

Towards Useful AI Interpretability for Humans via Interactive AI Explanations

Hua Shen

 @huashen218  huashen218@psu.edu

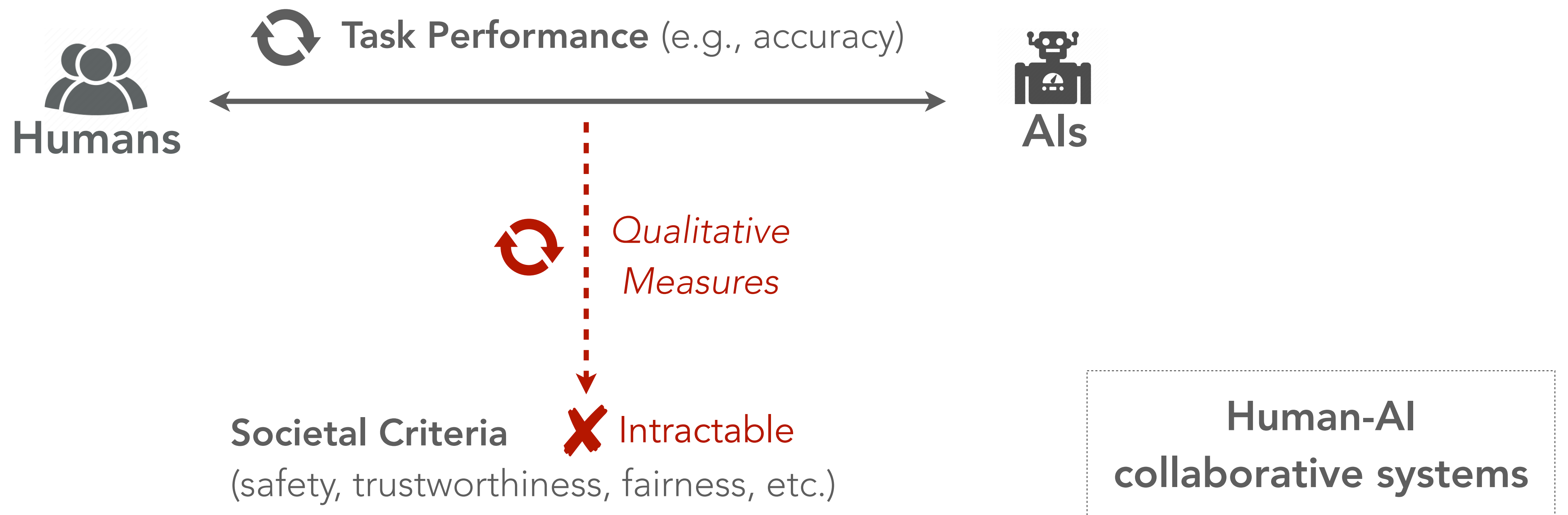
Penn State University

Background & Motivation



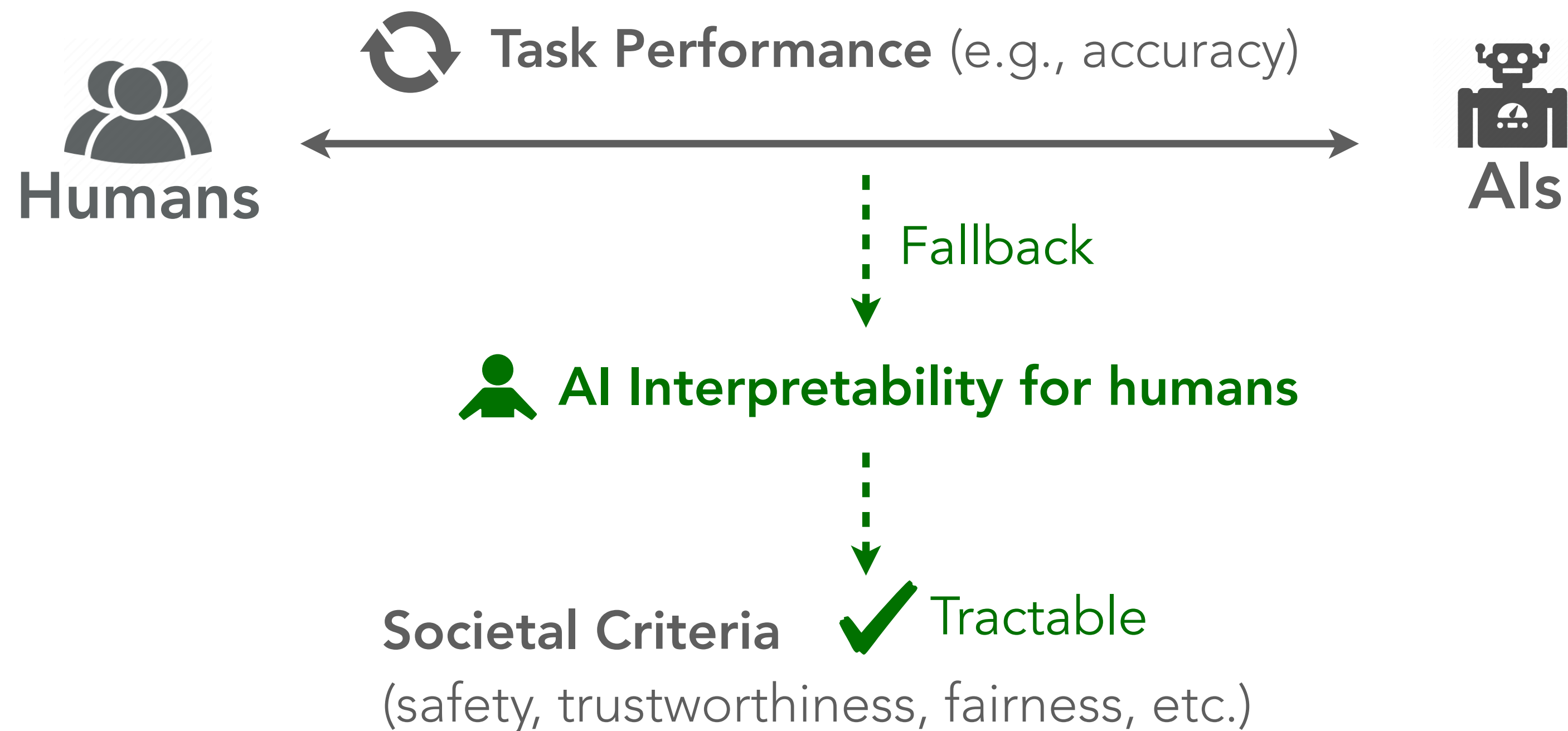
Why do we need AI interpretability?

Human-AI collaborative systems are not only **optimized** for **task performance** (e.g., accuracy), but also are required to **satisfy** vital **societal criteria** (e.g., trustworthiness, safety, fairness, etc.).

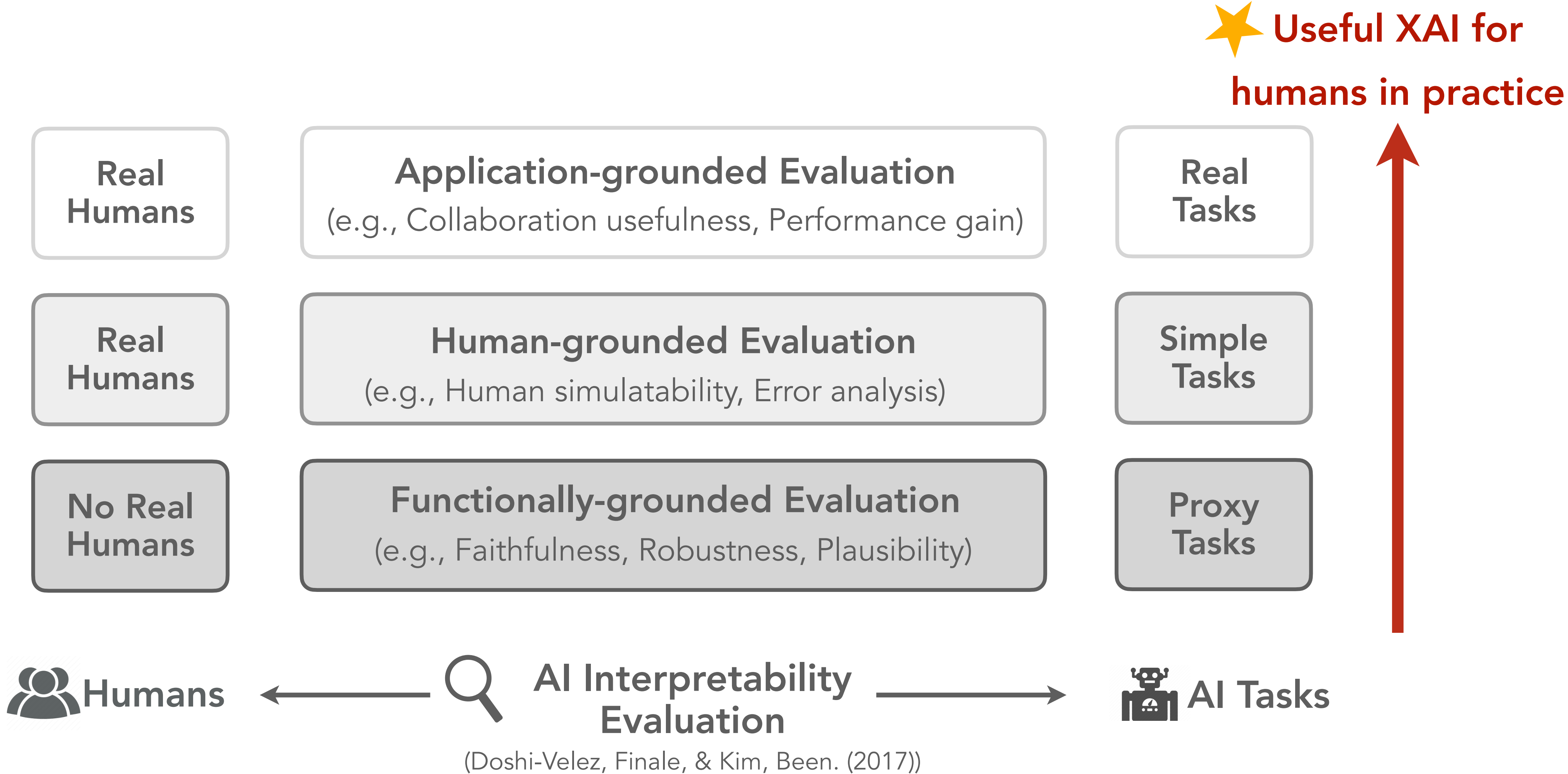


The **usefulness** of **XAI** for humans is **crucial**

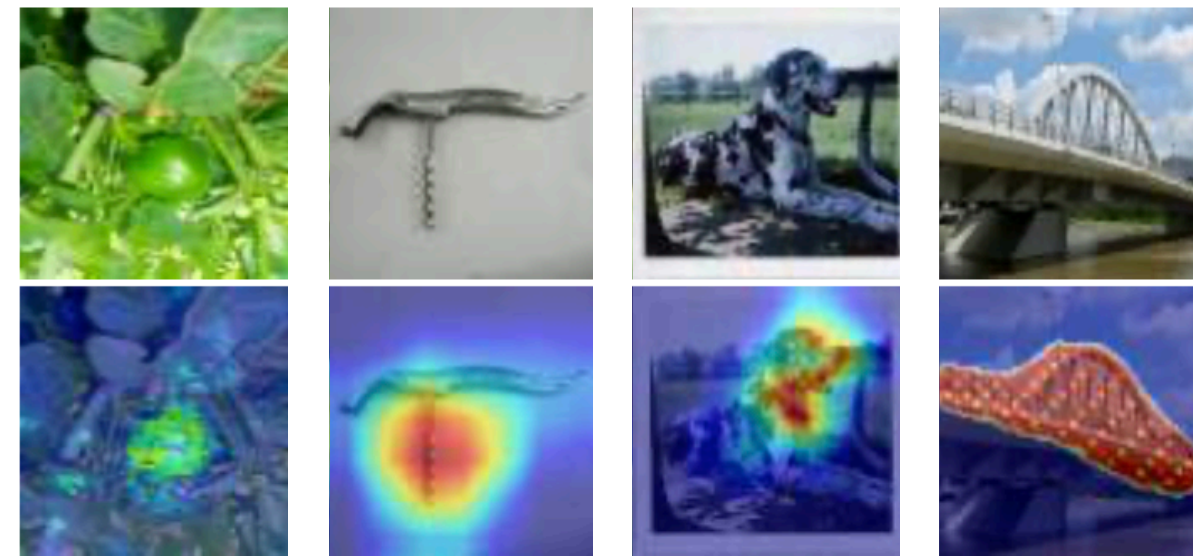
“AI interpretability is a **fallback** to be **used by humans** to **gauge the AI model reasoning** and **assess the societal measurements**”



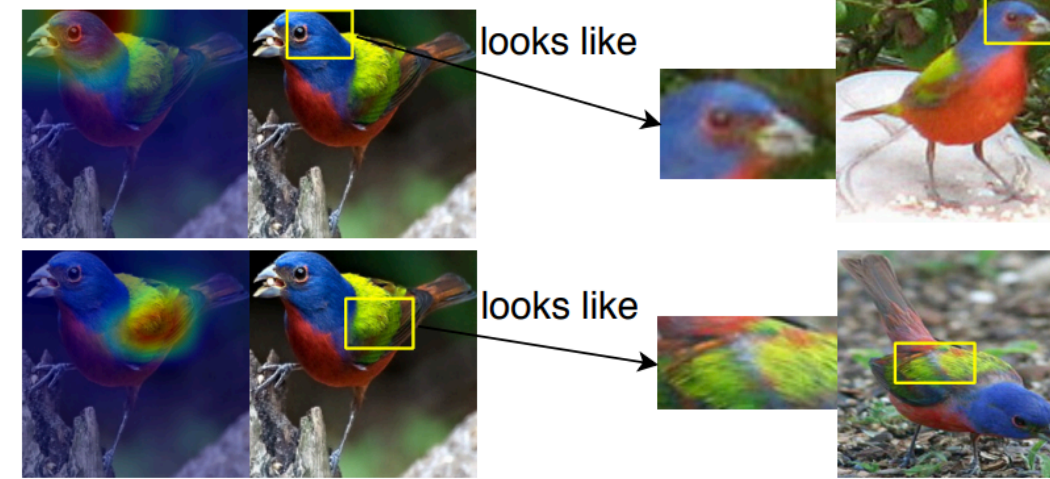
Evaluation of XAI usefulness



Trajectory of XAI algorithms



Gradient CAM Mask RTS
Attribution Heatmaps



(e.g. ProtoPNet)
Prototype Explanations

POSITIVE Examples

train id19684
ox predicted as cow



NEGATIVE Examples

train id4642
cow predicted as ox



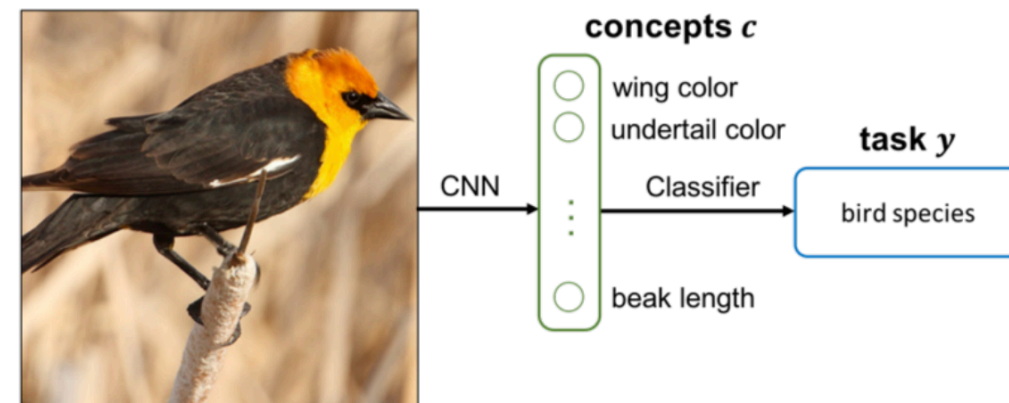
(e.g. Influence Functions, Representer Point)
Training Examples



