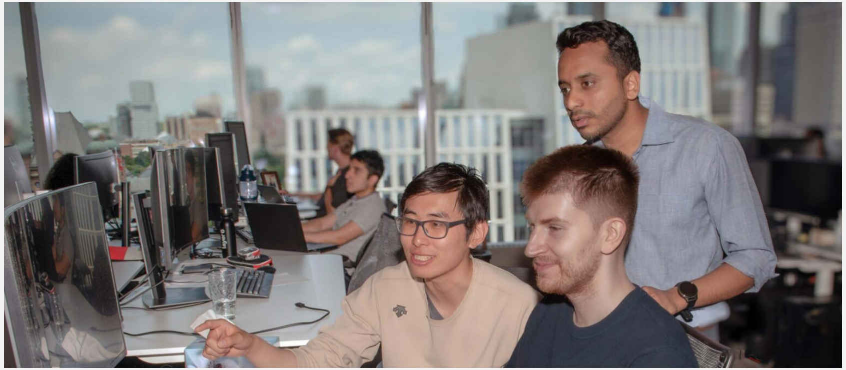
Layer 6 AI at TD

Layer 6 acts as the AI/ML brain center of the bank.

Our team of >100 scientists, engineers, and product owners serves all parts of TD, and we are hiring aggressively now - visit layer6.ai.

We handle the hardest modelling problems across the entire bank.

Research | Scale | Impact



Part I

Conformal Prediction as Uncertainty Quantification

Part II

Conformal Prediction for Statistical Guarantees of Correctness

Part I

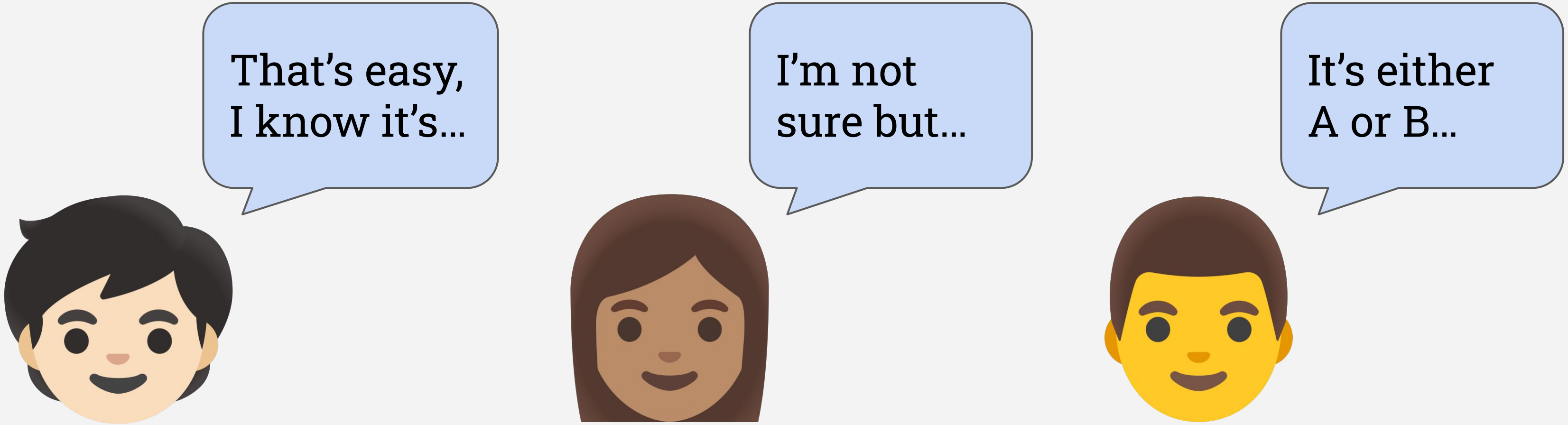
Conformal Prediction as Uncertainty Quantification

Part II

Conformal Prediction for Statistical Guarantees of Correctness

Why should we quantify uncertainty?

When humans answer questions, we naturally state how confident we are.
It's a crucial aspect of decision making.



We **signal confidence**, and **offer alternatives.** Models do not.
They give you one answer every time, even when they shouldn't

Conformal Prediction

Conformal prediction is a general purpose method for transforming heuristic notions of uncertainty into rigorous ones.

Instead of outputting a single prediction, conformal prediction returns a **set**.

Conformal Prediction

Conformal prediction is a general purpose method for transforming heuristic notions of uncertainty into rigorous ones.

Instead of outputting a single prediction, conformal prediction returns a **set**.

Image Classification Example:

Conformal Prediction

Conformal prediction is a general purpose method for transforming heuristic notions of uncertainty into rigorous ones.

Instead of outputting a single prediction, conformal prediction returns a **set**.

{Container Ship}

Conformal Prediction

Conformal prediction is a general purpose method for transforming heuristic notions of uncertainty into rigorous ones.

Instead of outputting a single prediction, conformal prediction returns a **set**.

{Container Ship}

{Squirrel Monkey, Spider Monkey, Lemur}

layer 6
AI at TD

Conformal Prediction

{Container Ship}

{Squirrel Monkey,
Spider Monkey, Lemur}

The **size** of a prediction set **quantifies how uncertain** the model is.

When the model is more uncertain, the prediction set is larger, and this **provides alternatives** to the point prediction.

layer 6 AI at TD

Conformal Prediction

Conformal prediction provides a statistical guarantee:

"The correct answer is in the prediction set with probability at least 1-α."

$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \alpha$$

1-α can be thought of as the **success rate -** we choose it based on our error tolerance. The technical term is **coverage**.

Conformal Prediction

Conformal prediction provides a statistical guarantee:

"The correct answer is in the prediction set with probability at least 1-α."

$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \alpha$$

1-α can be thought of as the **success rate -** we choose it based on our error tolerance. The technical term is **coverage**.

Mistakes are reduced by making prediction sets larger (and therefore less useful).

Along with **statistical rigour**, conformal prediction is **versatile**, and **simple** to apply.

Vovk, Gammerman, Shafer. "Algorithmic Learning in a Random World". Springer 2005.
Shafer & Vovk. "A tutorial on conformal prediction". JMLR 2008

1 Define a conformal score function

$$s(x, y)$$

Larger scores indicate **worse** agreement between x and y. Classification example:

$$s(x, y) = 1 - f(x)_y$$

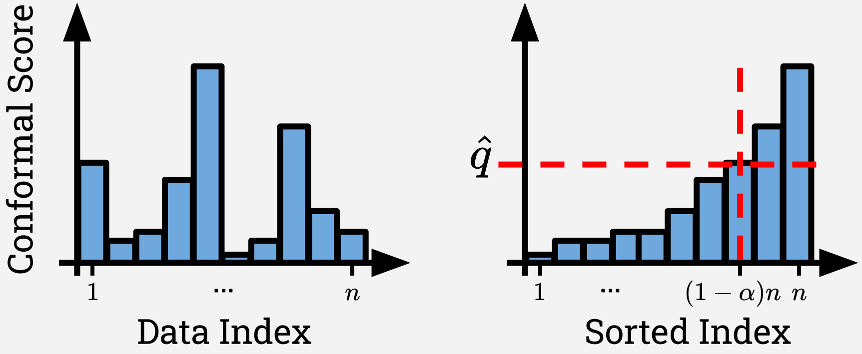
2 Given a calibration dataset

$$\{x_i, y_i\}_{i=1}^{n}$$

Compute conformal scores

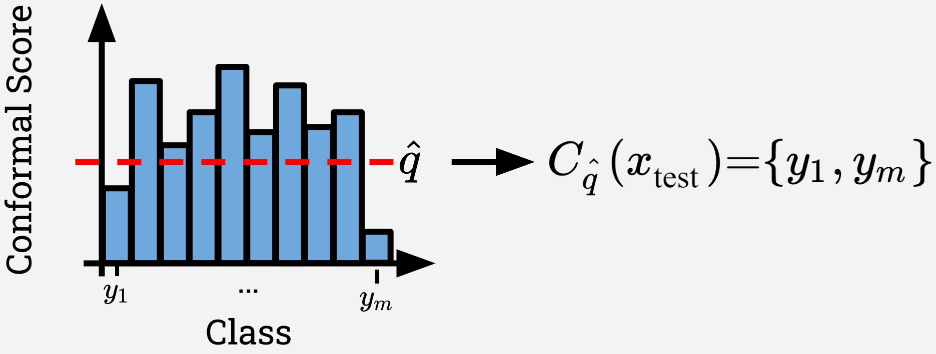
$$s_i = s(x_i, y_i)$$

3 Sort scores from low to high



Compute the $1 - \alpha$ quantile. This value of the score is called the conformal threshold $\hat{q}$

