

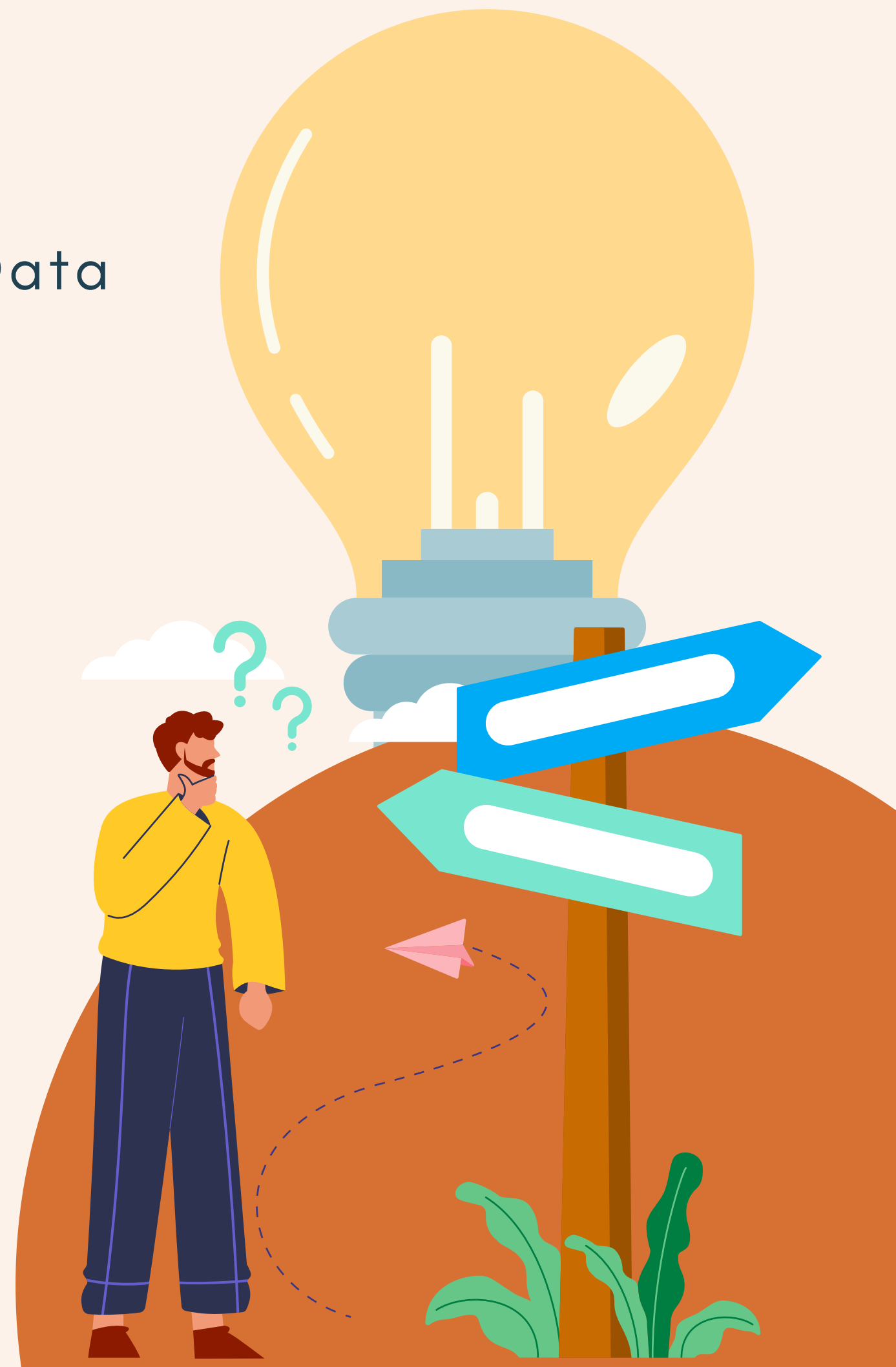
University of Central Florida

CAP6545 – Machine Learning for Biomedical Data

CONFORMAL PREDICTION SETS IMPROVE HUMAN DECISION MAKING

Authors: Jesse C. Cresswell, Yi Sui, Bhargava Kumar, Noel Vouitsis

Presented by: Hieu Chu



Humans Signal Uncertainty, Machines Often Don't

- People often express uncertainty, explain reasons when doubtful, and offer alternative solutions when unsure.
- Most machine learning models output a **single answer** without indicating uncertainty.
- This limits their usefulness in real-world decision-making tasks.