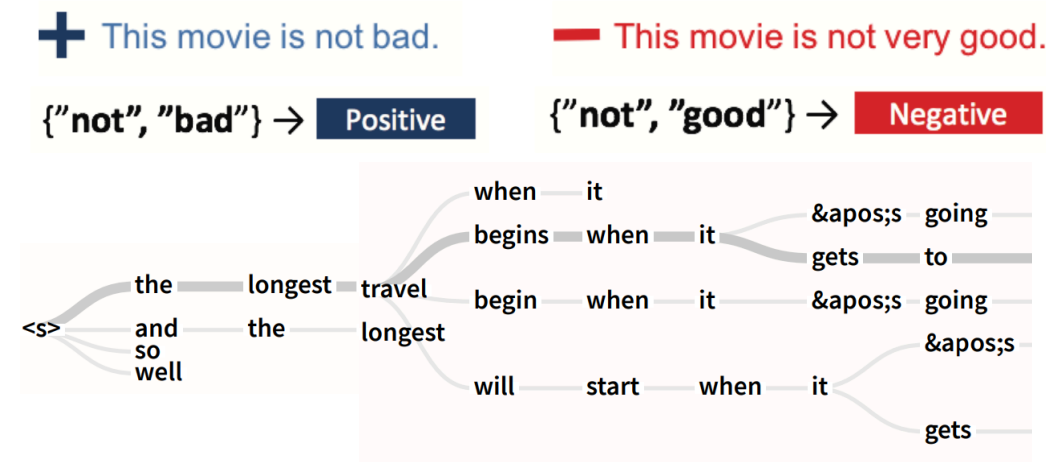
(e.g. SCOUT, Triggers)
Counterfactuals

2012

Concept-based Explanations
(e.g. TCAV, Concept Bottleneck)



Decision Rules / Graphs
(e.g. Anchors)



Free-text Rationales
(e.g. Chain-of-thought)

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

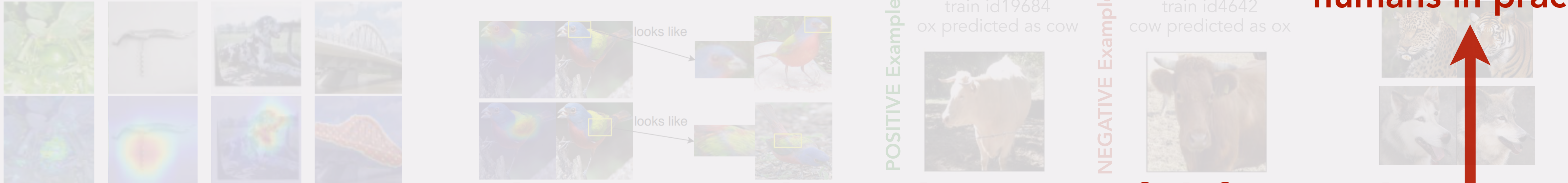
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

...

2023

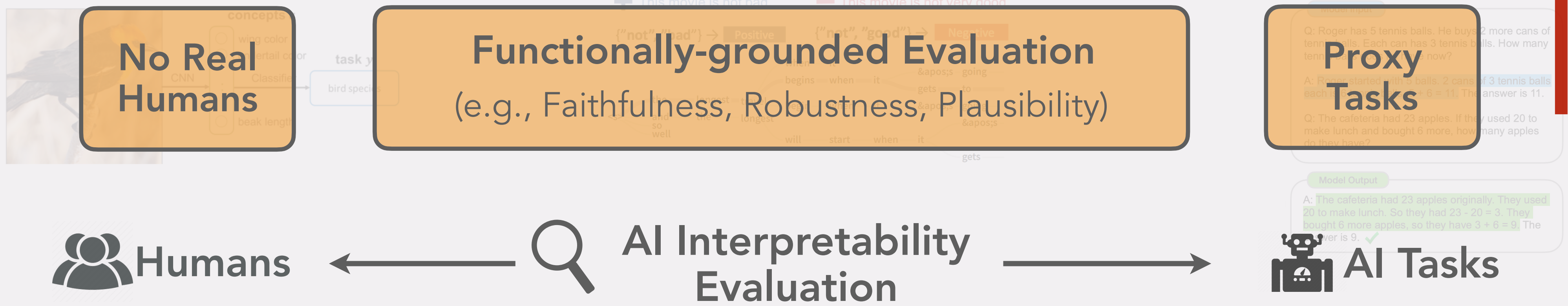
Under-Explored: human evaluation of XAI usefulness

★ Useful XAI for humans in practice



RQ1: Are these XAI algorithms Useful for real humans in practical simple tasks?

2012 ← Concept-based Explanations (e.g. TCAV, Concept Bottleneck) | Decision Rules / Graphs (e.g. Anchors) | Free-text Rationales (e.g. Chain-of-thought) → 2023



Background &
Motivation



**RQ1: Are XAI Useful
for Humans?**

Post-hoc Interpretability

Intrinsic Interpretability

How Useful Are the Machine-Generated Interpretations to General Users?



Hua Shen

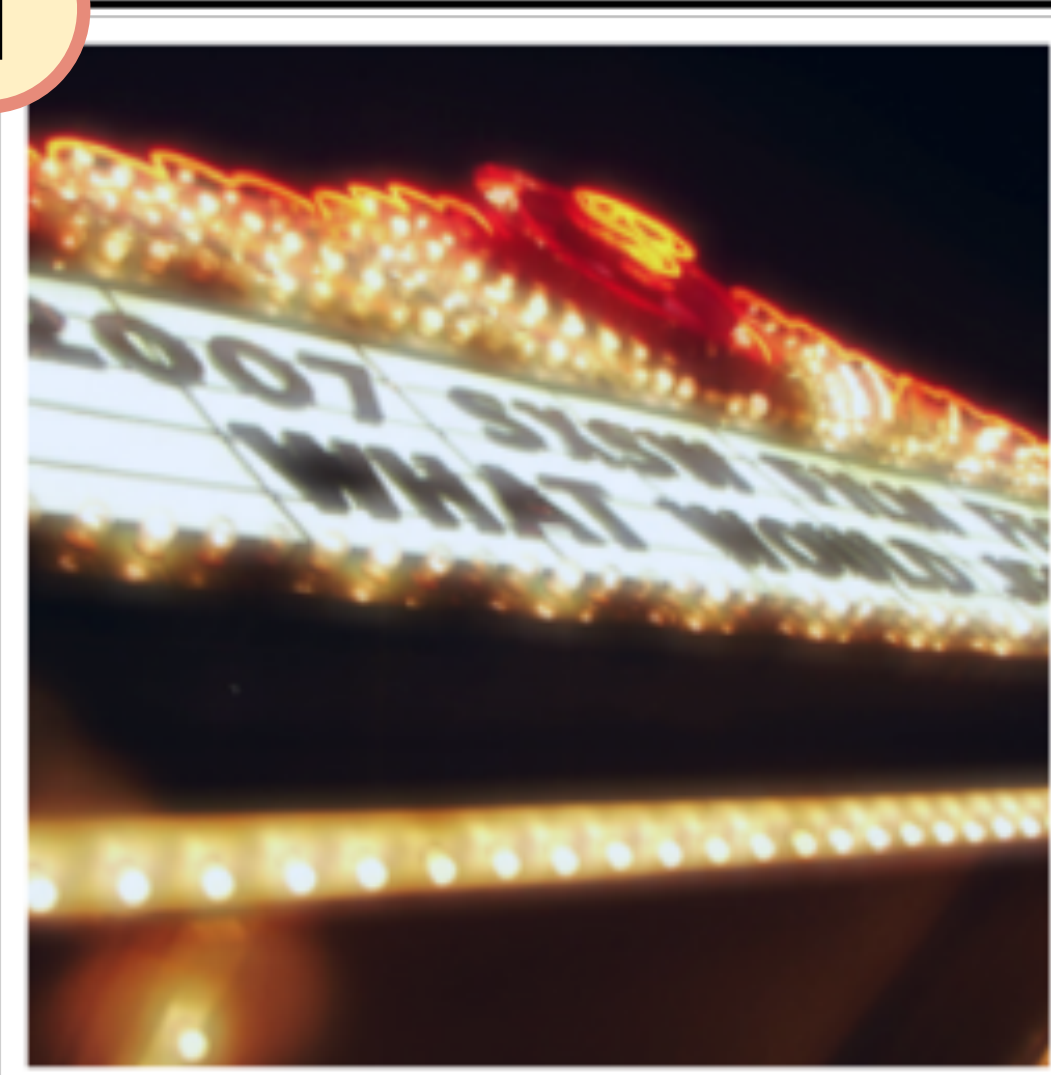


Kenneth Huang

Humans Analyze **Model Errors** in Image Classification

The model **misidentified** this image:

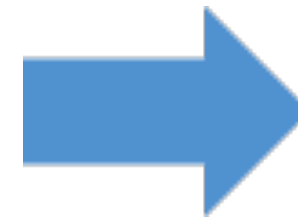
1



2

Correct Label:
Movie Theater

Input Image



4

Guess which label the model **incorrectly predicted?**

- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web



Multiple Choice Question



What AI explanations are **used**?

The model **misidentified** this image:

1 

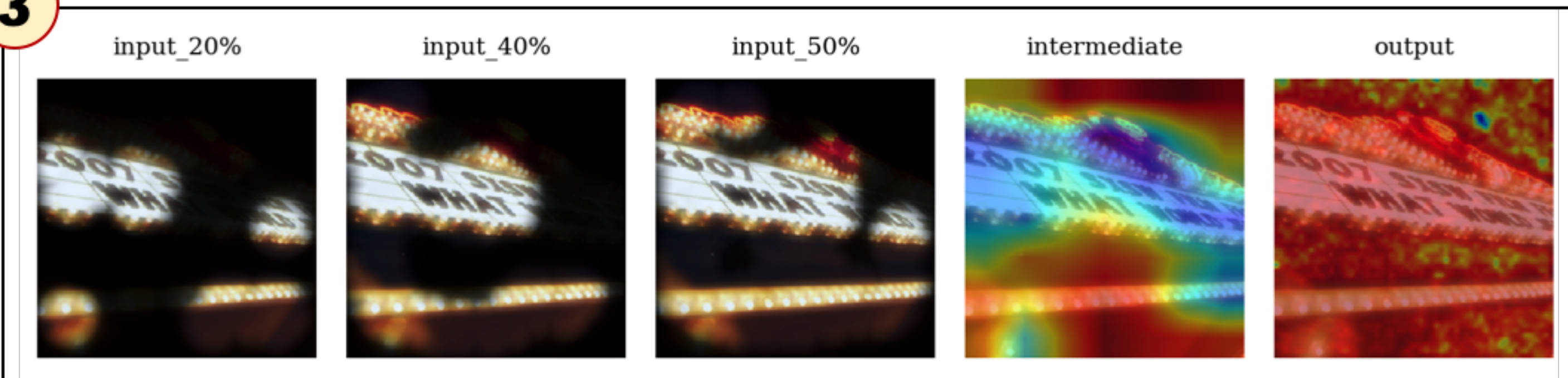
2 Correct Label: **Movie Theater**

Input Image

3

ExtremalPerturb GradCAM SmoothGrad

input_20% input_40% input_50% intermediate output



Machine-Generated Interpretations (Int)

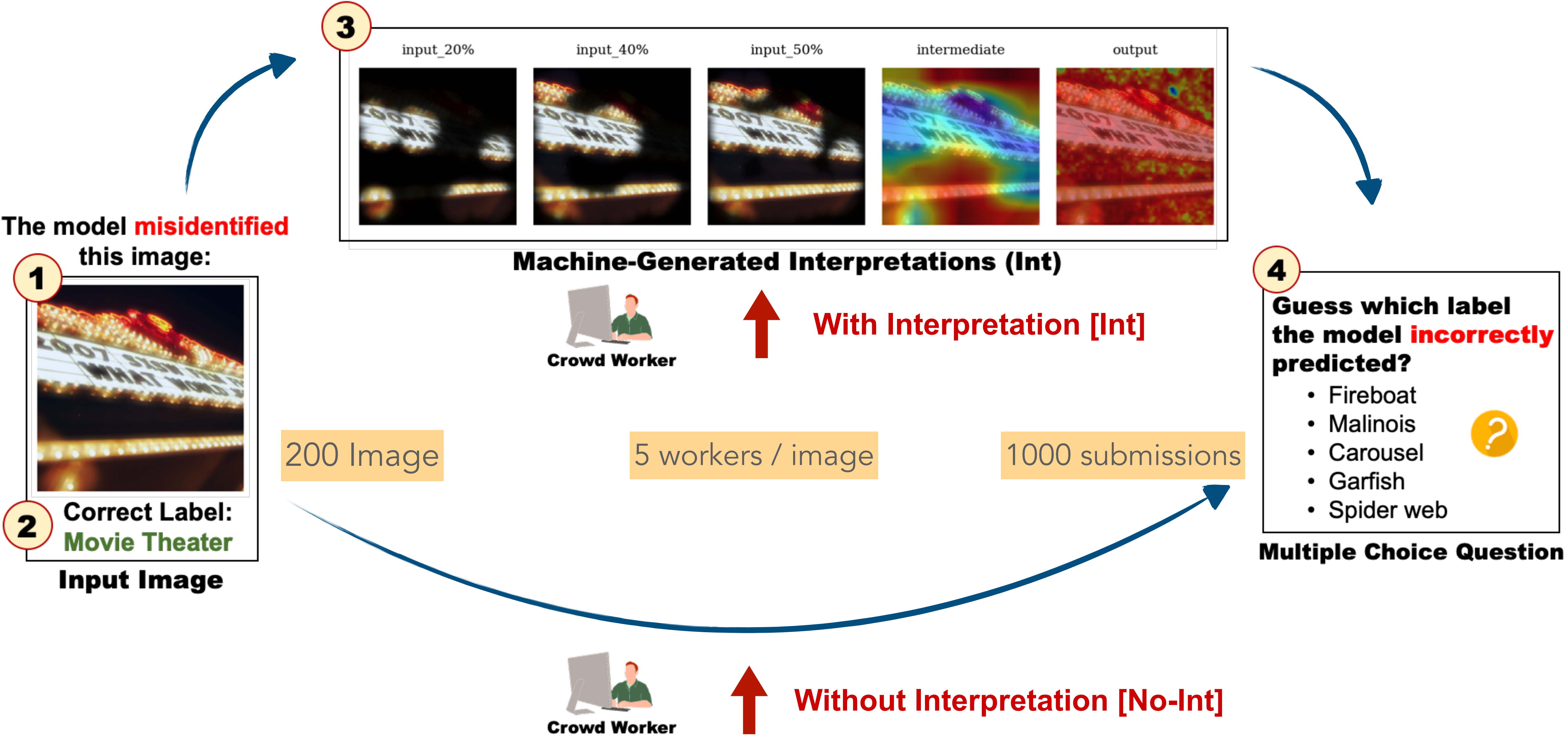
Widely-adopted Saliency Maps as AI Explanations

4 **Guess which label the model **incorrectly** predicted?**

- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web

Multiple Choice Question

Design of Human Study



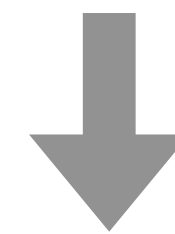
Results

	C1	C2	C3	C4	C5	Overall
Int	0.57	0.74	0.66	0.41	0.67	0.63
No-Int	0.52	0.71	**0.84	*0.59	0.77	**0.73
#images	44	20	112	18	6	200

Table: Average Human on Inferring Model Misclassification (non-overlap users).