4 For new data x_{test}, return all classes with scores below the threshold



Larger set sizes indicate greater model uncertainty

layer 6
AI at TD

Conformal Prediction

1. Define a score function
2. Compute scores on calibration data
3. Find 1-α quantile - conformal threshold
4. On test data, return all classes with scores below threshold

Prediction sets constructed this way will satisfy coverage

$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \alpha$$

No assumptions on:

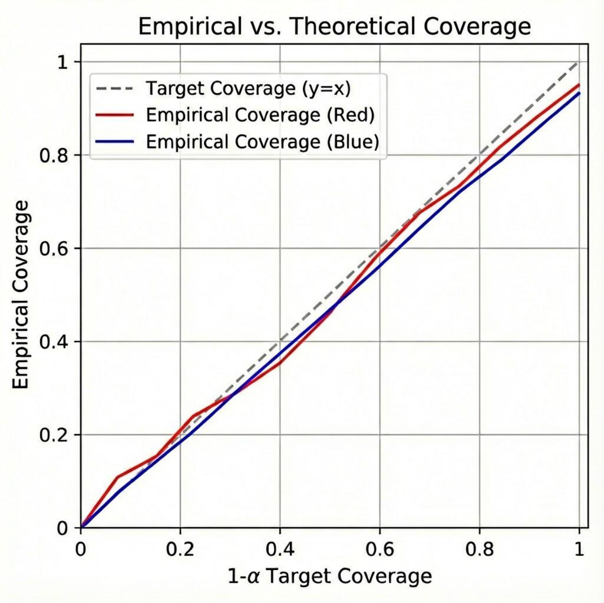
- the model architecture
- how it was trained
- the data distribution (Gaussianity)
- how many datapoints you have (infinite data regime)

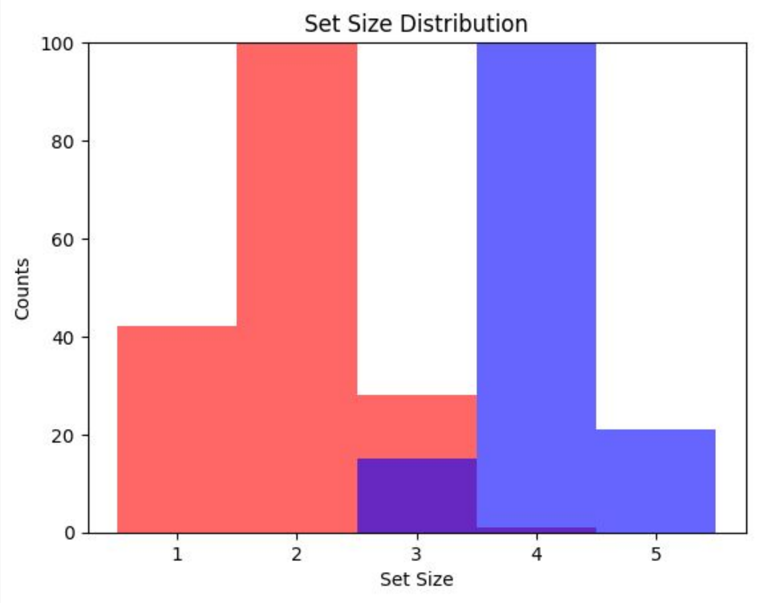
Angelopoulos & Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". arXiv:2107.07511

layer 6
AI at TD

Conformal Prediction

A conformal algorithm will give coverage (in expectation) for any score function, but score functions influence the quality of sets.





Conformalized Vision Tasks

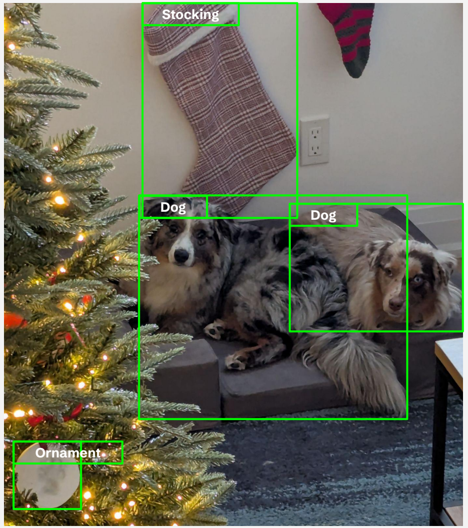
Conformal Prediction for Vision Tasks

Object Detection: Isolate and classify objects in an image

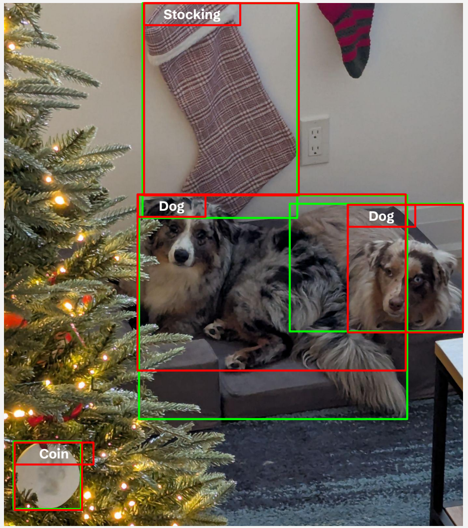
Produce a bounding box that *covers* the ground truth box, and give a label set that *covers* the ground truth label.

Larger bounding box relative to original prediction → more uncertain about location

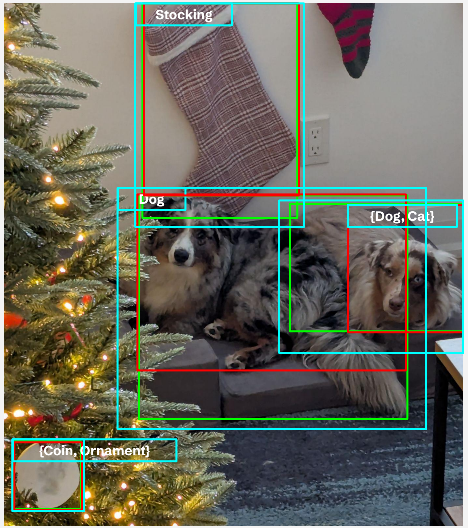
Ground Truth



Raw Prediction



Conformalized



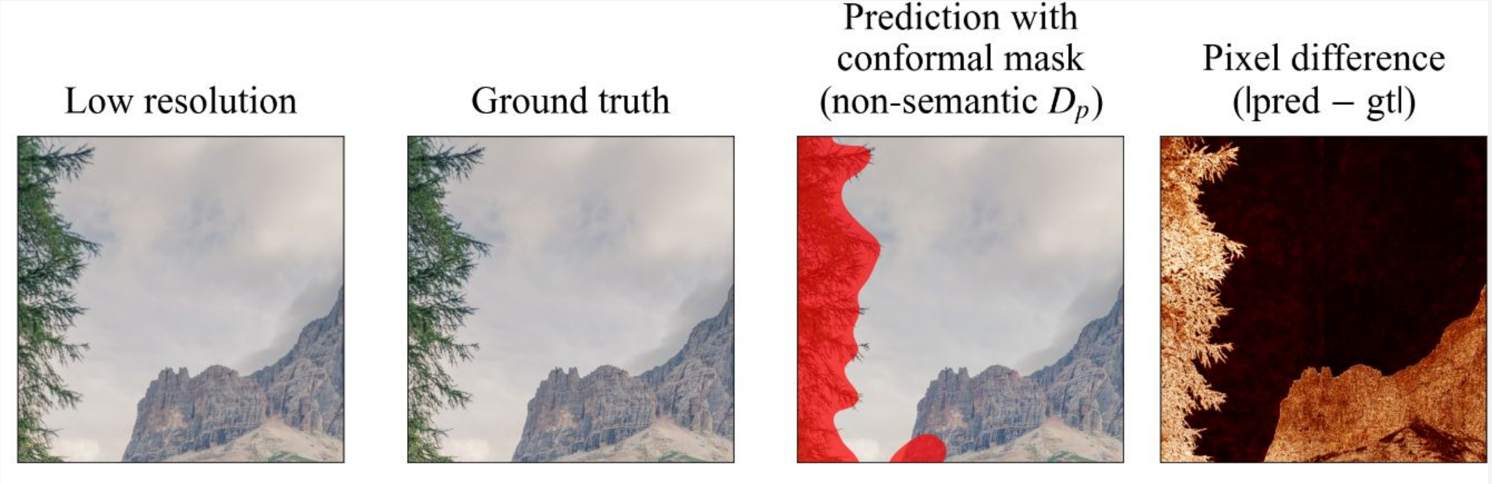
Grancey, Adam, et al. "Object Detection With Probabilistic Guarantees: a Conformal Prediction Approach". SAFECOMP 2022 Workshop.

layer 6
AI at TD

Conformal Prediction for Vision Tasks

Super-resolution: Reconstruct a high-res image from low-res inputs

Produce a mask over regions of low confidence in the reconstruction.
Masked regions should *cover* regions where the fidelity error is above α.