Existing Approaches and Their Limitations

Existing methods

- ! Standard ML models provide single outputs with no uncertainty measurement.
- ! Top-k prediction methods offer fixed alternative outcomes but lack confidence estimation.
- ! Prior methods do not effectively communicate uncertainty to human decision-makers.

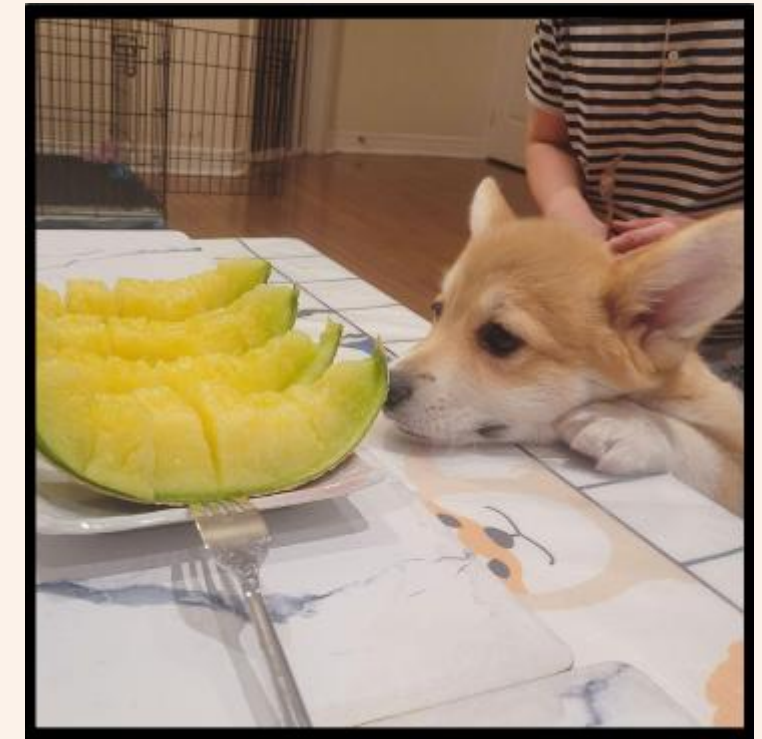
Introduced Solution

Conformal Prediction Sets
mimic human uncertainty
expression by providing
multiple possible outcomes
and signaling confidence
through set size.

What Are Conformal Prediction Sets?

Guarantee accuracy with variable prediction size sets

- Conformal prediction generates a set of possible predictions rather than a single answer with a guarantee of accuracy.
- Provides a formal coverage guarantee (e.g., 95% certainty that the true answer is within the set).
- The size of the prediction set reflects the uncertainty of the model:
 - Smaller sets → High confidence
 - Larger sets → Low confidence



{Dog, Melon}

{Dog, Fork, Melon, Tablecloth, T-shirt}

Research Purpose and Methodology

Do Prediction Sets Help Humans Make Better Decisions?

Research Goal:

Investigate whether providing humans with conformal prediction sets can improve their accuracy and response time in decision-making compared to traditional methods.

Metrics Measured:

- **Accuracy:** Did participants select the correct answer?
- **Response Time:** How quickly did participants respond accuracy?

Participants completed three tasks:

Image Classification, Sentiment Analysis and Named Entity Recognition

Dataset and pre-trained model was used

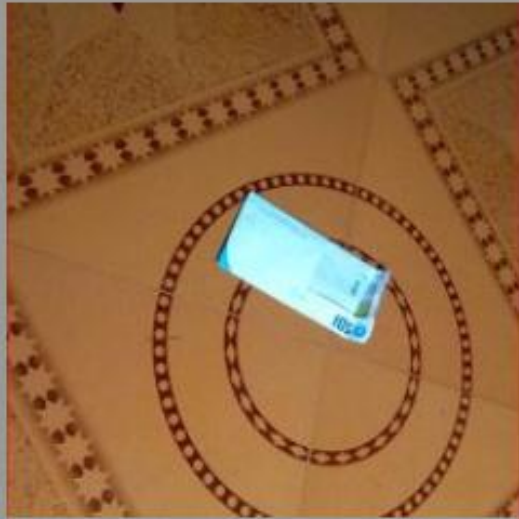
- **Image Classification - ObjectNet:** A challenging image classification dataset of common objects with different angle. The 20 most common classes were selected. Pre-trained model is **CLIP ViT-L/14**.
- **Sentiment Analyst - GoEmotions:** A sentiment/emotion classification dataset from Reddit comments, using 10 commonly occurring emotions. Pre-trained model is fine-tuned **RoBERTa-Base**.
- **Named Entity Recognition - Few-NERD:** A dataset of sentences from Wikipedia with named entity annotations. The 10 most common classes were selected. Pre-trained model is fine-tuned **SpanMarker RoBERTa-Large**.
- **Data Format:** The datasets were pre-processed and split into calibration (Dcal) and test (Dtest) sets. The calibration dataset is used to compute the conformal threshold, while the test dataset is shown to humans along with the prediction sets.