Conclusion

[No-interpretation] condition > **[Interpretation]** condition



(statistically significant)

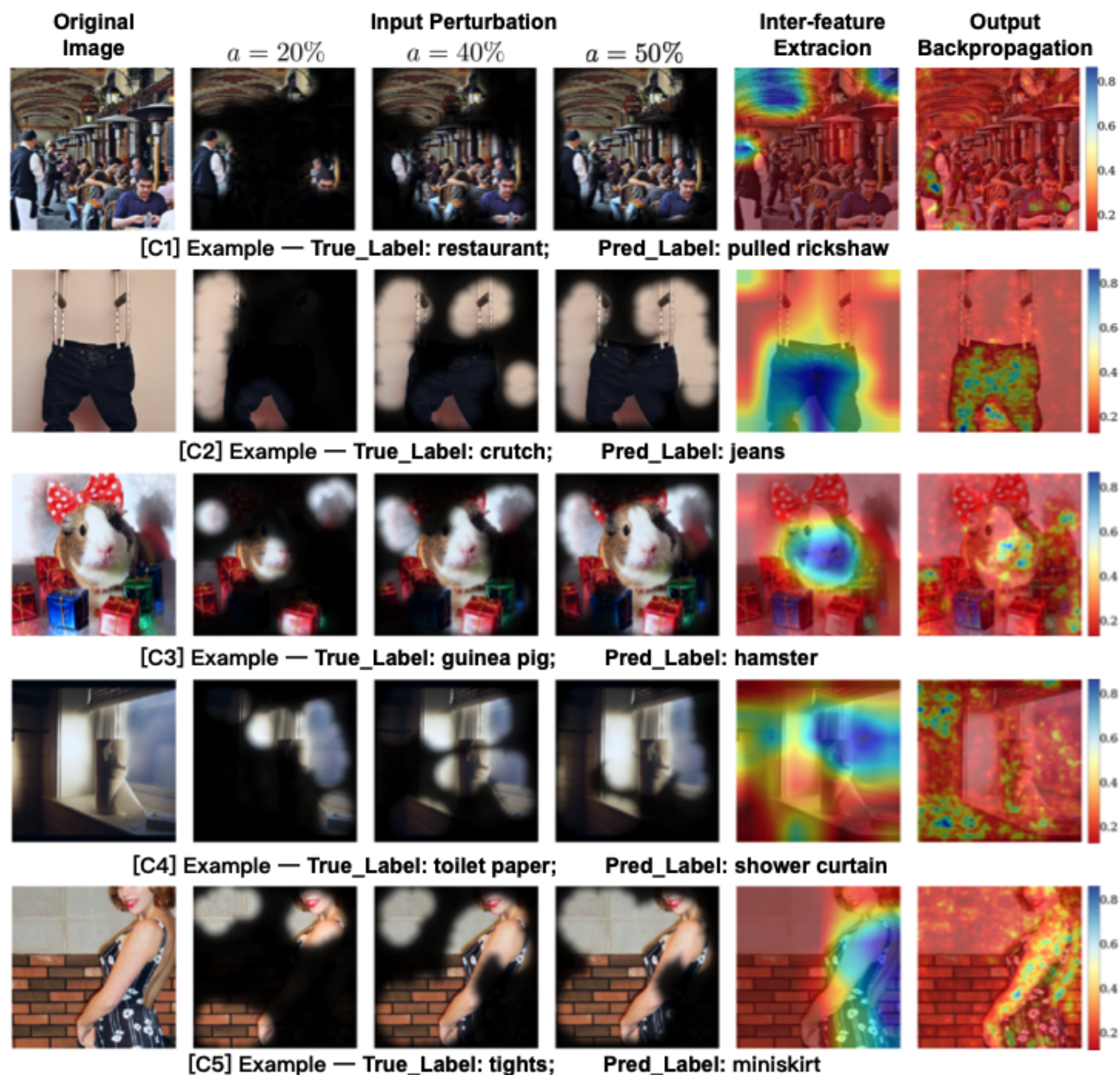
10% Accuracy Drop

Key Findings

Displaying the visual **AI interpretations** to **general users** did not increase, but rather **decreased**, the average **accuracy** on guessing **incorrectly predicted labels** by roughly **10%**.

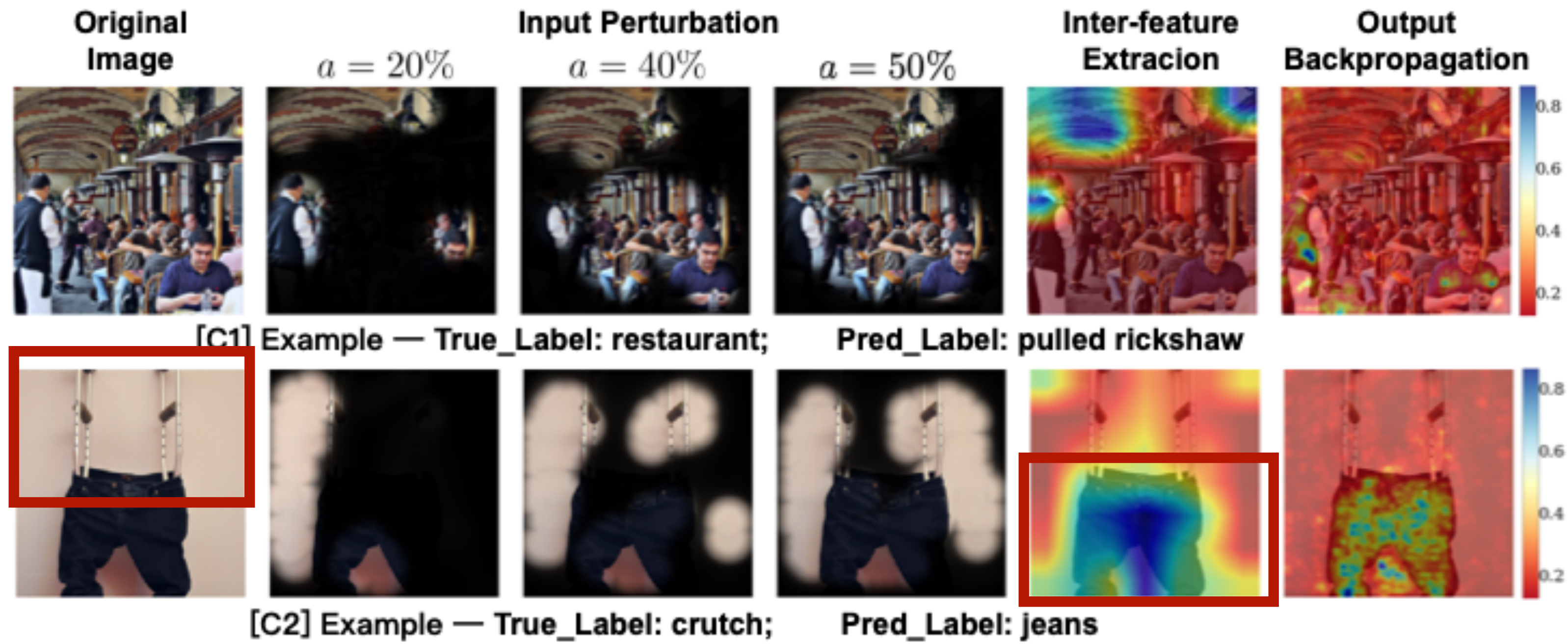
WHY?

Model **error categories** for fine-grained analysis



- **C1:** Local Character Inference
- **C2:** Multiple Objects Selection
- **C3:** Similar Appearance Inference
- **C4:** Correlation Learning
- **C5:** Incorrect Gold-standard Labels

XAI can be **useful** in **some** model **error** categories



	C1	C2	C3	C4	C5	Overall
Int	0.57	0.74	0.66	0.41	0.67	0.63
No-Int	0.52	0.71	**0.84	*0.59	0.77	**0.73
#images	44	20	112	18	6	200

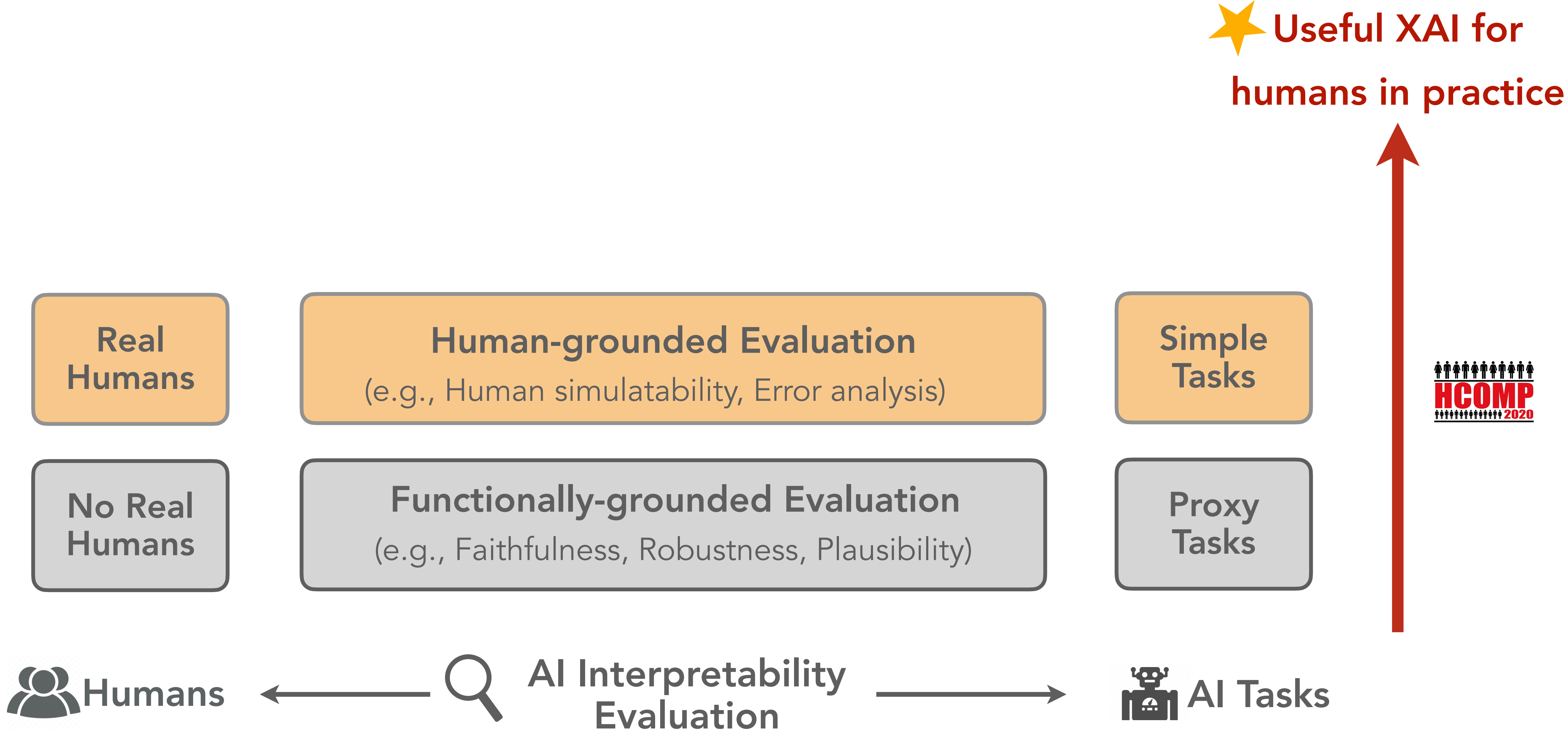
Table: Average Human on Inferring Model Misclassification (non-overlap users).

- **C1:** Local Character Inference
- **C2:** Multiple Objects Selection

Take-away Message

Be **cautious** when **displaying machine-generated interpretations** to explain models' mistakes, because it is **not always helpful** for understanding by general users.

Evaluation of XAI usefulness



Background &
Motivation



**RQ1: Are XAI Useful
for Humans?**



Are Shortest Rationales the Best Explanations for Human Understanding?



Hua Shen



Sherry Wu



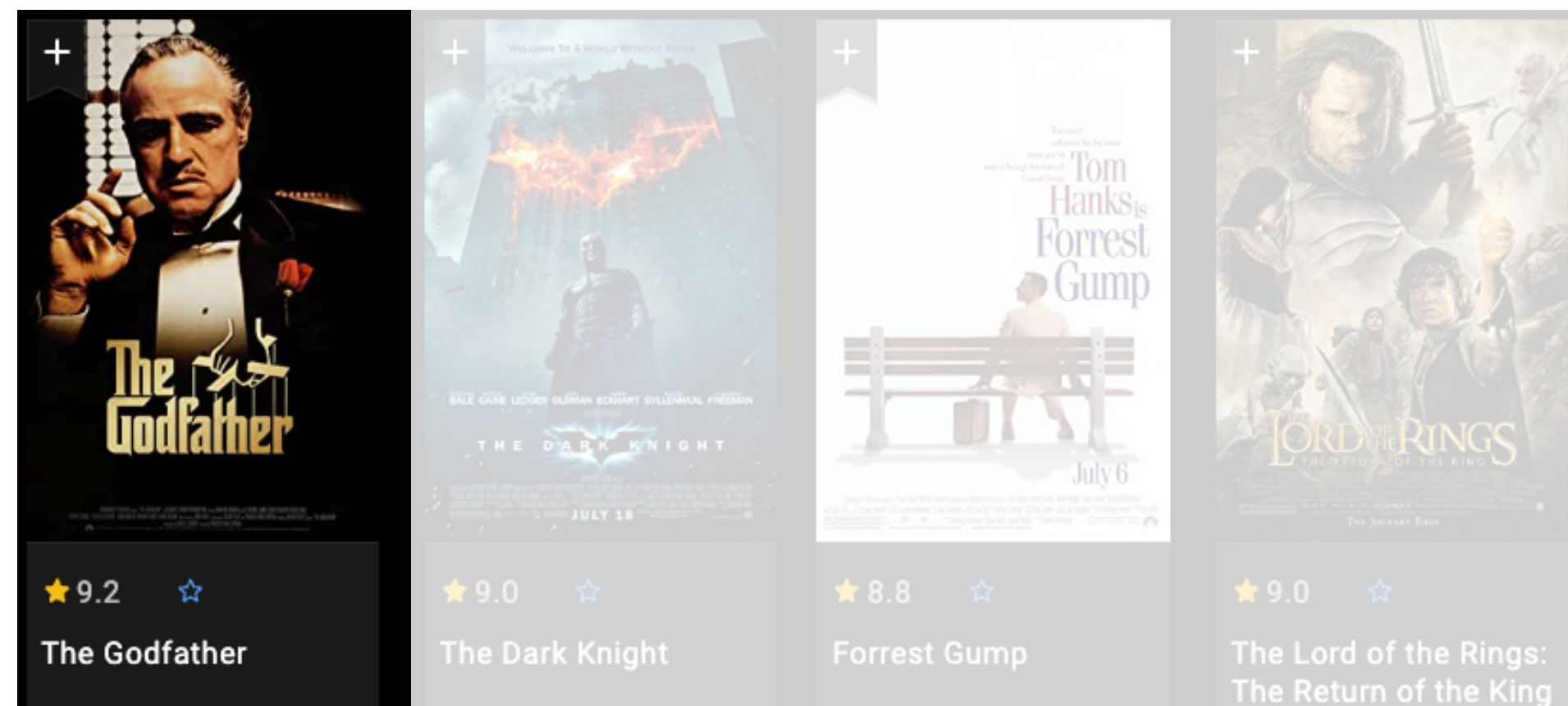
Wenbo Guo



Kenneth Huang

Rationales in Text Classification Tasks

Sentiment Analysis For Movie Reviews



Explanations:

A sufficient **subset** of input **words**, that are **short** and **coherent**, yet **sufficient** to make the **correct** model's **prediction**.