Masked regions → more uncertain about infill

Adame, Csillag, Goedert. "Image Super-Resolution with Guarantees via Conformal Generative Models". NeurIPS 2025.

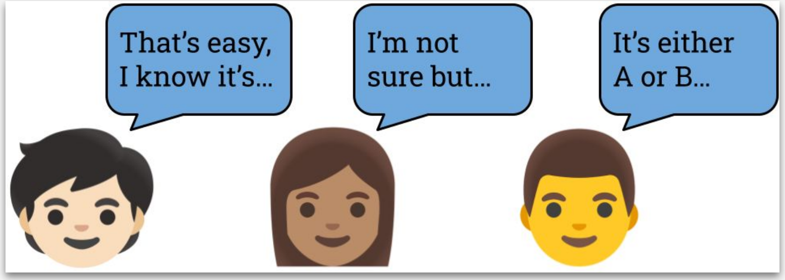
What do you do with prediction sets?

Actionable Model Outputs - Image Classification

Conformal prediction sets parallel how humans express uncertainty:

Sets signal model confidence through size.

When the model is less confident, it offers alternative.

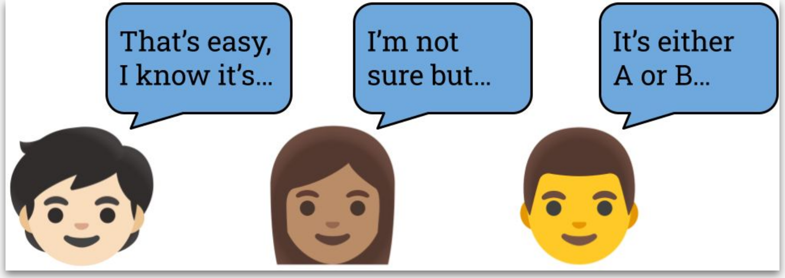


Actionable Model Outputs - Image Classification

Conformal prediction sets parallel how humans express uncertainty:

Sets signal model confidence through size.

When the model is less confident, it offers alternative.



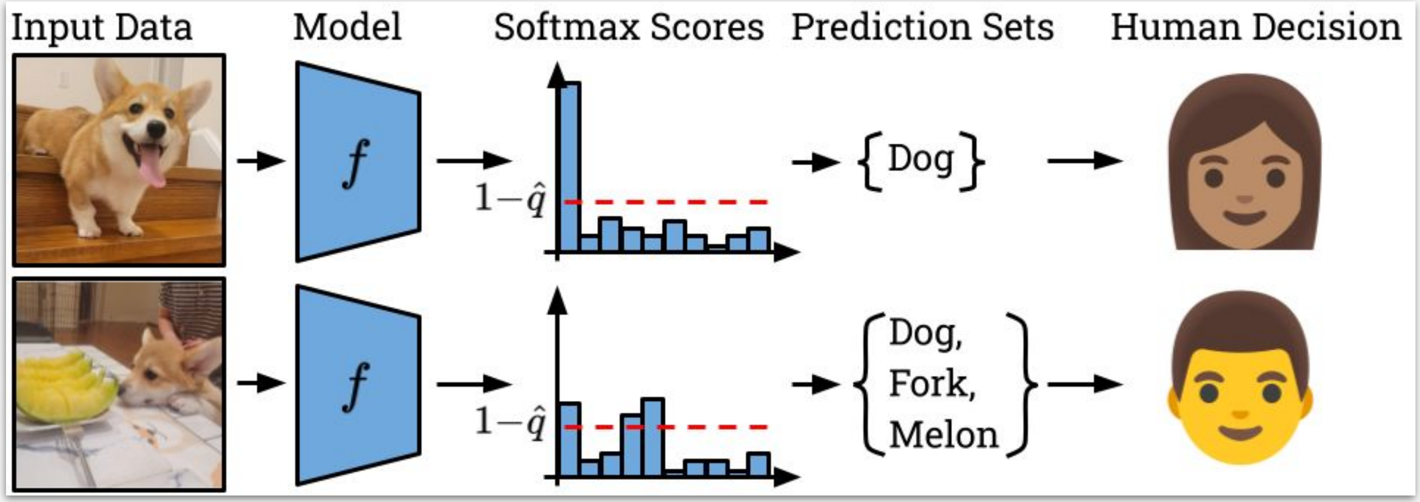
But, prediction sets are **not inherently actionable**.

Generally, we need a model to output a single prediction to automatically take an action.

Then how is conformal prediction meant to be used?

Human in the Loop Conformal Prediction

Since conformal prediction allows models to communicate in a more human way, it is natural to incorporate humans into conformal decision making pipelines.



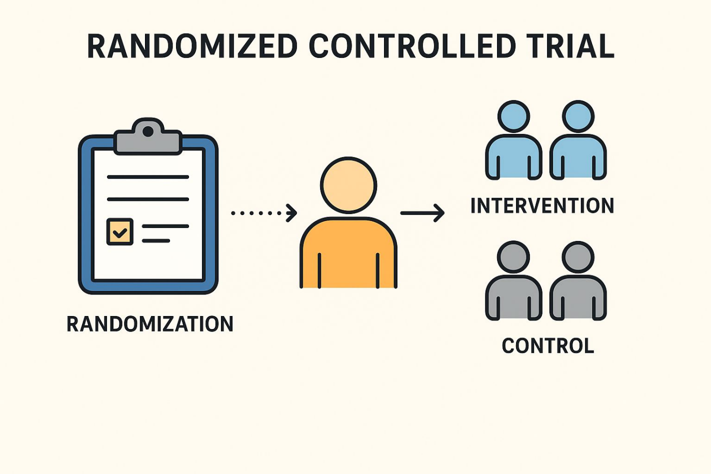
Do humans benefit from receiving conformal prediction sets?

layer 6 AI at TD

Human in the Loop Conformal Prediction

We designed and conducted a randomized controlled trial to test two things:

Do prediction sets improve human
- accuracy?
- speed?



JC, Sui, Kumar, Vouitsis. “Conformal Prediction Sets Improve Human Decision Making”. ICML 2024
Zhang, Chatzimparmpas, Kamali, Hullman. “Evaluating the Utility of Conformal Prediction Sets for AI-Advised Image Labeling” CHI 2024

layer 6 AI at TD

Test Design

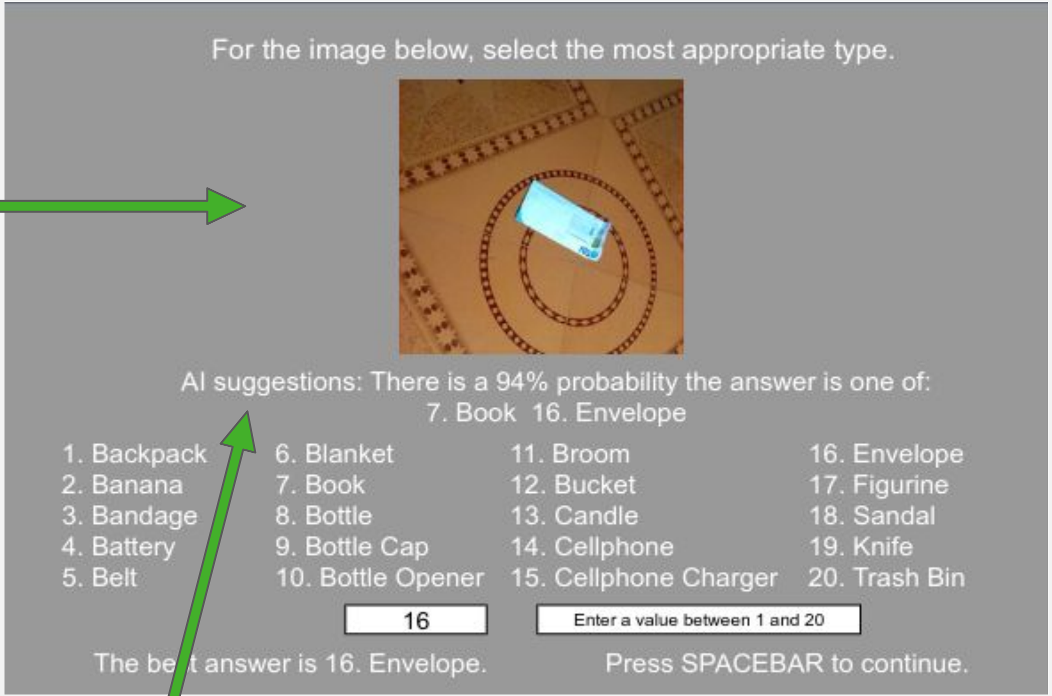
Human assigned to 1 of 3 tasks.