Experimental setups

Testing Levels of Machine Assistance

- **Control:** Humans make decisions independently, without any machine-generated aid or suggestions.
- **Top-k Prediction Sets:** Participants receive a fixed k-number of alternative predictions from the model without reflecting confidence levels.
- **Conformal Prediction:** Participants are provided with variable-sized prediction sets, where the size of the set reflects the model's uncertainty.

For the image below, select the most appropriate type.



AI suggestions: There is a 94% probability the answer is one of:
7. Book 16. Envelope

1. Backpack	6. Blanket	11. Broom	16. Envelope
2. Banana	7. Book	12. Bucket	17. Figurine
3. Bandage	8. Bottle	13. Candle	18. Sandal
4. Battery	9. Bottle Cap	14. Cellphone	19. Knife
5. Belt	10. Bottle Opener	15. Cellphone Charger	20. Trash Bin

The best answer is 16. Envelope. Press SPACEBAR to continue.

How does Conformal Prediction work?

Conformal Prediction works on all ML model

- **Compute Conformal Scores:** Apply the trained model to **calibration set** and calculate conformal scores for each sample.
- **Determine Error Threshold:** Sort the conformal scores and select a threshold that ensures the desired coverage (e.g., 90%).
- **Generate Prediction Set:** For a new input, include all labels whose conformal scores fall below the threshold.
- **Guarantee Coverage:** The prediction set will contain the true label with at least the desired coverage probability, ensuring reliable uncertainty quantification.

How does Conformal Prediction work?

Simple Example

Classifying an image into one of four categories: Dog, Melon, Fork, Chair.

- Apply the trained model to the **calibration set**, Compute Conformal Scores ($1 - \text{softmax}$) for the **true label** and sort the score: [0.05, 0.10, 0.12, 0.15, 0.20, 0.25, ..., 0.30, 0.35, 0.40, 0.45, ...]
- **Determine Error Threshold (90% Coverage):** Threshold at 90% = 0.45
- Model outputs probabilities with test set (softmax): Dog: 0.80 - Melon: 0.7 - Fork: 0.52 - Chair: 0.05
- Compute Conformal Scores: Dog: 0.20 - Melon: 0.3 - Fork: 0.48 - Chair: 0.95
- Conformal prediction sets: {Dog, Melon} with 90% confidence