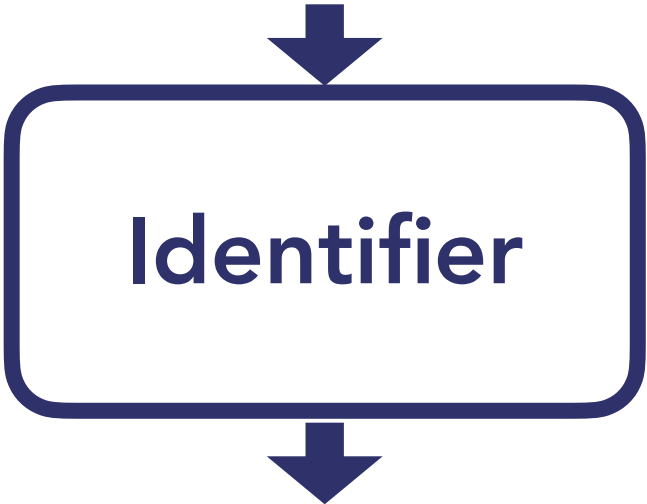
It 's not life - affirming — its vulgar and mean, but I liked it. ← - Explanations

Positive

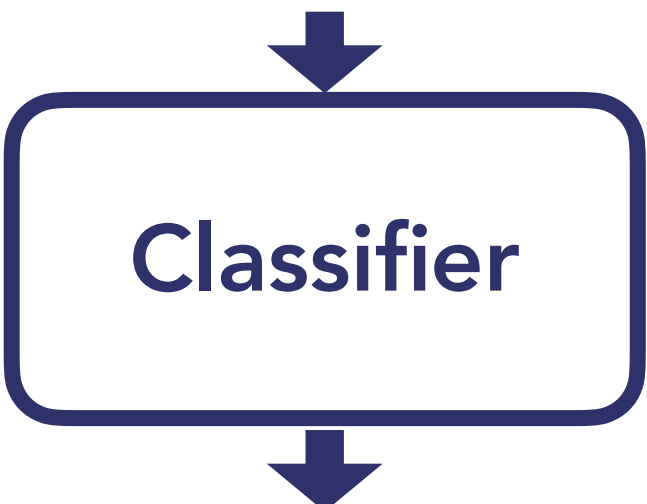
Negative

Self-Explaining Models

It 's not life - affirming — its vulgar and mean, but I liked it.



It 's not life - affirming — its vulgar and mean, but I liked it.



Positive

Prior Study

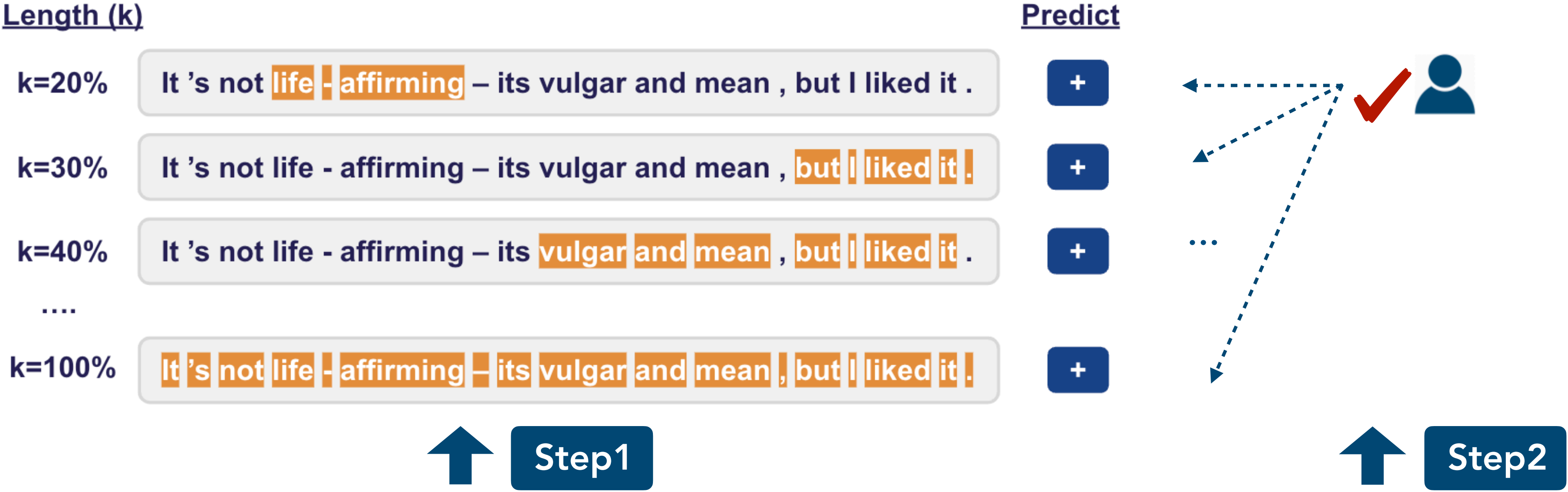
Shorter Explanations are Better.

? Yet to be validated by human studies!

Are *Shortest Explanations* the *Best* for *Human Understanding*?

Overview of Study Design

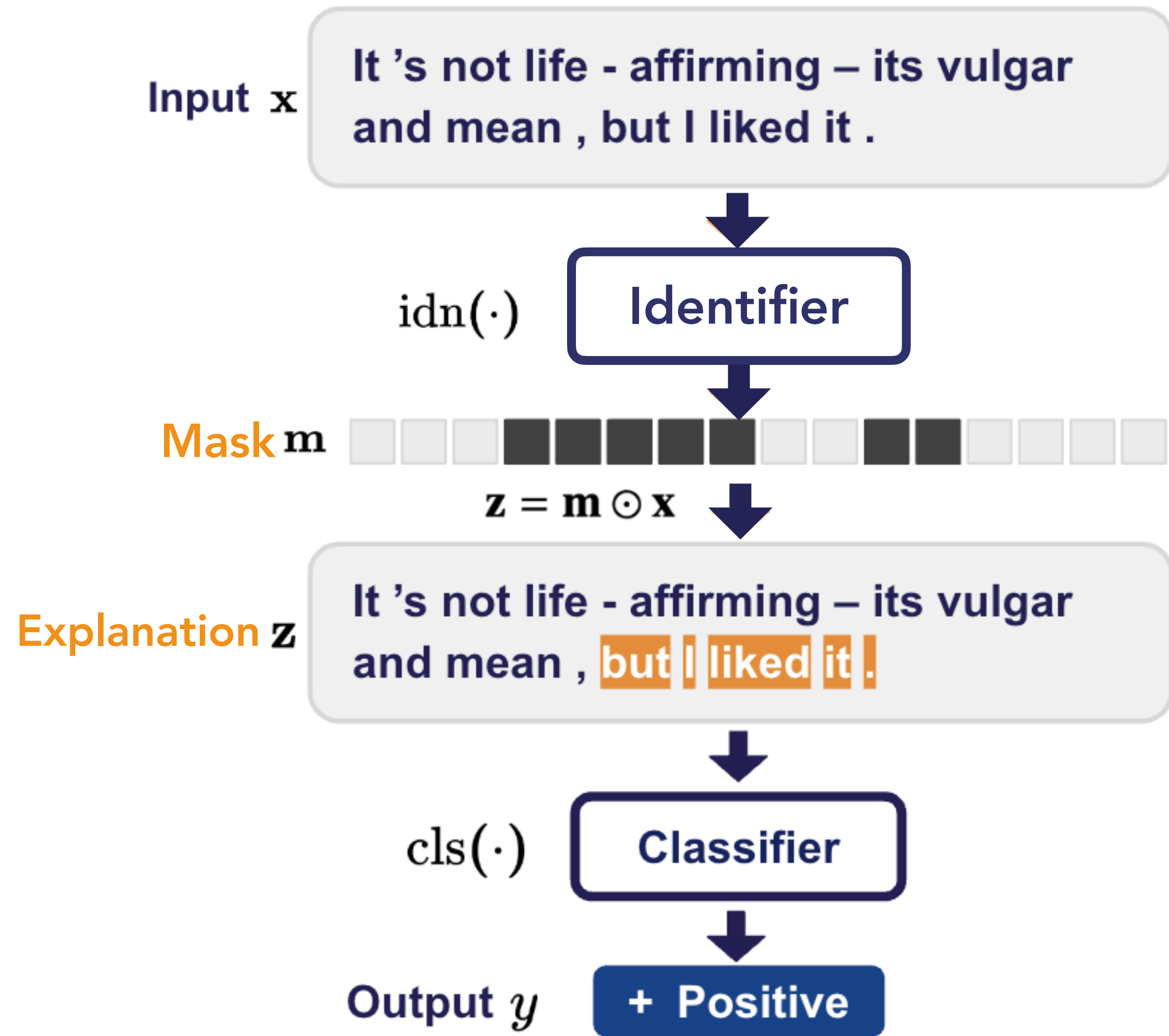
Goal: the impact of explanation lengths on human understanding:



Contribution

A novel self-explaining model **human evaluation on intrinsic interpretability usefulness**

LimitedInk model generates rationals with different length



Optimization Objective

$$\min_{\theta_{\text{idn}}, \theta_{\text{cls}}} \underbrace{\mathbb{E}_{z \sim \text{idn}(x)} \mathcal{L}(\text{cls}(z), y)}_{\text{sufficient prediction}} + \underbrace{\lambda \Omega(m)}_{\text{regularization}}$$