Tasks:

Image Classification “ObjectNet”

Sentiment Analysis
“Go-Emotions” - tweets

Named Entity Recognition
“Few-NERD” - Wikipedia

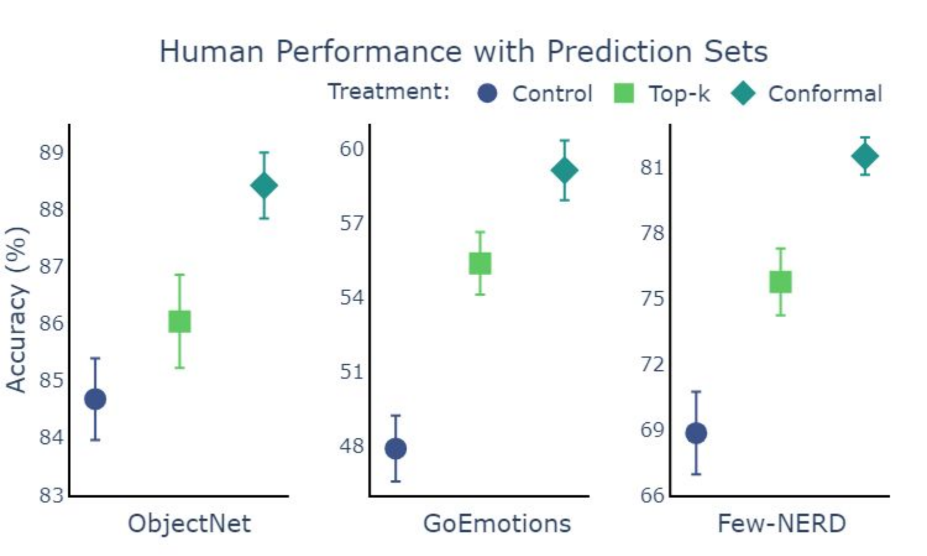


Treatment: Given either no help, top 3 model predictions, or conformal set (variable size)

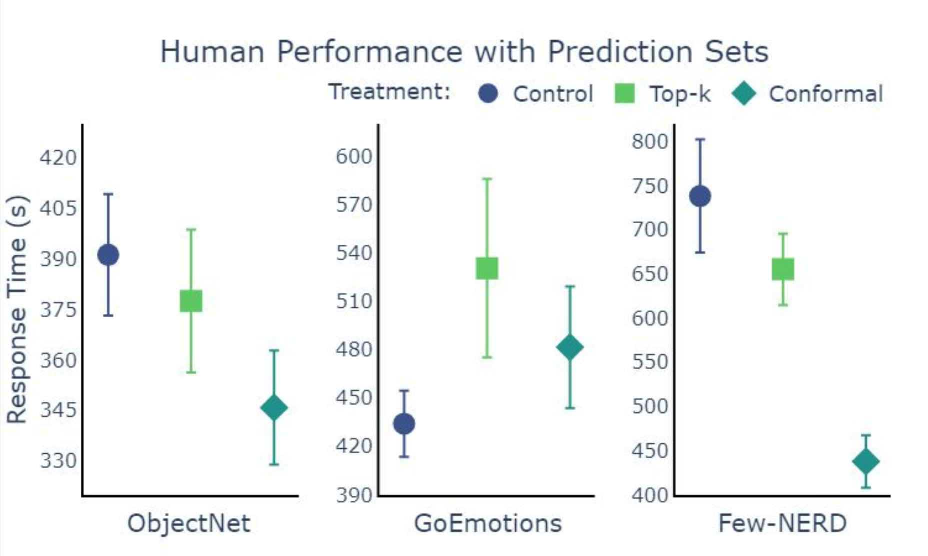
Main Results

Main results show that there is a **statistically significant and large difference in accuracy** between treatments, but no consistent trend for speed.

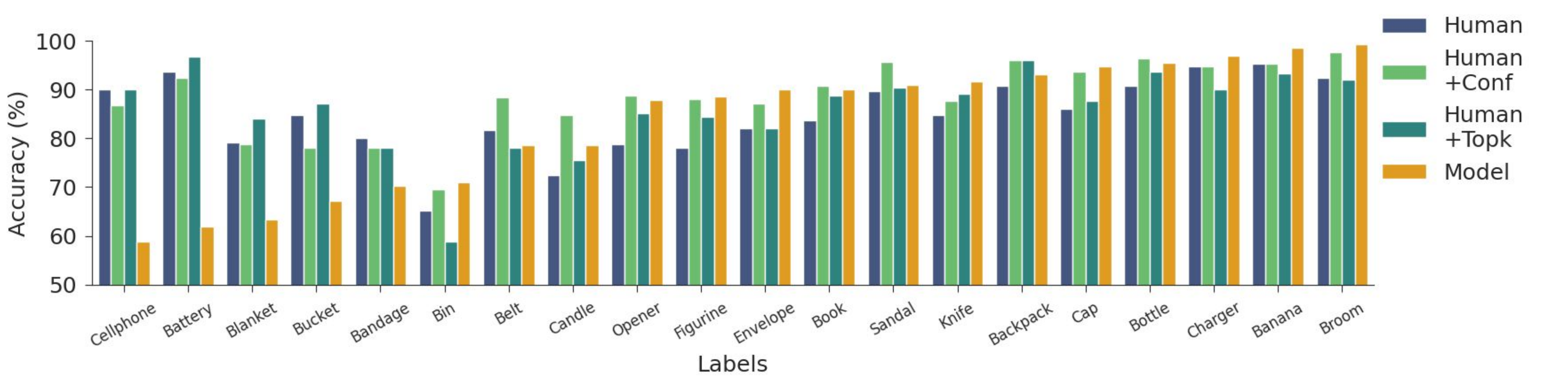
Accuracy



Speed



Fairness Concerns



Observation:

When model accuracy was **low,** Human+Conformal could be **worse** than Human.

When model accuracy was **high**, Human+Conformal was **better** than Human.

Conformal sets do not improve all classes equally - indicates Disparate Impact.

JC, Kumar, Sui, Belbahri. "Conformal Prediction Sets Can Cause Disparate Impact". ICLR 2025

Conformal Prediction + Fairness

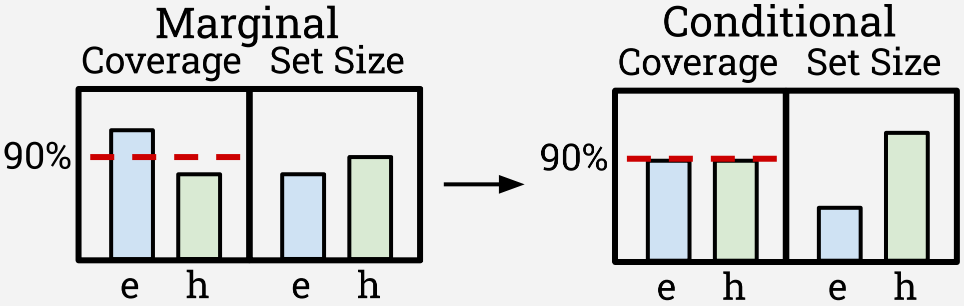
Suppose we have two groups within the data.
Coverage is valid *marginally*, not on each group *conditionally*.

How do you get *higher coverage on the undercovered group*?

Make their prediction sets **larger**

How do you get *lower coverage on the overcovered group*?