$$\mathcal{C}_{\hat{q}}(x) = \{y \mid s(x, y) < \hat{q}\}.$$

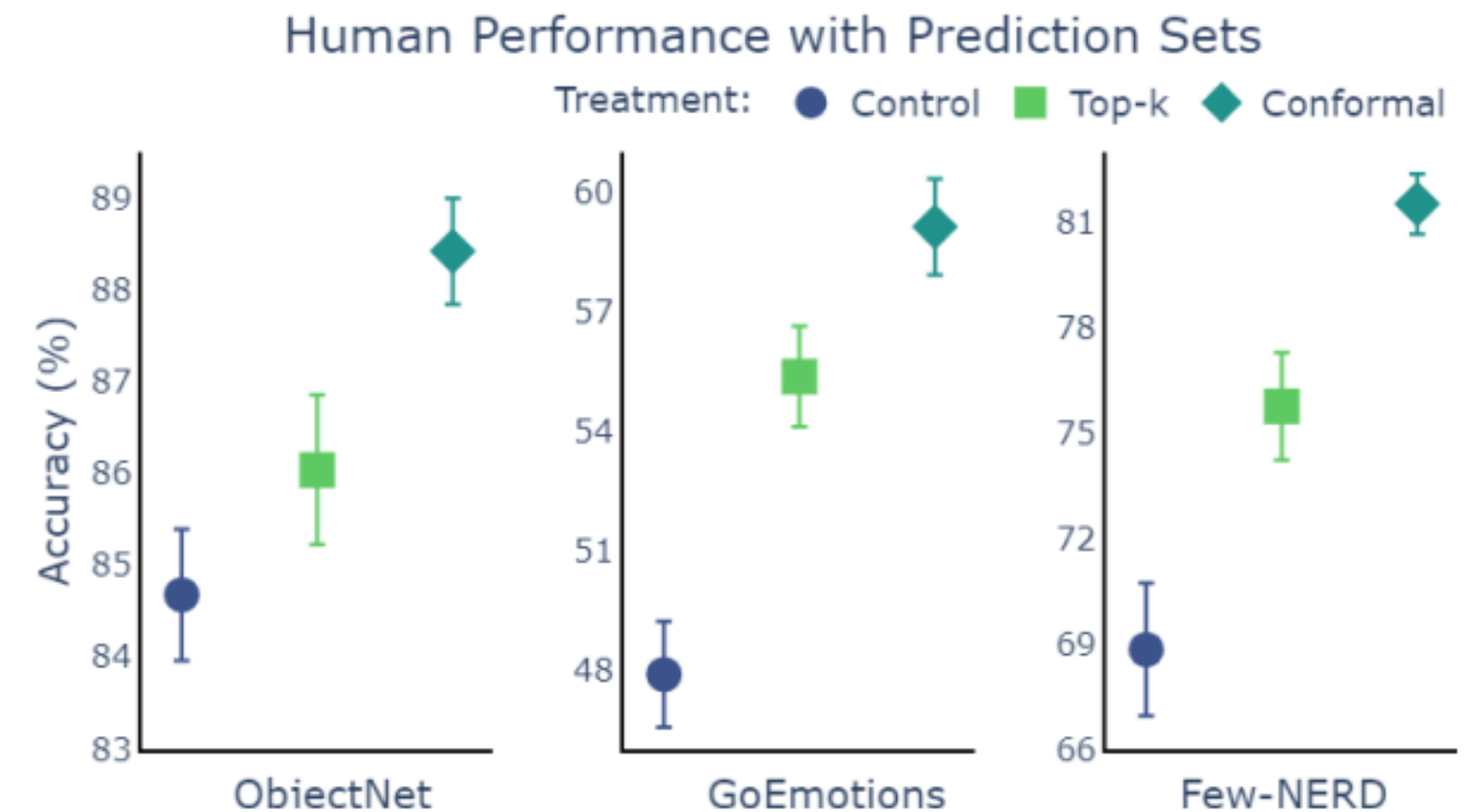
Boosting Decision Accuracy with Conformal Sets

Consistent Performance Improvement:

- Human participants using **conformal prediction sets** consistently outperformed those using top-k sets or no assistance.

Higher Accuracy Across Tasks:

- Across all three tasks—**ObjectNet** (image classification), **GoEmotions** (sentiment analysis), and **Few-NERD** (entity recognition)—conformal sets led to significant improvements in decision accuracy.



Speed and Trust Trade-offs with Conformal Sets

Faster Responses with Confident Predictions:

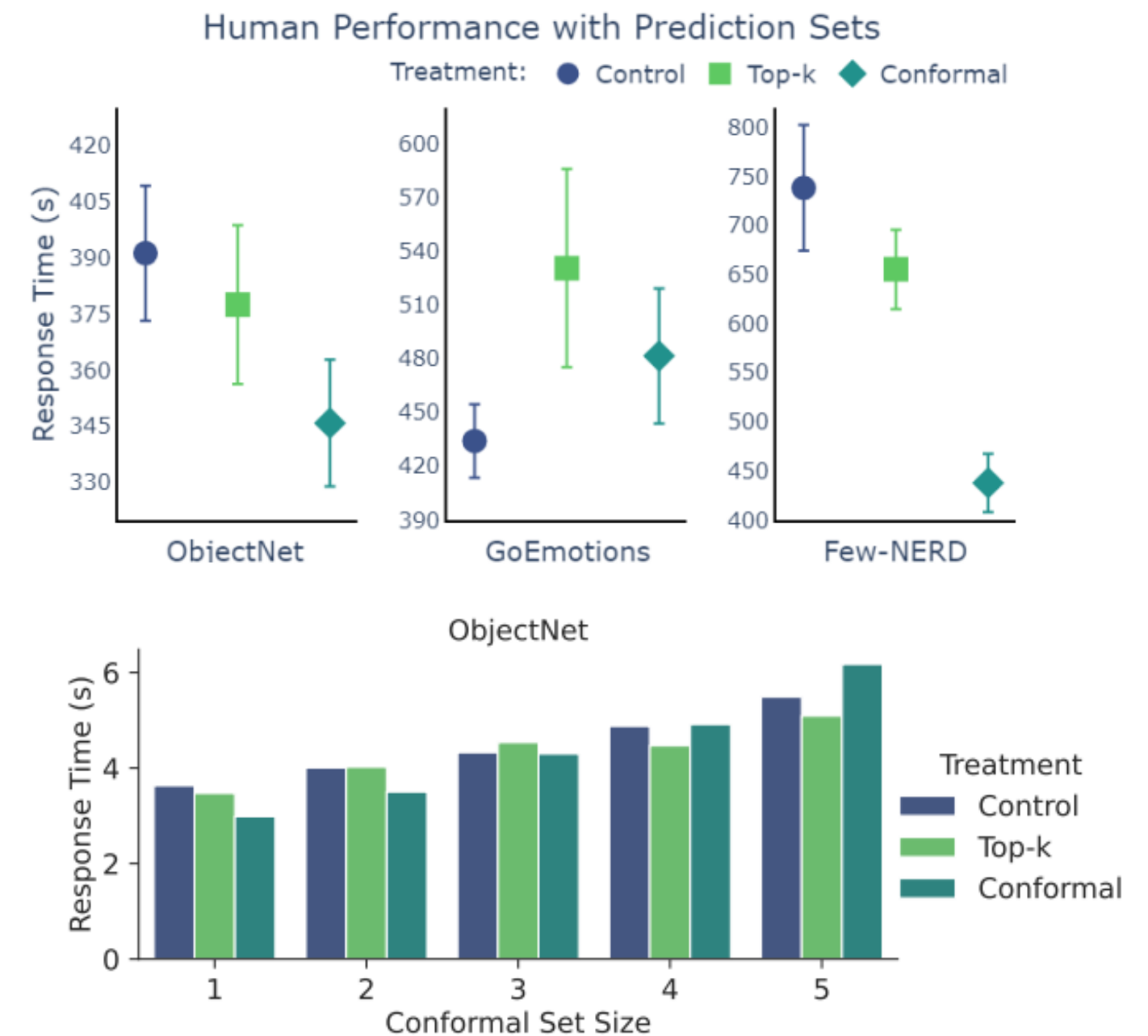
Conformal sets often led to quicker decisions when predictions were highly confident, resulting in smaller sets with fewer options to consider.

Cognitive Load Effects:

When predictions were uncertain, larger sets increased the cognitive load, occasionally slowing down response times due to additional information processing.

Trust in Coverage Guarantees:

Participants consistently relied on the provided coverage levels (e.g., 94%), aligning their final decisions with the prediction set's confidence.



Strengths and Limitations of Conformal Prediction

Advantages:

- **Improved Accuracy:** Enhances human decision-making compared to top-k sets.
- **Uncertainty Quantification:** Provides reliable measures of model uncertainty through set sizes.
- **Risk Control:** Allows users to predefine acceptable error rates.
- **Broad Applicability:** Compatible with any pre-trained model.

Disadvantages:

- **Cognitive Load:** Larger sets can increase mental effort during decision-making.
- **Bias Transfer Risk:** Inherited model biases can influence human decisions.
- **Human Oversight Needed:** Requires human for final decision actionable outcomes.

Where Can This Research Go Next?

Adaptive Prediction Sets Based on User Expertise

- Simplify choices for expert users (e.g., top 3 predictions) while displaying the **coverage guarantee**.
- Provide larger prediction sets for novices to support comprehensive decision-making.
- Allow users to adjust set sizes based on comfort and task complexity.

Testing in High-Stakes and Ethical Contexts

- **Medical Diagnosis:** Cancer detection, heart disease risk.
- **Legal Decision Support:** Recidivism prediction, evidence review.
- **Financial Risk Management:** Fraud detection, risk scoring.
- Evaluate fairness, bias transfer, and the impact on human trust in critical fields.