1. Gumbel-Softmax Sampling

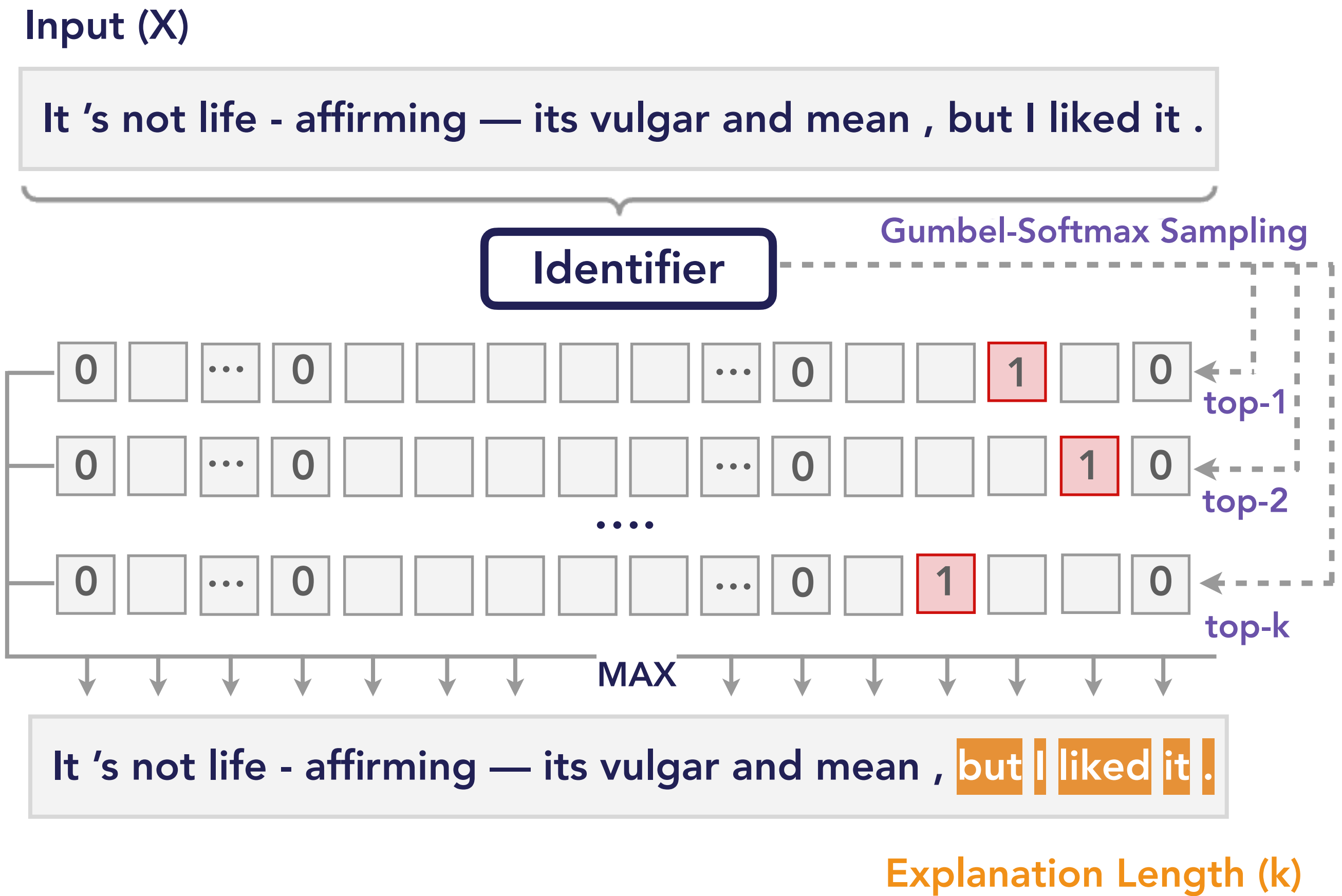
2. Vector and Sort Regularization



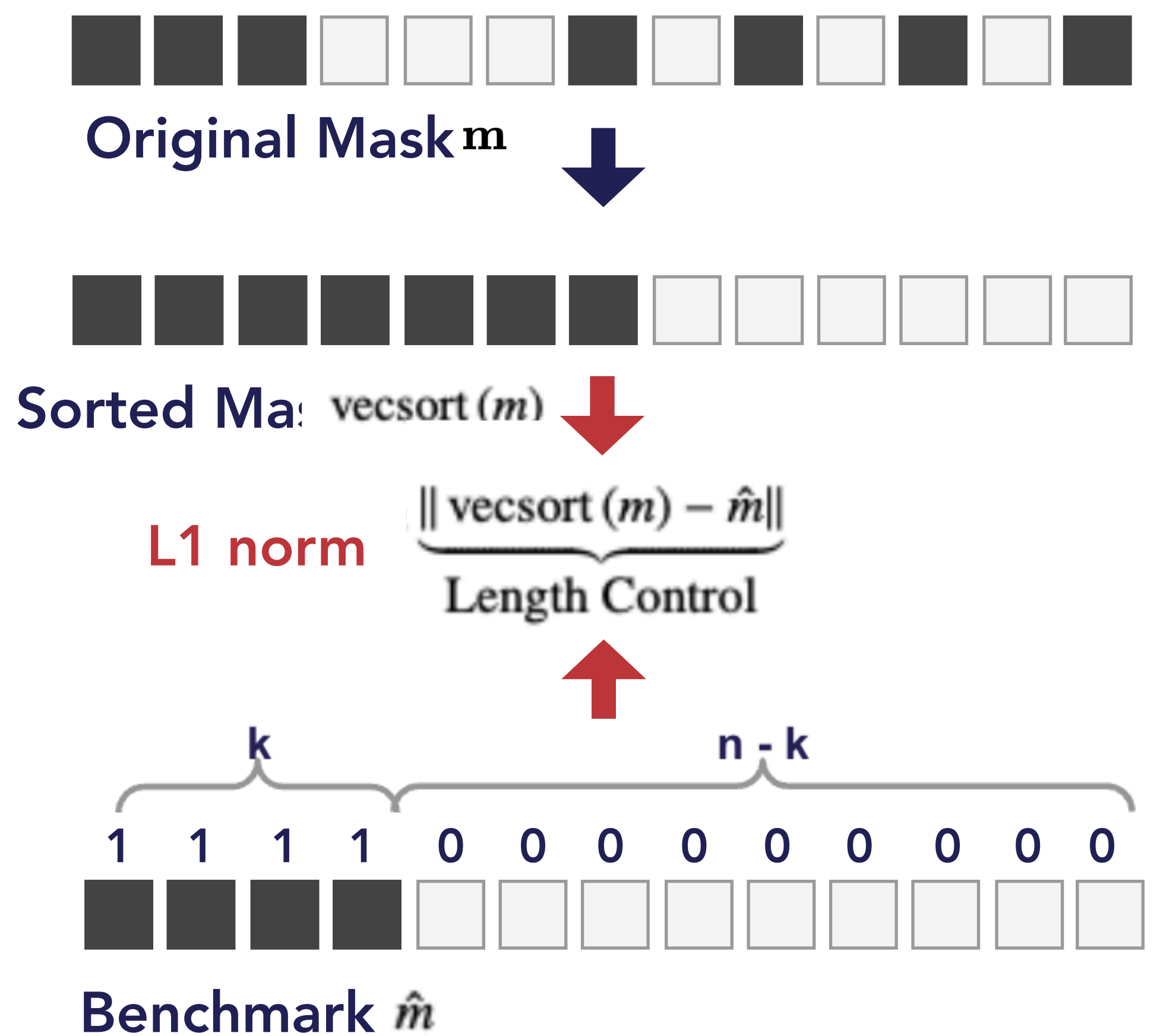
Control Different Explanation Length

How to control explanation length in LimitedInk

1. Gumbel-Softmax Sampling



2. Vector and Sort Regularization



Can LimitedInk perform well on classification?

- End-task classification: **T**ask, weighted average F1
- Human Plausibility with annotated dataset: **P**recision, **R**ecall, Token-level **F1**

Method	Movies				BoolQ				Evidence Inference				MultiRC				FEVER				
	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	
Full-Text	.91	-	-	-	.47	-	-	-	.48	-	-	-	.67	-	-	-	.89	-	-	-	
Sparse-N	.79	.18	.36	.24	.43	.12	.10	.11	.39	.02	.14	.03	.60	.14	.35	.20	.83	.35	.49	.41	
Sparse-C	.82	.17	.36	.23	.44	.15	.11	.13	.41	.03	.15	.05	.62	.15	.41	.22	.83	.35	.52	.42	
Sparse-IB	.84	.21	.42	.28	.46	.17	.15	.15	.43	.04	.21	.07	.62	.20	.33	.25	.85	.37	.50	.43	
LIMITEDINK	.90	.26	.50	.34	.56	.13	.17	.15	.50	.04	.27	.07	.67	.22	.40	.28	.90	.28	.67	.39	
Length Level		50%				30%				50%				50%				40%			

LimitedInk **performed compatible with three SOTA baselines** on the two common rationale metrics in five ERASER text classification benchmark datasets.

Step2 - Human Study Setups

LimitedInk Explanations

k=10%
... not

k=20%
..... life - affirming

k=30%
..... but | liked it

k=40%
..vulgar and mean but | liked it

k=50%
..... its vulgar and mean , but | liked it , ...

Random text spans (similar length)

k=10%
... affirming

k=20%
..... affirming - its

k=30%
.... its vulgar and mean

k=40%
..... not life - affirming ... | liked it

k=50%
... life - affirming - its vulgar and mean , but

Only highlight explanations & hide other texts!

Five-level explanations:
10%, 20%, 30%, 40%, 50%

We conducted **user studies** to investigate the **human understanding** on **LimitedInk** and **Baseline** (random sampled tokens).

Human Task Design

Select Sentiment and Confidence of the Displayed Parts of Movie Review

Please select the **sentiment label** of the displayed parts of the movie review and provide your **confidence** on the selection.

Parts of the Movie Review 1

..... recall hearing species 2 described as "erotic." I would love to know who used with that adjective for this a woman's abdomen as an alien baby claws its way free , splat blood and gore in all directions . anyone turned on by that

Question1: Is the movie review **Positive** or **Negative**? Please guess based on the parts of texts you see.

(Empty reviews are usually caused by data processing errors)

Question2: How **Confident** are you in your above selection?

- The displayed texts show clear attitude, and reflects the core sentiment (like/dislike) of the full review.

- The displayed texts show attitude towards the movie, but not very clear to reflect the core sentiment.

- The displayed texts seem positive/negative, but I cannot guess if it's representative of the full review.

- The displayed texts are ambiguous. I am not confident on the attitude towards the movie.

- The displayed texts are too trivial and does not reflect on the larger themes.



Sentiment Analysis:

we randomly sampled **100** reviews (correct prediction) from the **Movie review** test set



prediction



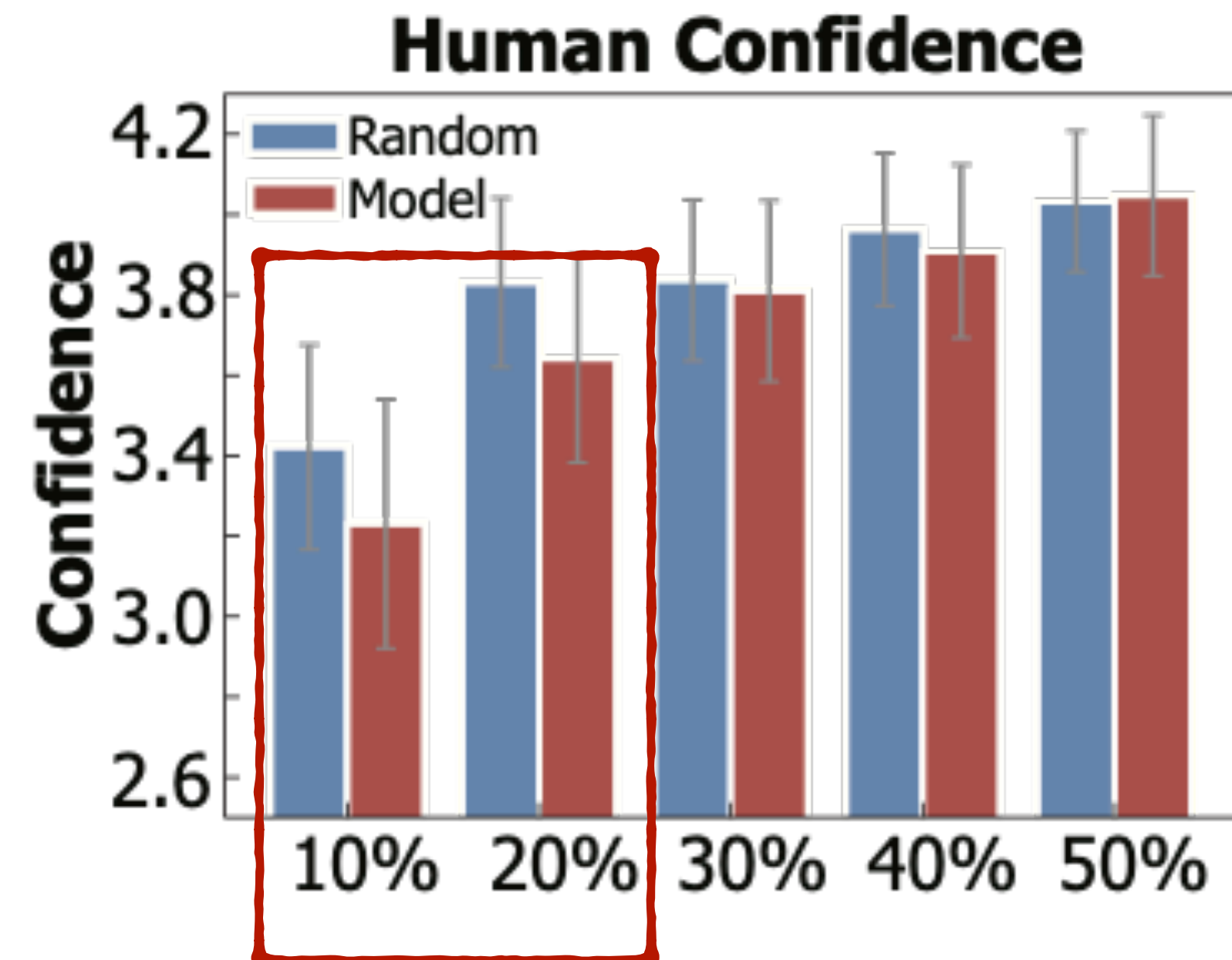
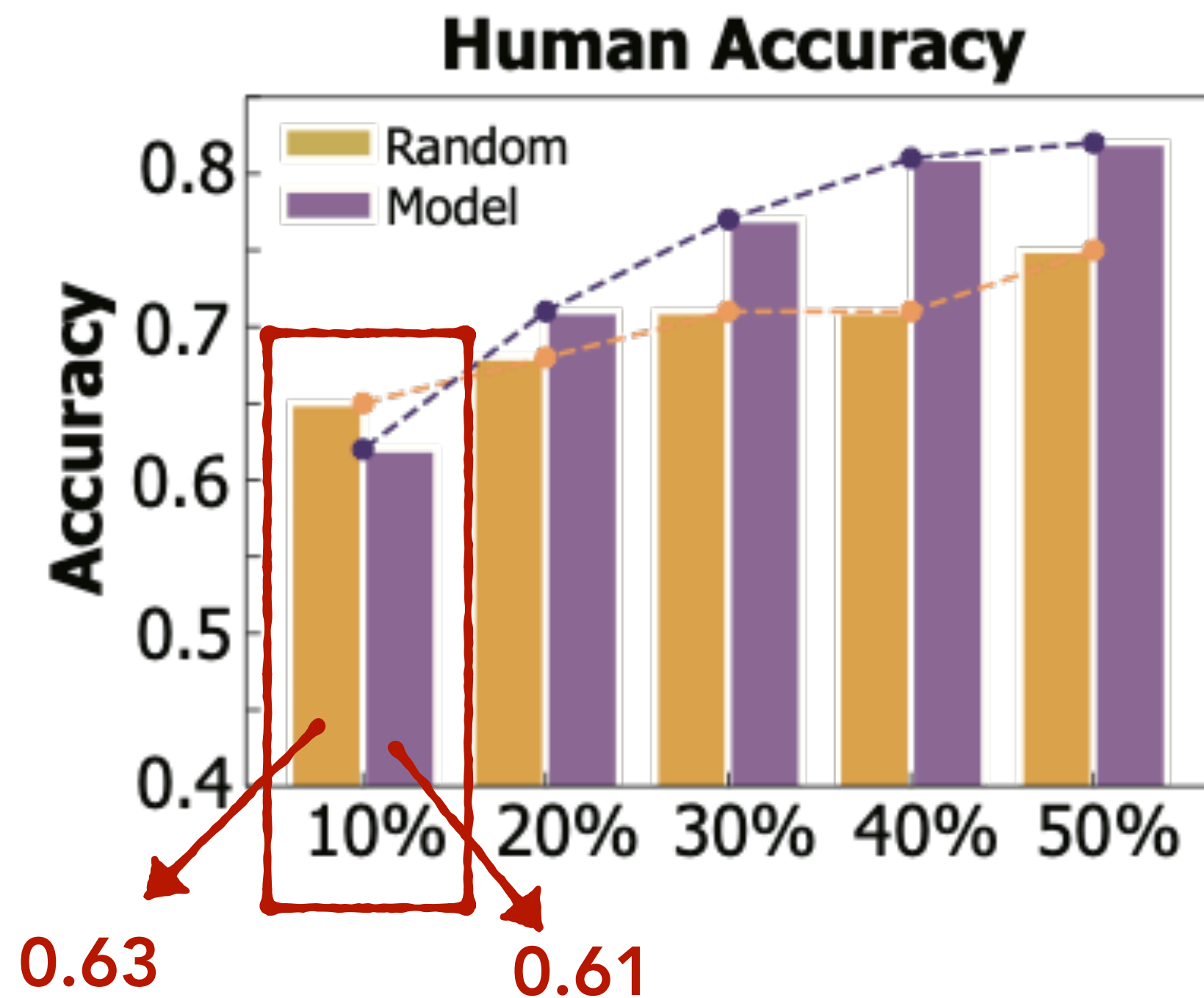
confidence



MTurk
Workers

We asked MTurk Workers to **simulate model predictions** and provide the **confidence** on **100 random instances** from **Movie Review** dataset. Each worker sees a review only once.

Key Findings



Human **accuracy** and **confidence**, at the shortest level (i.e., 10% length), are **lower than** the random baseline.