Make their prediction sets **smaller**



Under-covered

{Squirrel Monkey}

Add more labels to sets

{Squirrel Monkey, Spider Monkey, **Lemur**}

Over-covered

{Cruise Ship, **Container Ship**}

Remove labels from sets

{Cruise Ship}

Conformal Prediction + Fairness

Prior work argues that coverage should be equalized across groups for fairness.

Our experimental data shows that **set size matters more** for outcomes**.**

Equalizing ***Coverage*** *across groups makes CP less fair*

Equalizing ***Set Size*** *across groups makes CP more fair*

Hypothesis 1 - Prediction sets given to decision makers **can cause disparate impact** in the human's performance.

Hypothesis 2 - Sets with Equalized Coverage **will cause greater disparate impact** than marginal coverage.

JC, Kumar, Sui, Belbahri. "Conformal Prediction Sets Can Cause Disparate Impact". ICLR 2025
Romano, Barber, Sabatti, Candès. "With malice towards none: Assessing uncertainty via equalized coverage". HDSR 2020

layer 6
AI at TD

Test Design

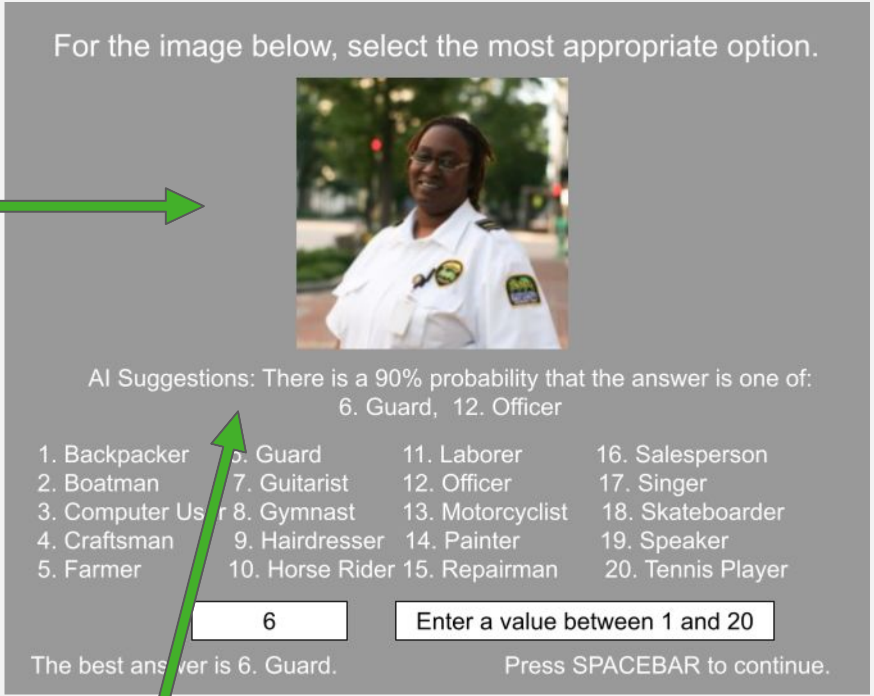
Human assigned to 1 of 3 tasks.

Tasks (Classification):

Image Classification “FACET”
Age = {Young, Middle, Old, Unknown}

Emotion Recognition “RAVDESS”
Gender= {Male, Female}

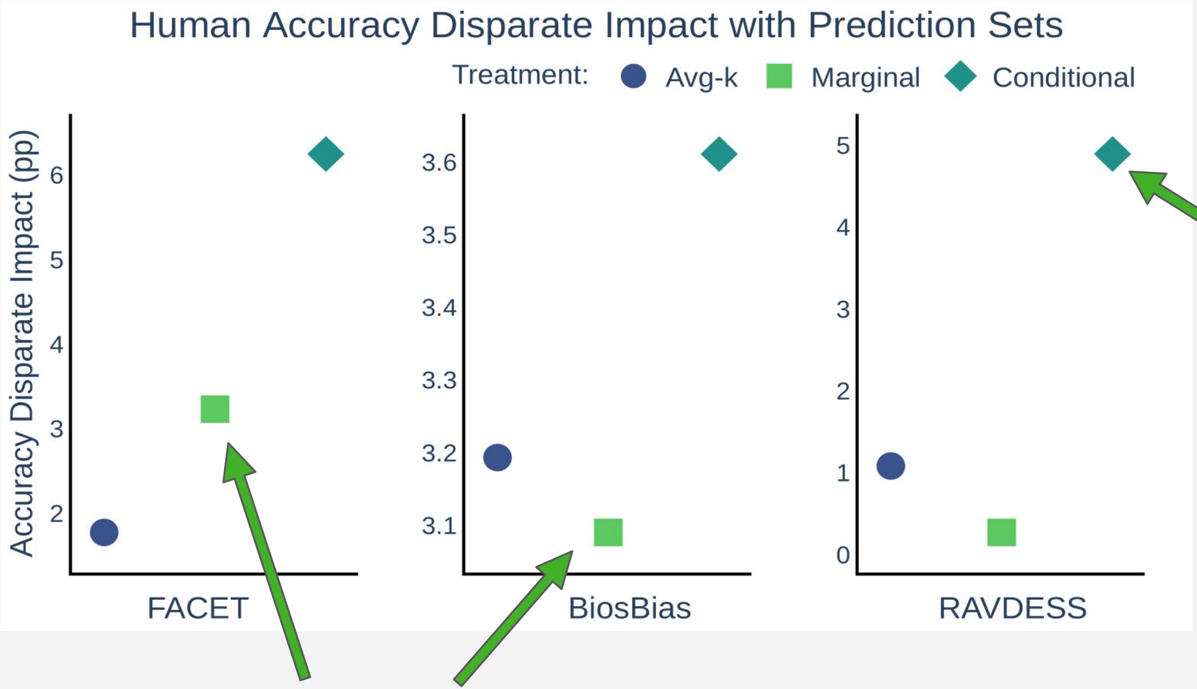
Text Classification “BiosBias”
Gender= {Male, Female}



Treatment: Given either no help, *marginal* conformal, *conditional* conformal

layer 6
AI at TD

Main Results



H2: Conditional sets caused much greater **DI** in all tests.

H1: Marginal sets caused some **DI** in 2 of 3 tests.

layer 6
AI at TD

Part I

Conformal Prediction as Uncertainty Quantification

Size of prediction sets indicates uncertainty

Coverage guarantee comes from calibration

Uncertainty is useful for downstream tasks

Part II

Conformal Prediction for Statistical Guarantees of Correctness

CP for Language Tasks

We can do classification, regression, set prediction. Straightforward.
How can CP be used for language tasks?

Uncertainty quantification for LLMs is an extremely important and unsolved problem.

- Hallucinations
- Abstention
- Probabilistic generation

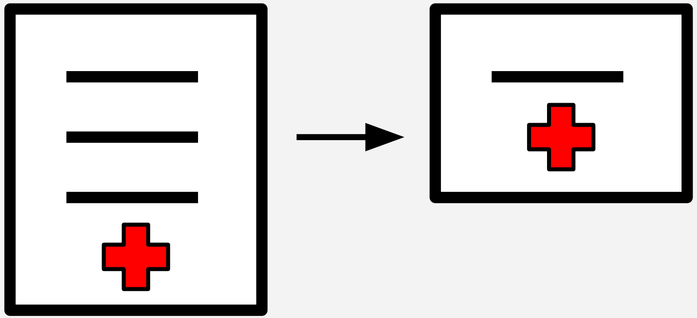
But it may not be obvious how to apply CP...
Should we generate multiple answers and return some of them as a set???

Document Summarization

Document Summarization

Summarization is an easy task nowadays - just give a document to an LLM.

But what if the document contains some critical information that must be retained?



LLM summarization gives you no guarantees that the summary

- Will contain all critical information
- Will not contain hallucinations

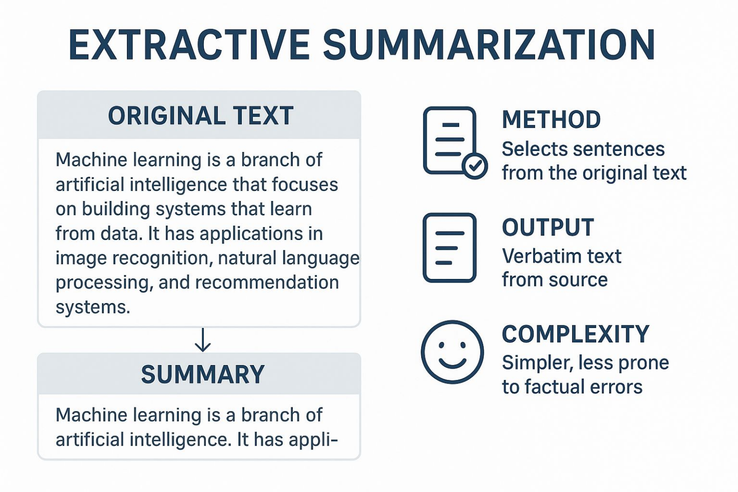
layer 6 AI at TD

Extractive Summarization with Guarantees

Extractive summarization does not paraphrase, it directly extracts phrases.