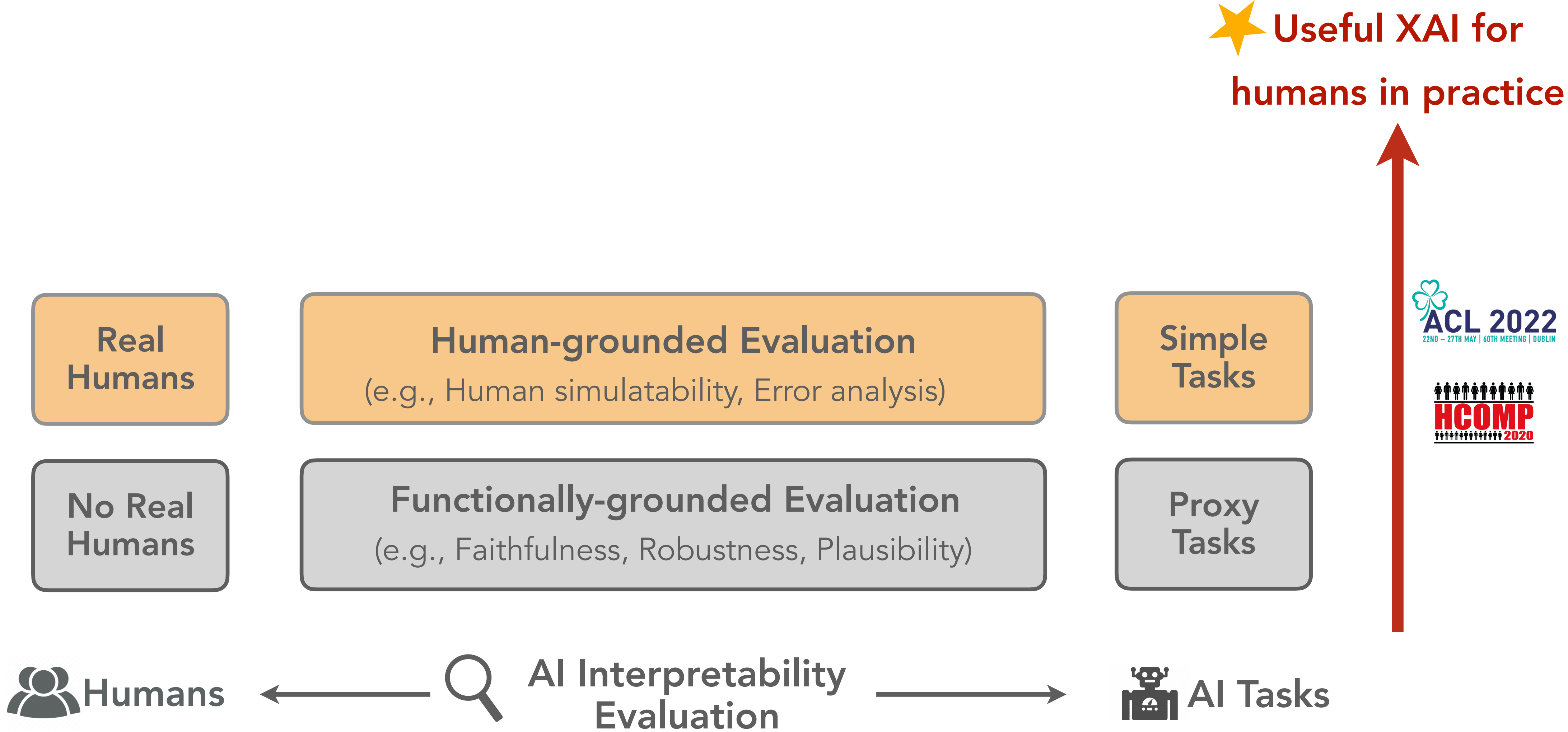
The **shortest rationales** are **NOT always the best** for human understanding.

Take-away Message

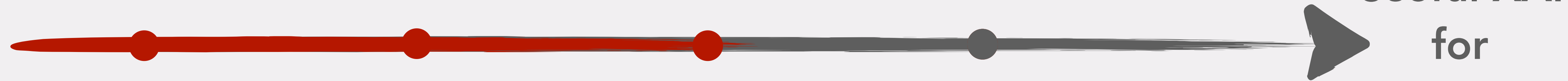
Shortest explanations are largely **NOT the best** for humans to simulate model predictions.

With very **short rationales**, the explanations might **NOT be helpful** for humans.

Evaluation of XAI usefulness



Background &
Motivation



RQ1: Are XAI Useful
for Humans?

RQ2: Why?

Useful XAI
for
Humans





@



Explaining the Road Not Taken



Hua Shen



Ting-Hao 'Kenneth' Huang

What are the real-world user needs for XAI?

XAI Question Bank

43 XAI User Questions

<p>Input</p>	<ul style="list-style-type: none"> • What kind of data does the system learn from? • What is the source of the data? • How were the labels/ground-truth produced? • * What is the sample size? • * What data is the system NOT using? • * What are the limitations/biases of the data? • * How much data [like this] is the system trained on? 	<p>Why</p>	<ul style="list-style-type: none"> • Why/how is this instance given this prediction? • What feature(s) of this instance leads to the system's prediction? • Why are [instance A and B] given the same prediction?
<p>Output</p>	<ul style="list-style-type: none"> • What kind of output does the system give? • What does the system output mean? • How can I best utilize the output of the system ? • * What is the scope of the system's capability? Can it do...? • * How is the output used for other system component(s) ? 	<p>Why not</p>	<ul style="list-style-type: none"> • Why/how is this instance NOT predicted...? • Why is this instance predicted P instead of Q? • Why are [instance A and B] given different predictions?
<p>Performance</p>	<ul style="list-style-type: none"> • How accurate/precise/reliable are the predictions? • How often does the system make mistakes? • In what situations is the system likely to be correct/incorrect? • * What are the limitations of the system? • * What kind of mistakes is the system likely to make? • * Is the system's performance good enough for... 	<p>What If</p>	<ul style="list-style-type: none"> • What would the system predict if this instance changes to...? • What would the system predict if this feature of the instance changes to...? • What would the system predict for [a different instance]?
<p>How (global)</p>	<ul style="list-style-type: none"> • How does the system make predictions? • What features does the system consider? <ul style="list-style-type: none"> • * Is [feature X] used or not used for the predictions? • What is the system's overall logic? <ul style="list-style-type: none"> • How does it weigh different features? • What rules does it use? • How does [feature X] impact its predictions? • * What are the top rules/features it uses? • * What kind of algorithm is used? <ul style="list-style-type: none"> • * How are the parameters set? 	<p>How to be that</p>	<ul style="list-style-type: none"> • How should this instance change to get a different prediction? • How should this feature change for this instance to get a different prediction? • What kind of instance gets a different prediction?
		<p>How to still be this</p>	<ul style="list-style-type: none"> • What is the scope of change permitted to still get the same prediction? • What is the [highest/lowest/...] feature(s) one can have to still get the same prediction? • What is the necessary feature(s) present or absent to guarantee this prediction? • What kind of instance gets this prediction?
		<p>Others</p>	<ul style="list-style-type: none"> • * How/what/why will the system change/adapt/improve/drift over time? (change) • * How to improve the system? (change) • * Why using or not using this feature/rule/data? (follow-up) • * What does [ML terminology] mean? (terminological) • * What are the results of other people using the system? (social)

*How **well** can existing **XAI** algorithms **respond** to these questions that **users care in practice** ?*

We surveyed 200+ XAI Papers related to NLP

ID	Title	Year	Venue	Paper URL
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938.pdf
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/anthology/N16-1082.pdf
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taolei/papers/emnlp16_rationale.pdf
4	BERT Rediscovered the Classical NLP Pipeline	2019	ACL	https://www.aclweb.org/anthology/P19-1452.pdf
5	Attention is not Explanation	2019	NAACL	https://arxiv.org/pdf/1902.10186.pdf

Matching XAI Papers with XAI Question Bank?

43 User Questions

218 XAI Papers

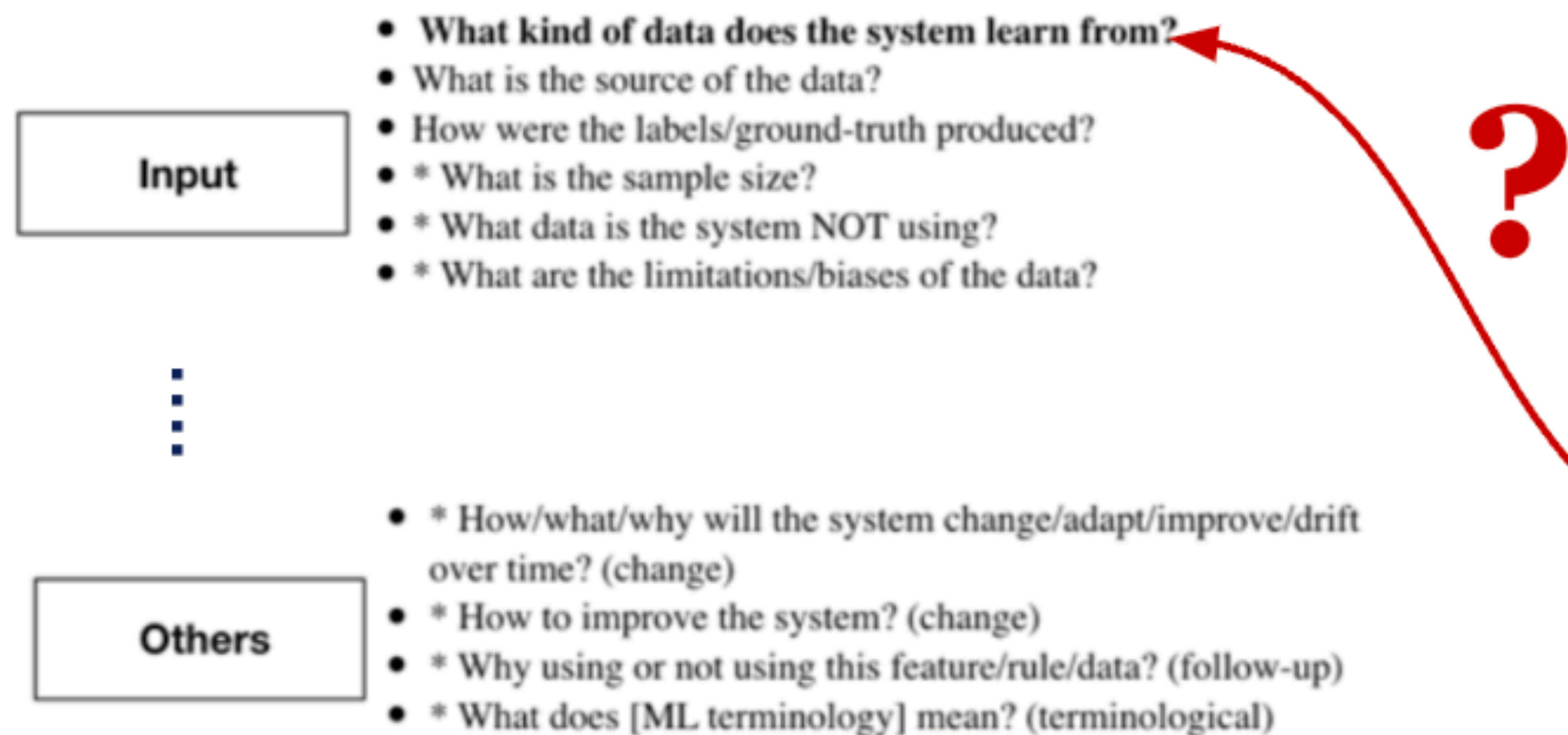
Input

- What kind of data does the system learn from?
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?

⋮

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)



ID	Title	Year	Venue	Paper URL
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/antholog
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
4	BERT Rediscovered the Classical NLP Pipeline	2019	ACL	https://www.aclweb.org/antholog
5	Attention is not Explanation	2019	NAACL	https://arxiv.org/pdf/1902.10186
⋮				
214	How much should you ask? On the question structure in QA systems	2018	BlackboxNLP	https://arxiv.org/pdf/1809.03734
215	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	https://arxiv.org/pdf/2011.06854
216	A Survey of the State of Explainable AI for Natural Language Processing	2020	AAACL-IJCNLP	https://arxiv.org/pdf/2010.00711
217	Explaining Simple Natural Language Inference	2019	ACL	https://www.aclweb.org/antholog
218	Understanding Neural Abstractive Summarization Models via Uncertainty	2020	EMNLP	https://arxiv.org/pdf/2010.07882

Manually Matching: $218 * 43 = 9,374 \dots$

Matching Each User Question with XAI Forms in NLP

43 User Questions

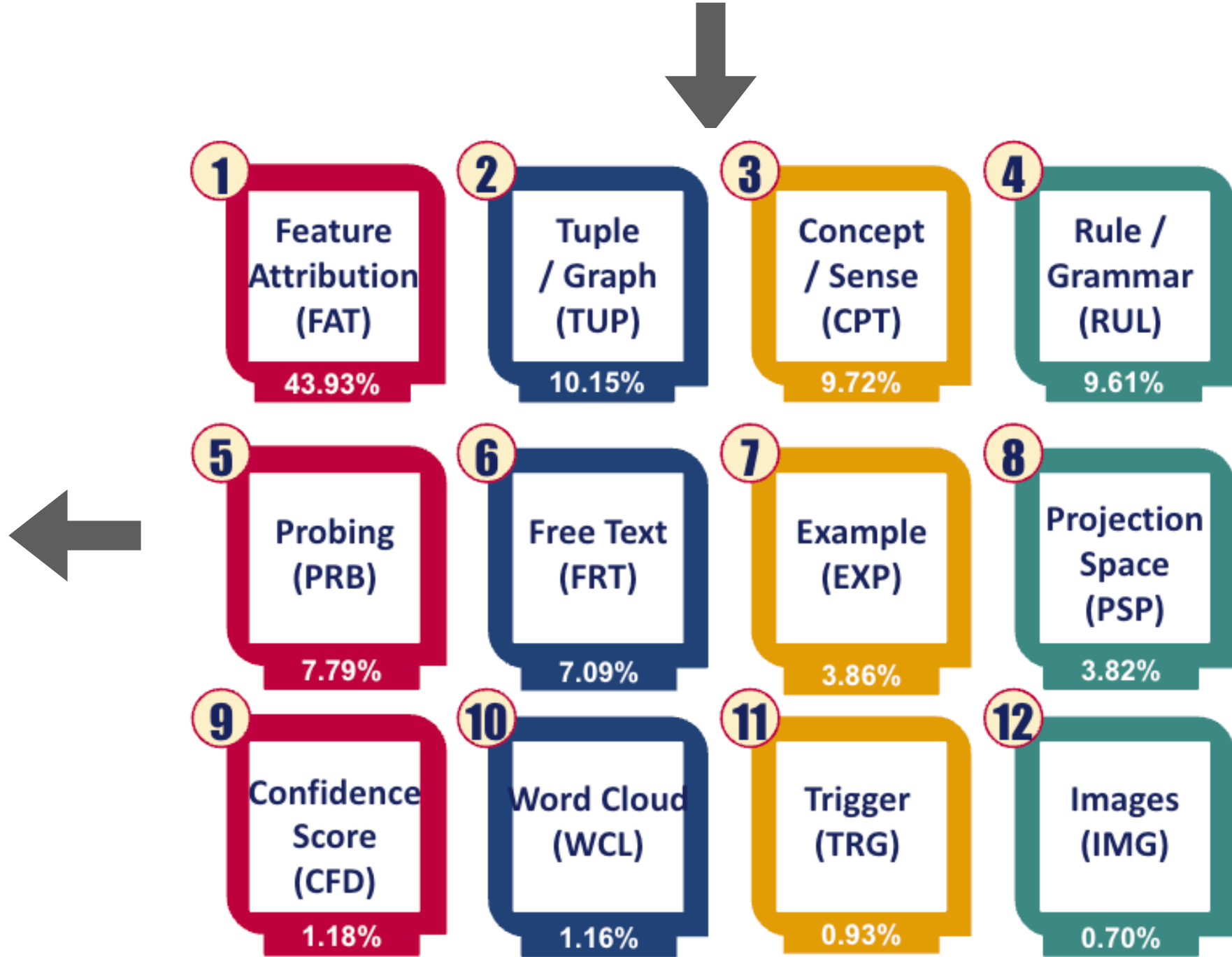
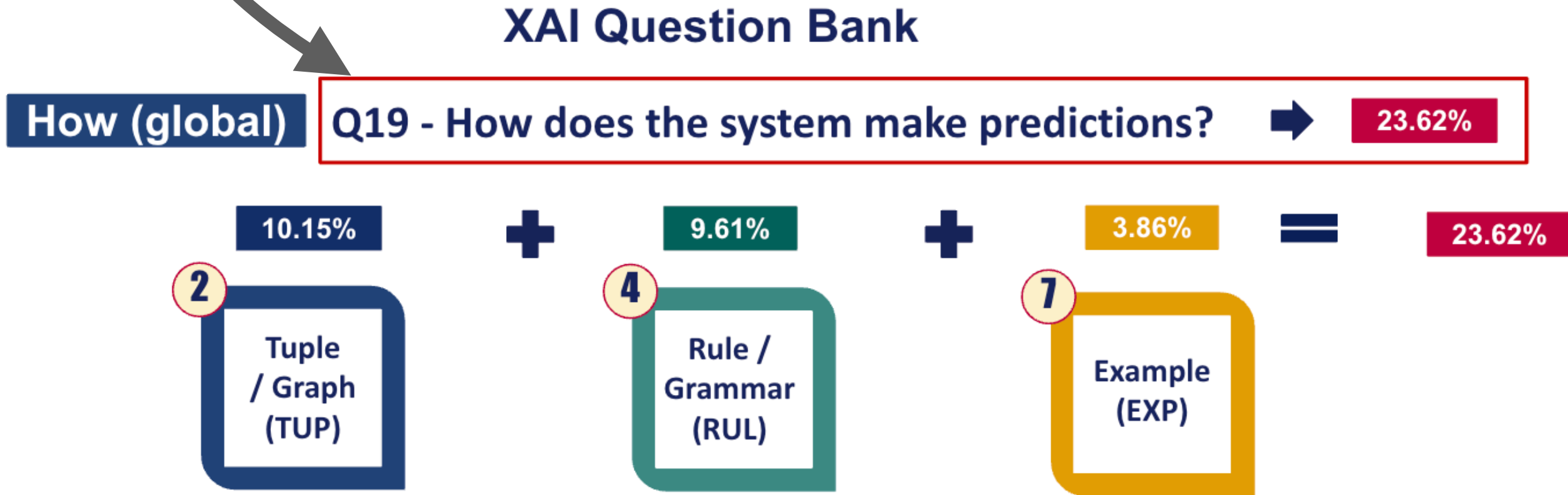
- Input

 - What kind of data does the system learn from
 - What is the source of the data?
 - How were the labels/ground-truth produced?
 - * What is the sample size?
 - * What data is the system NOT using?
 - * What are the limitations/biases of the data?
- ⋮
- Others

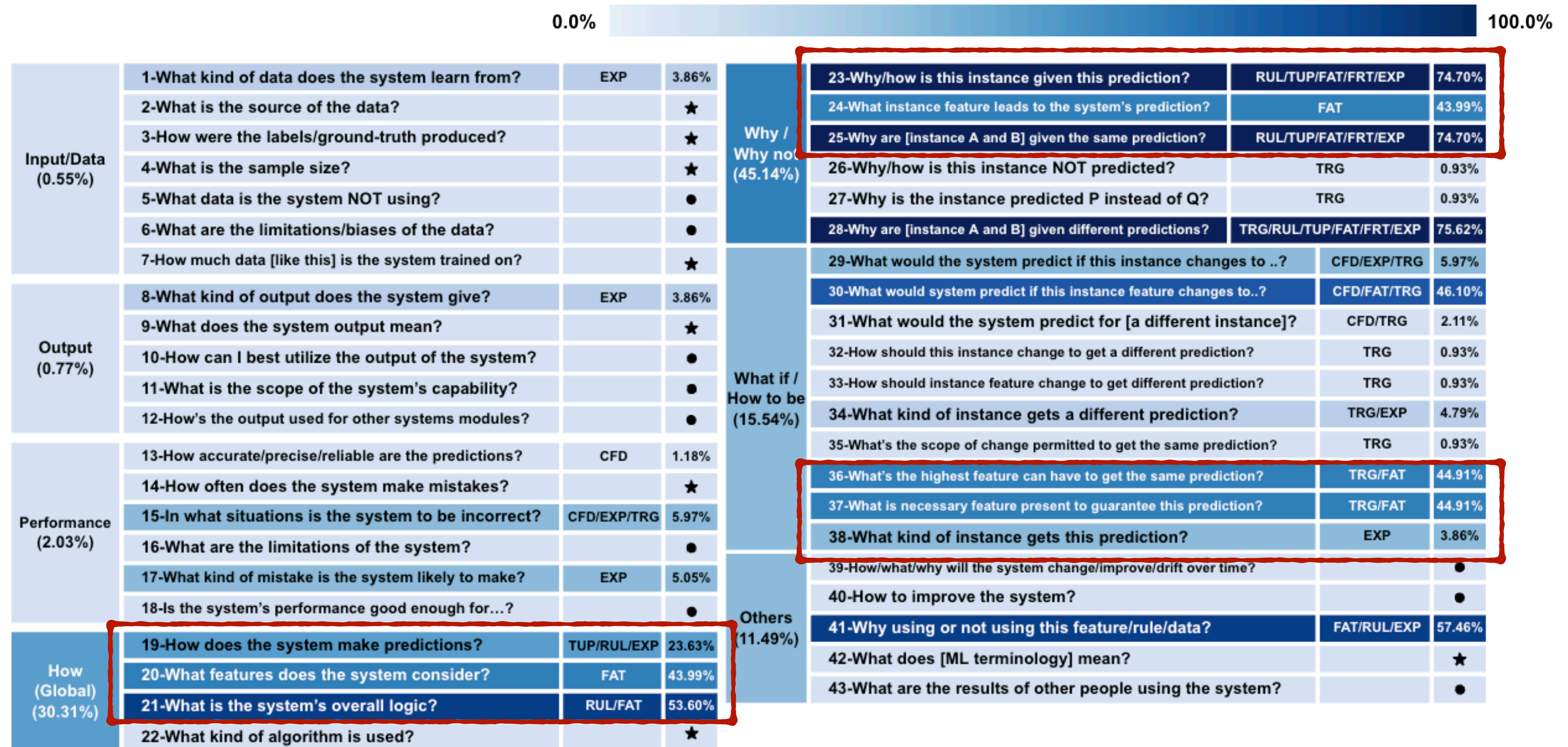
 - * How/what/why will the system change/adapt/improve/drift over time? (change)
 - * How to improve the system? (change)
 - * Why using or not using this feature/rule/data? (follow-up)
 - * What does [ML terminology] mean? (terminological)

200+ XAI Papers

ID	Title	Year	Venue	Paper URL
1	"Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/anthology
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
4	BERT Rediscovered the Classical NLP Pipeline	2019	ACL	https://www.aclweb.org/anthology
5	Attention is not Explanation	2019	NAACL	https://arxiv.org/pdf/1902.10186
⋮				
214	How much should you ask? On the question structure in QA systems	2018	BlackboxNLP	https://arxiv.org/pdf/1809.03734
215	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	https://arxiv.org/pdf/2011.06854
216	A Survey of the State of Explainable AI for Natural Language Processing	2020	AAACL-IJCNLP	https://arxiv.org/pdf/2010.00711
217	Explaining Simple Natural Language Inference	2019	ACL	https://www.aclweb.org/anthology
218	Understanding Neural Abstractive Summarization Models via Uncertainty	2020	EMNLP	https://arxiv.org/pdf/2010.07882

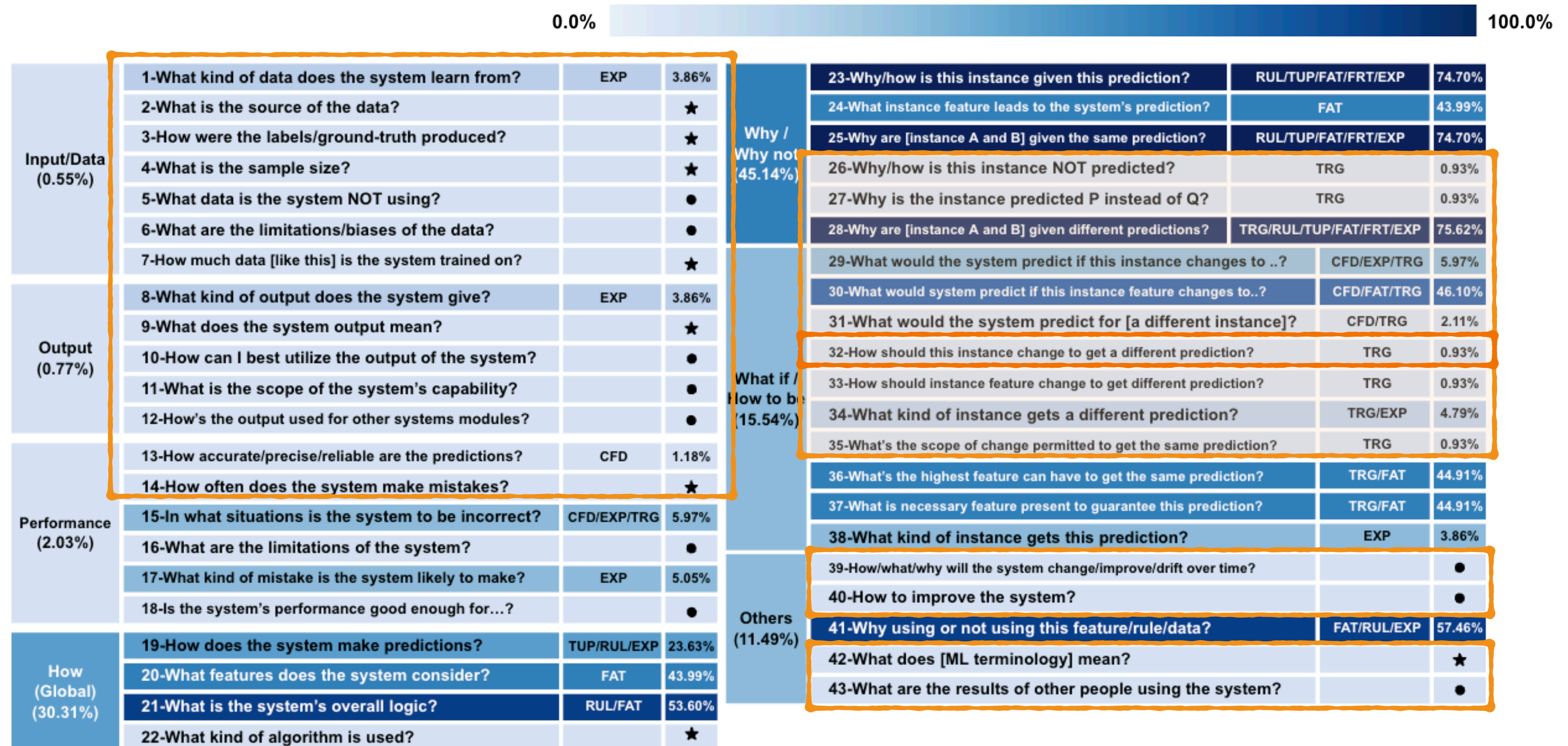


Results: Heatmap for XAI Question Bank



➔ XAIs are **skewed to**: how AI systems **CAN** provide specific outputs

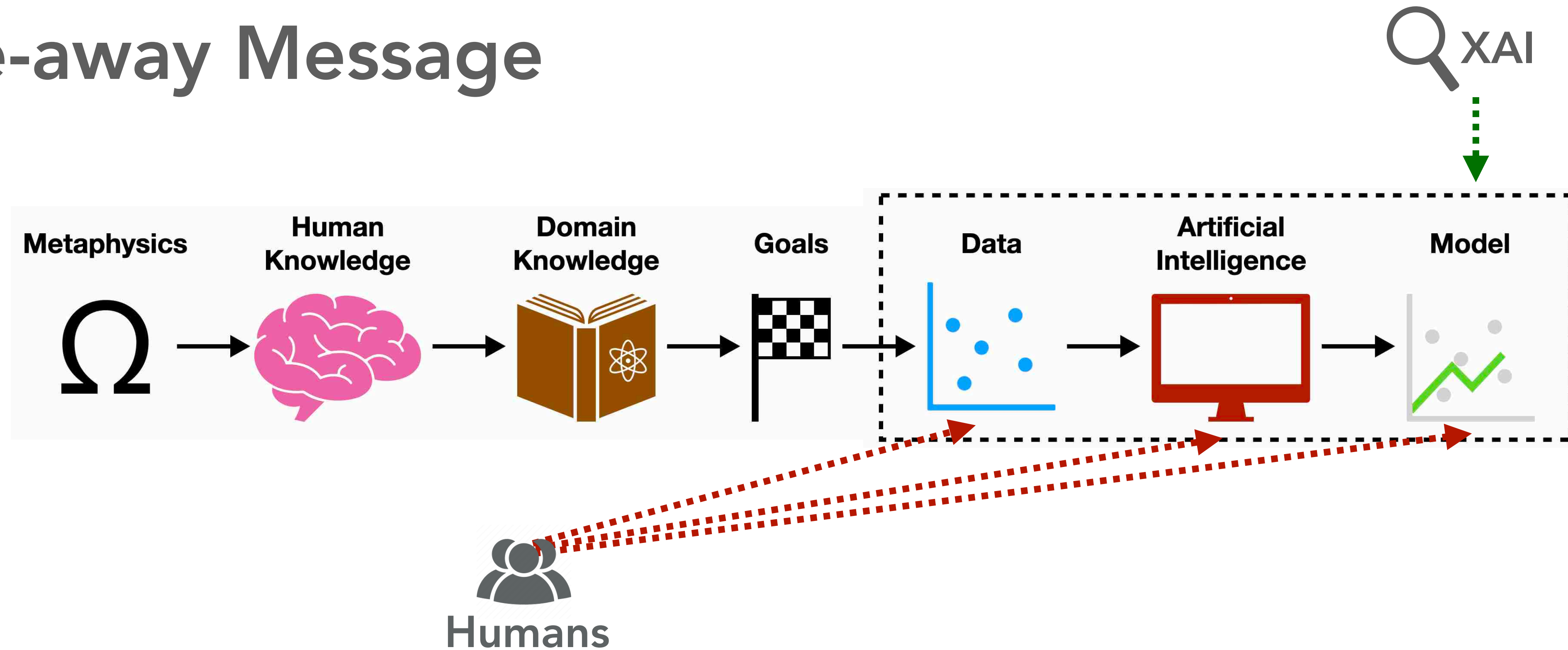
Results: Heatmap for XAI Question Bank



➔ XAI largely **ignored**: what AI systems **CANNOT** achieve (e.g., counterfactuals).

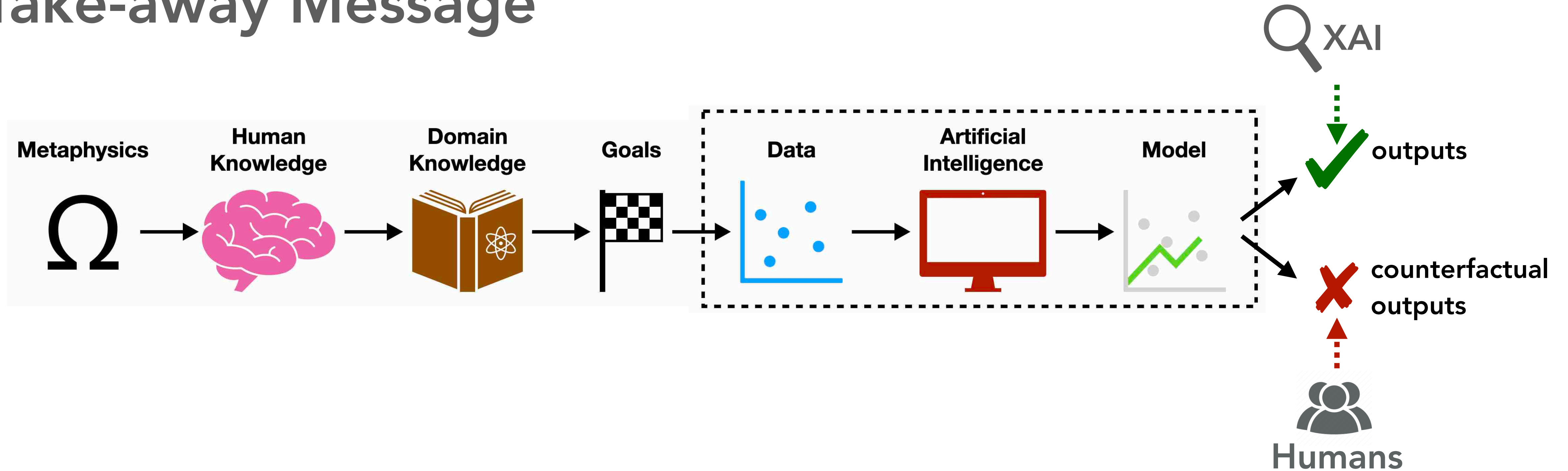
➔ XAI largely **ignored**: **diverse information** across AI development process (data, model, deployment, etc.)