By performing extraction only, no hallucinations.

We can use the principles of conformal prediction to give statistical guarantees that critical information will be retained.



layer 6 AI at TD

Extractive Summarization with Guarantees

Given a document x, consisting of p sentences

$$x = \{c_1, \ldots, c_p\}$$

where a subset $y^* \subseteq x$ is GT important, we want to produce a summary y which contains all important information with high probability

$$\mathbb{P}[y^* \subseteq y] \geq 1 - \alpha$$

Kuwahara, Lin, Huang, Leung, Yapeter, Stanevich, Perez, JC. "Document Summarization with Conformal Importance Guarantees". NeurIPS 2025

layer 6
AI at TD

Extractive Summarization with Guarantees

Given a document x, consisting of p sentences

$$x = \{c_1, \ldots, c_p\}$$

where a subset $y^* \subseteq x$ is GT important, we want to produce a summary y which contains all important information with high probability

$$\mathbb{P}[y^* \subseteq y] \geq 1 - \alpha$$

We assign an "importance score" to each sentence, and filter out sentences based on a calibrated conformal threshold.

Shorter, more concise summaries are more helpful, so aim to retain few sentences.

Kuwahara, Lin, Huang, Leung, Yapeter, Stanevich, Perez, JC. "Document Summarization with Conformal Importance Guarantees". NeurIPS 2025

Conformal Importance - Calibration

1. Collect calibration data where each sentence has a binary label of importance.
2. Assign an “importance score” to each sentence (using a model).

Calibration Document #1

“Patient presented with heavy cough.”	GT=1	Importance Score = 0.7
“Patient was wearing blue socks.”	GT=0	Importance Score = 0.1
“Patient diagnosed with pneumonia.”	GT=1	Importance Score = 0.9

3. Set document-level conformal score as lowest importance score for any GT=1 sentence.

layer 6
AI at TD

Conformal Importance - Calibration

1. Collect calibration data where each sentence has a binary label of importance.
2. Assign an “importance score” to each sentence (using a model).

   Calibration Document #1 - Overall conformal score = **0.7**

“Patient presented with heavy cough.”	GT=1	Importance Score = **0.7**
“Patient was wearing blue socks.”	GT=0	Importance Score = 0.1
“Patient diagnosed with pneumonia.”	GT=1	Importance Score = 0.9

3. Set document-level conformal score as lowest importance score for any GT=1 sentence.

layer 6 AI at TD

Conformal Importance - Calibration

1. Collect calibration data where each sentence has a binary label of importance.
2. Assign an "importance score" to each sentence (using a model).

Calibration Document #1 - Overall conformal score = **0.7**

"Patient presented with heavy cough."	GT=1	Importance Score = **0.7**
"Patient was wearing blue socks."	GT=0	Importance Score = 0.1
"Patient diagnosed with pneumonia."	GT=1	Importance Score = 0.9

3. Set document-level conformal score as lowest importance score for any GT=1 sentence.
4. Find the 1-α quantile $\hat{q}$ of conformal scores over the calibration dataset.

Conformal Importance - Prediction

1. Assign an "importance score" to each sentence in the same way as before.

   Test Document

"Patient had a sandwich for lunch."	GT=?	Importance Score = 0.1
"Patient to take acetaminophen daily."	GT=?	**Importance Score = 0.9**
"Patient left hospital at 2:09 pm."	GT=?	Importance Score = 0.6
"Patient advised to not consume alcohol."	GT=?	**Importance Score = 0.8**

2. Sentences with importance greater than the threshold $\hat{q}$ are kept.

Theorem: Summaries created this way contain all important information with high probability

$$\mathbb{P}[y^* \subseteq y] \geq 1 - \alpha$$

layer 6
AI at TD

Conformal Importance - Scoring

Beyond proving mathematically that coverage holds, we proposed and evaluated various importance scoring functions.

LLM Scoring: Prompt an LLM to judge how important a sentence is from 0 to 1.

Embedding Similarity: Create sentence-level embeddings and compute distances between them to form a graph. Various NLP algorithms can compute a score based on the distance.

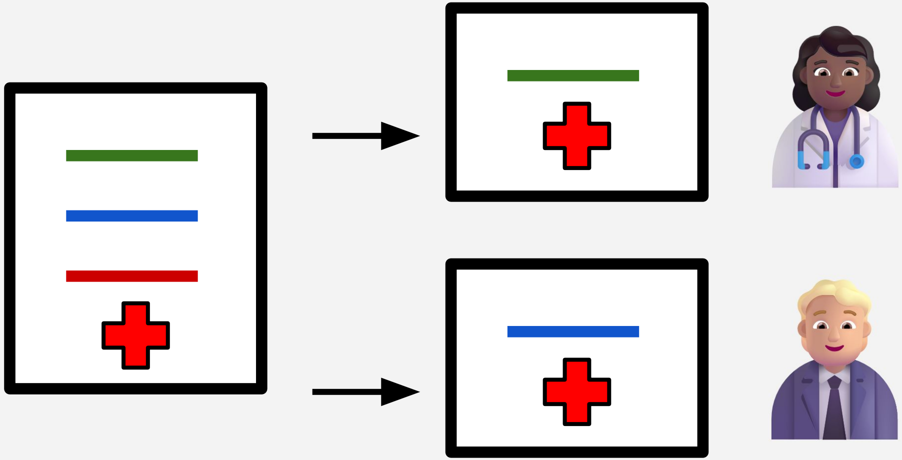
LLM scoring works best!

Importance Score	AUPRC ↑					Fraction of Sentences Removed ↑				
	ECT	CSDS	CNN/DM	SciTLDR	MTS	ECT	CSDS	CNN/DM	SciTLDR	MTS
Original Article	0.10	0.27	0.10	0.06	0.81	0.00	0.00	0.00	0.00	0.00
Cos. Sim. Centrality	0.22	0.34	0.34	0.35	0.86	0.22	0.11	0.18	0.29	0.18
Sentence Centrality	0.14	0.34	0.29	0.28	0.86	0.17	0.08	0.22	0.30	0.10
GUSUM	0.21	0.44	0.33	0.21	0.90	0.11	0.24	0.27	0.15	0.13
LexRank	0.22	0.43	0.32	0.32	*[3]	0.16	0.12	0.20	0.37	*[3]
GPT-4o mini (binary)	0.12	0.34	0.13	0.08	0.83	0.24	0.22	0.26	0.22	0.08
GPT-4o mini	0.30	0.49	0.34	0.33	0.93	0.24	0.25	**0.30**	0.40	0.16
Llama3-8B	0.18	0.39	0.22	0.15	0.92	0.13	0.11	0.14	0.11	0.14
Qwen3-8B	0.17	0.38	0.22	0.16	0.91	0.13	0.11	0.09	0.14	**0.22**
Gemini 2.0 Flash-Lite	0.35	0.68	**0.42**	**0.39**	**0.95**	0.28	0.46	0.25	0.40	0.13
Gemini 2.5 Flash	**0.43**	**0.69**	0.36	0.34	0.94	**0.37**	**0.49**	0.26	**0.41**	0.14

layer 6
AI at TD

Conformal Importance - Customization

Two users may have different opinions on what is important



Doctor: Diagnoses, Medications, ...

Admin: Length of Stay, Costs, ...

Conformal Importance can accommodate different opinions by

- Having each user annotate their own calibration set
- Defining what is considered important in the LLM scoring prompt (e.g. ICL)

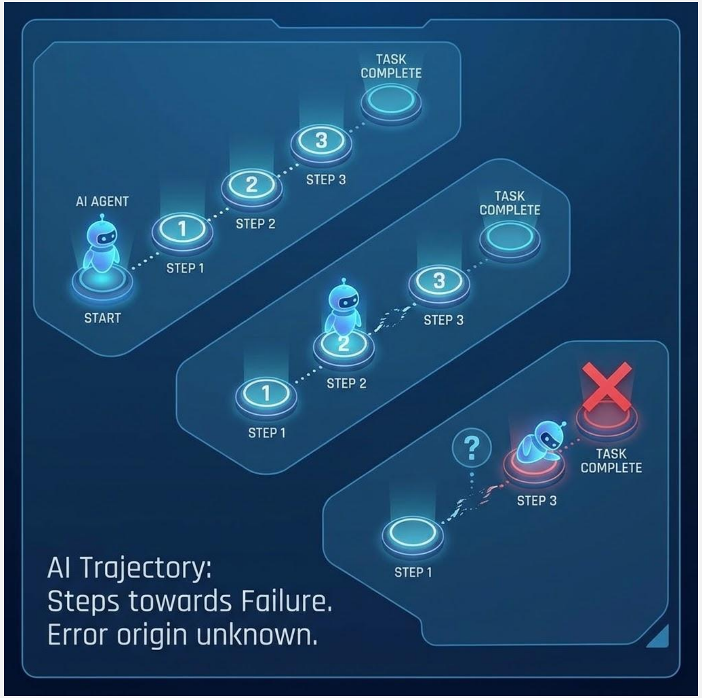
layer 6
AI at TD

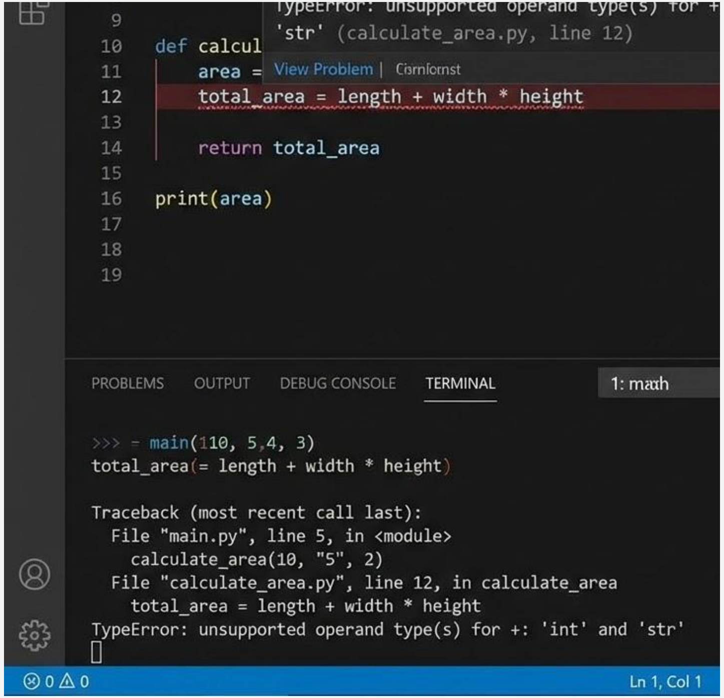
Agent Error Attribution

Agent “Debugging”

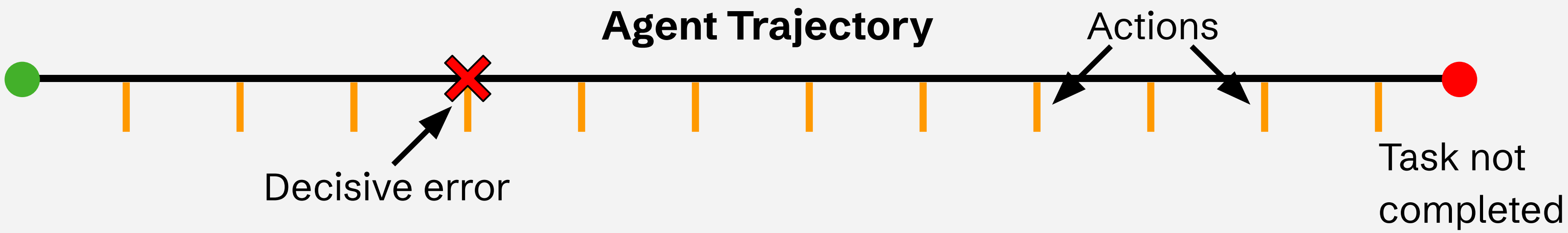
When an AI agent fails at a task, how can we determine what went wrong?

Unlike traditional code, we do not see exactly which step caused the error.





Agent Error Attribution

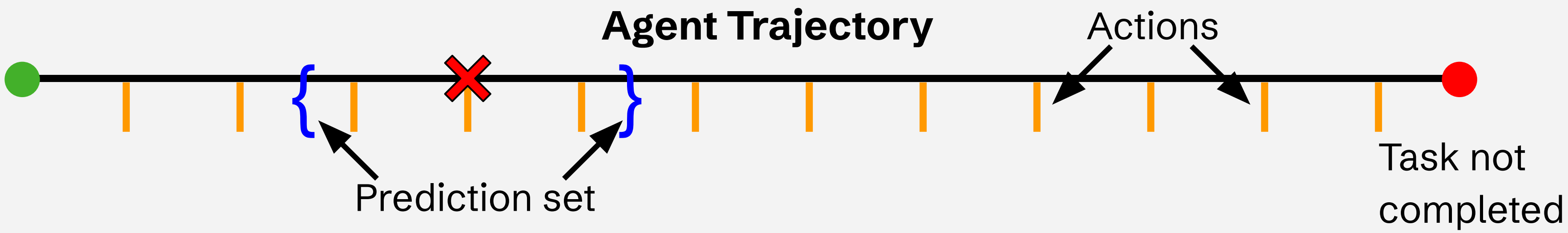


Imagine a web-shopping agent with a task to purchase formal black shoes. At each step it can observe a textual webpage and perform actions like clicking a mouse or typing.

Classifying exactly which step was the "decisive error" has proven to be challenging, with ~9% accuracy rates in first studies.

Zhang, Yin, et al. "Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems". ICML 2025

Agent Error Attribution



Instead, we aim to predict a (contiguous) set of steps which contains the error with high probability:

$$\mathbb{P}[y^* \subseteq y] \geq 1 - \alpha$$

Zhang, Yin, et al. "Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems". ICML 2025

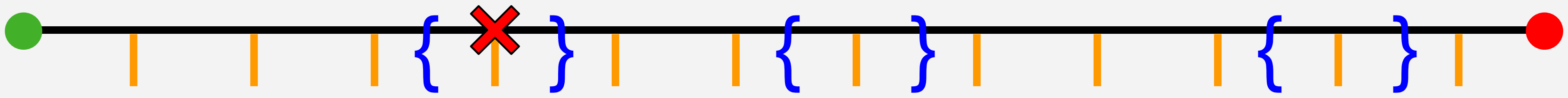
Conformal Error Attribution

This setting perfectly demonstrates the two main components of conformal systems:

1. The algorithm producing sets with a coverage guarantee,
2. The scoring function, a model of the predictive task.

Hou, Feng, Sui, Ga, JC. “Conformal Agent Error Attribution”. To appear Feb 2026

Conformal Error Attribution - Algorithms



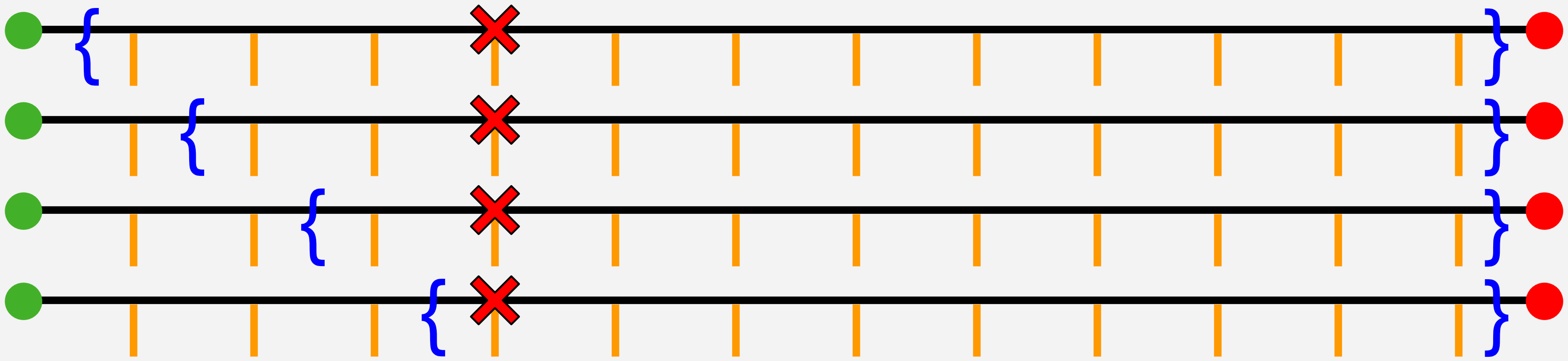
Ordinary conformal prediction for classification can be applied:

1. Define score function - how likely is a step to be the decisive error?
2. Compute scores on calibration data.
3. Find 1-α quantile.
4. On test data, add all steps with high enough conformal score.

But this does not have the property that predicted sets be contiguous.
We are not taking advantage of the data's structure.

Hou, Feng, Sui, Ga, JC. "Conformal Agent Error Attribution". To appear Feb 2026

Conformal Error Attribution - Algorithms



Left Filtering:

1. Define score function - how likely is decisive error to be in a subsequence.
2. Filter out steps from the left that are not the decisive error. Record final score.
3. Find 1-α quantile.
4. On test data, filter out steps until just before it drops below calibrated threshold.

Hou, Feng, Sui, Ga, JC. "Conformal Agent Error Attribution". To appear Feb 2026

Conformal Error Attribution - Algorithms

Root-to-Leaf Tree Traversal:

1. Define score function - how likely is decisive error to be in a subsequence.
2. View the trajectory as a tree - entire trajectory = root - single step = leaf.
3. Start with single most likely leaf. Traverse up tree until decisive error contained.
4. Find 1-α quantile.
5. On test data, traverse up tree until score surpasses calibrated threshold.

Hou, Feng, Sui, Ga, JC. "Conformal Agent Error Attribution". To appear Feb 2026

Conformal Error Attribution - Score Functions

We need to assign a score to a subsequence.
You essentially need to use an LLM to handle the textual information of the trajectory:

- Task statement
- Final output (failure)
- Agent's chosen action/tool at subsequence steps
- Response from environment at subsequence steps
- Any other relevant metadata.

Hou, Feng, Sui, Ga, JC. "Conformal Agent Error Attribution". To appear Feb 2026

Conformal Error Attribution - Score Functions

All-at-Once:
Provide all the information in a single LLM call.

Pros:

Minimal LLM calls

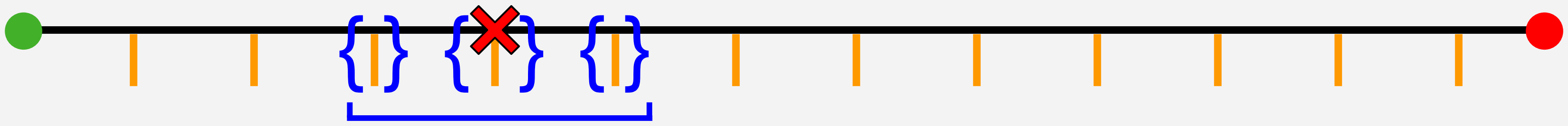
Cons:

LLM can be overwhelmed by amount of detail

Scores for nested subsequence may not be monotonic

Hou, Feng, Sui, Ga, JC. “Conformal Agent Error Attribution”. To appear Feb 2026

Conformal Error Attribution - Score Functions



Aggregated Scoring:

Score each step in a subsequence individually, then aggregate the results.

E.g. **Sum** the individual scores; or take the **maximum** over scores.

Pros:

- LLM can focus on details of each step.
- Scores for nested subsequence can be monotonic by design.

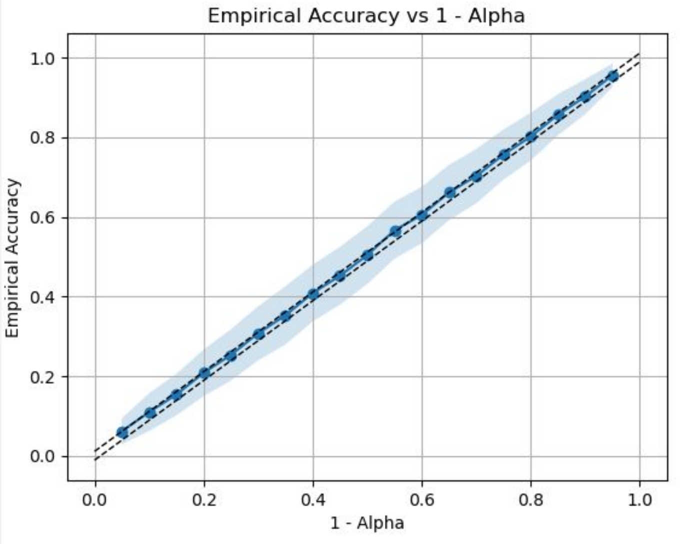
Cons:

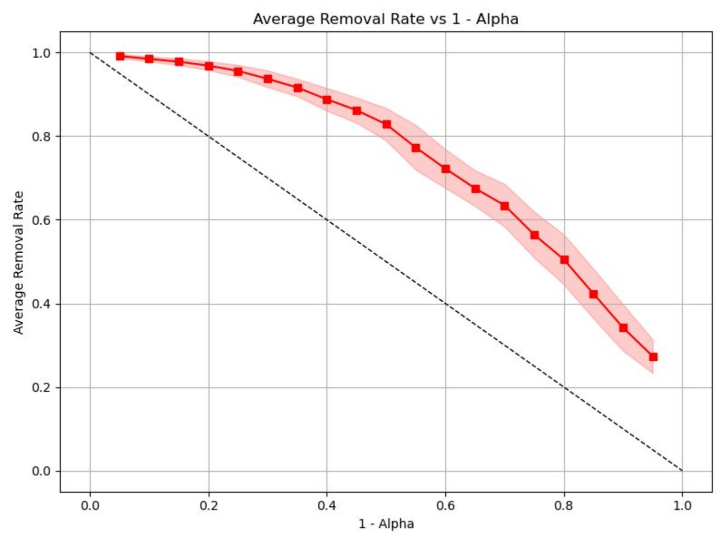
- Several LLM calls needed.

Hou, Feng, Sui, Ga, JC. “Conformal Agent Error Attribution”. To appear Feb 2026

Conformal Error Attribution

Can achieve valid statistical guarantees of coverage while narrowing down the prediction set to 20-30% of total steps.





Hou, Feng, Sui, Ga, JC. "Conformal Agent Error Attribution". To appear Feb 2026

Conclusions

layer 6
AI at TD

Conclusions

Conformal prediction is not just a method for generating sets, or quantifying uncertainty.

It is a flexible framework for providing statistical guarantees on various forms of correctness. Tight guarantees hold with minimal assumptions.

By being creative with how score functions are defined, we can adapt conformal prediction to new settings, like document summarization, and agent evaluation.

Layer 6 is hiring for

- Research Machine Learning Scientists
- Machine Learning Engineers
- Technical Product Owners

Visit layer6.ai and chat with me at the break!

layer 6
AI at TD

Thank you!

References

[1] Vovk, Gammerman, and Shafer. “Algorithmic Learning in a Random World”. Springer 2005.

[2] Shafer & Vovk. “A tutorial on conformal prediction”. JMLR 2008

[3] Angelopoulos & Bates. “A gentle introduction to conformal prediction and distribution-free uncertainty quantification”. arXiv:2107.07511

[4] Grancey, Adam, et al. “Object Detection With Probabilistic Guarantees: a Conformal Prediction Approach”. SAFECOMP 2022 Workshop.

[5] Adame, Csillag, and Goedert. “Image Super-Resolution with Guarantees via Conformal Generative Models”. NeurIPS 2025.

[6] Cresswell, Sui, Kumar, and Vouitsis. “Conformal Prediction Sets Improve Human Decision Making”. ICML 2024

[7] Zhang, Chatzimparmpas, Kamali, and Hullman. “Evaluating the Utility of Conformal Prediction Sets for AI-Advised Image Labeling” CHI 2024

[8] Cresswell, Kumar, Sui, and Belbahri. “Conformal Prediction Sets Can Cause Disparate Impact”. ICLR 2025

[9] Romano, Barber, Sabatti, and Candès. “With malice towards none: Assessing uncertainty via equalized coverage”. HDSR 2020

[10] Kuwahara, Lin, Huang, Leung, Yapeter, Stanevich, Perez, and Cresswell. “Document Summarization with Conformal Importance Guarantees”. NeurIPS 2025

[11] Zhang, Yin, et al. “Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems”. ICML 2025

layer 6
AI at TD