Take-away Message



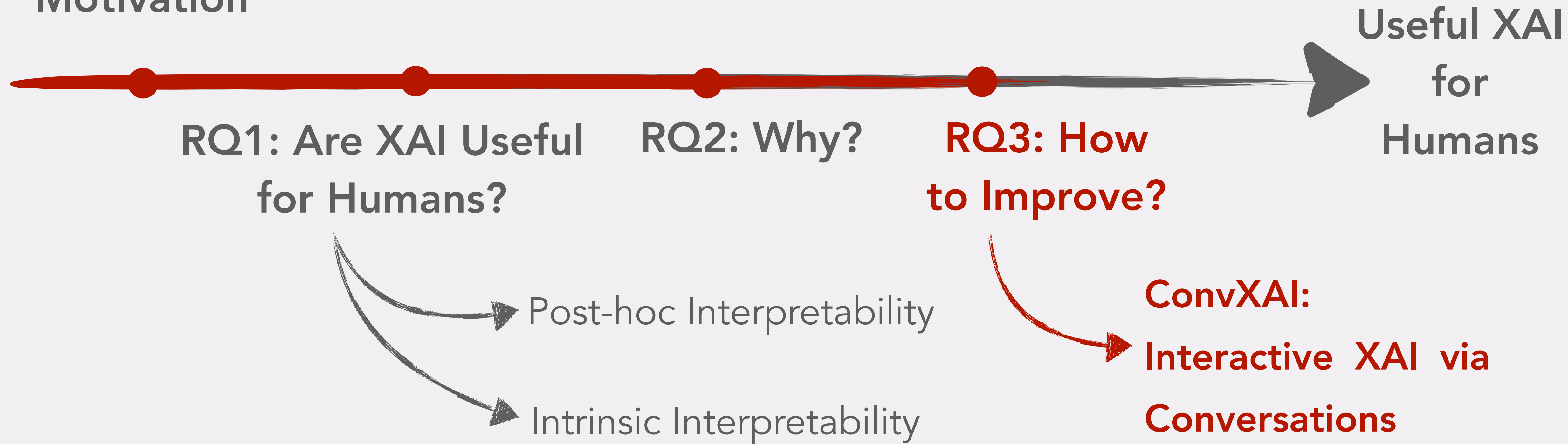
Users demand diverse XAI questions to understand AI models across AI development lifecycle, whereas existing XAI methods commonly answer only **ONE XAI** question.

Take-away Message



Users are widely interested in **what AI systems cannot achieve** other than what AI already succeeded, indicating the necessary of **interactive XAI** for **counterfactual explanations**.

Background & Motivation



ConvXAI 🗣️: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing



Hua Shen



Chieh-Yang Huang

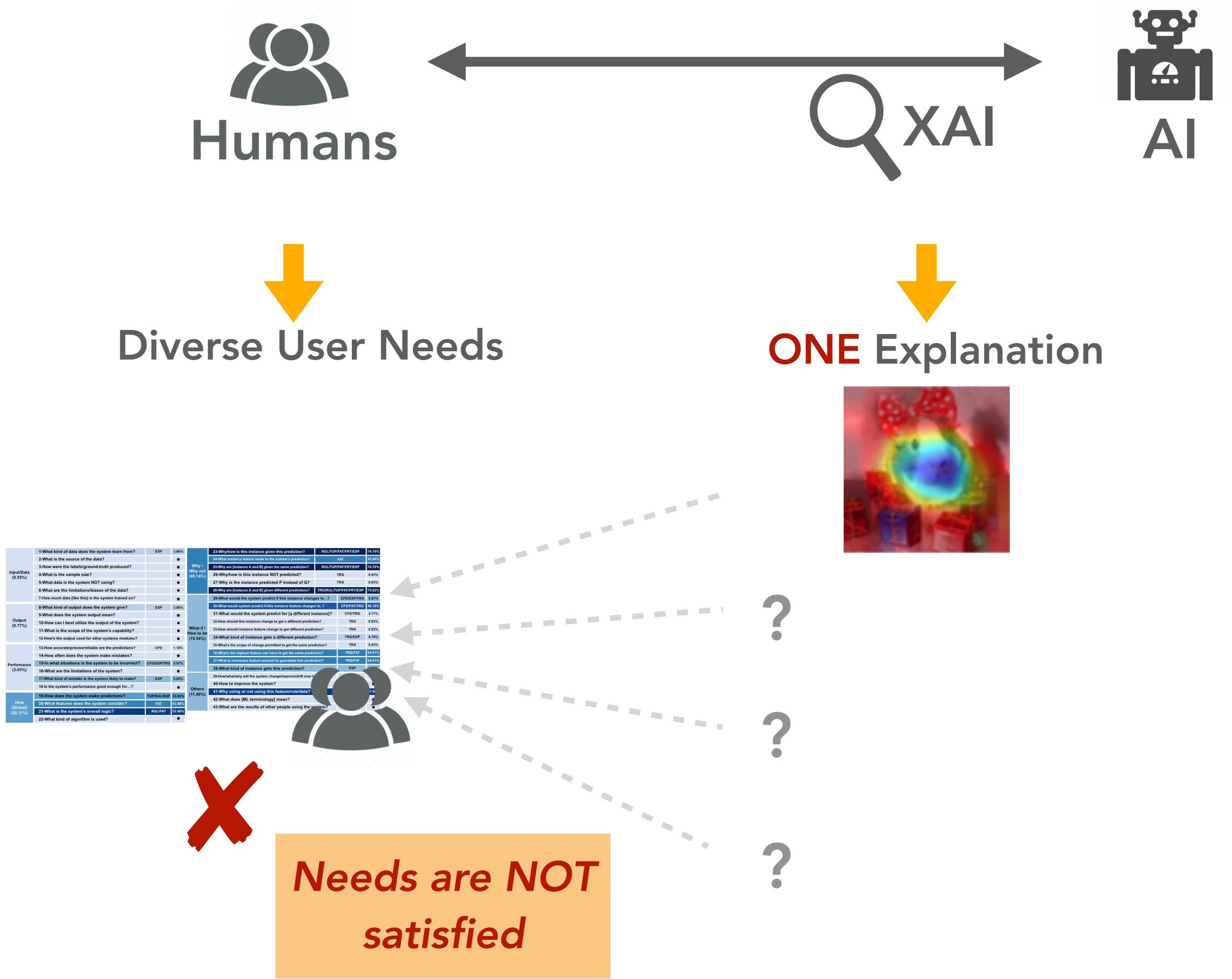


Sherry Wu



Kenneth Huang

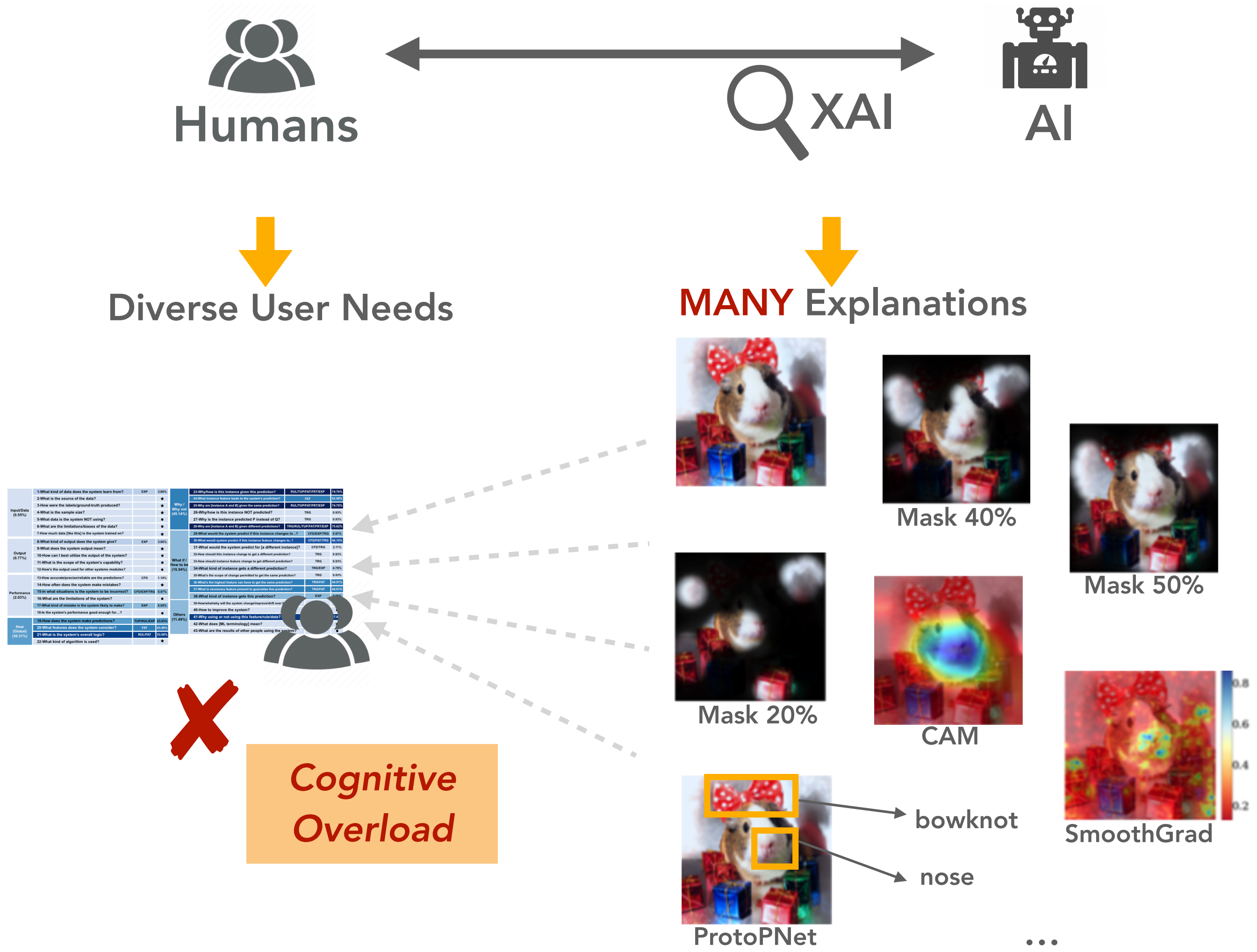
Challenges of Existing XAI



- Showing **ONE** specific **explanation** might **NOT** meet **diverse XAI user needs**.

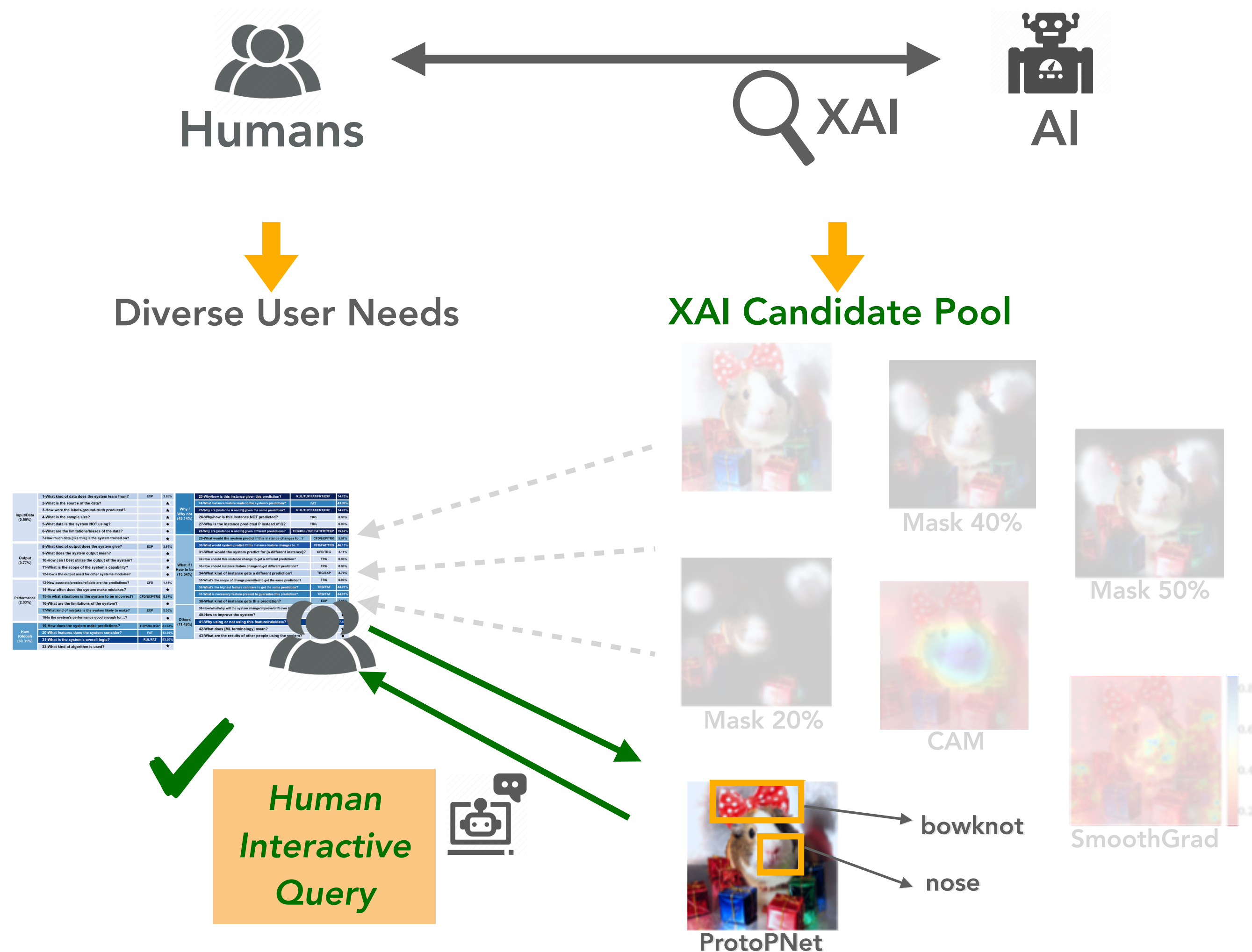
Challenges of Existing XAI

- Showing ONE specific explanation might NOT meet diverse XAI user needs.
- Showing **MANY explanations** at one time may lead to **cognitive overload** for humans



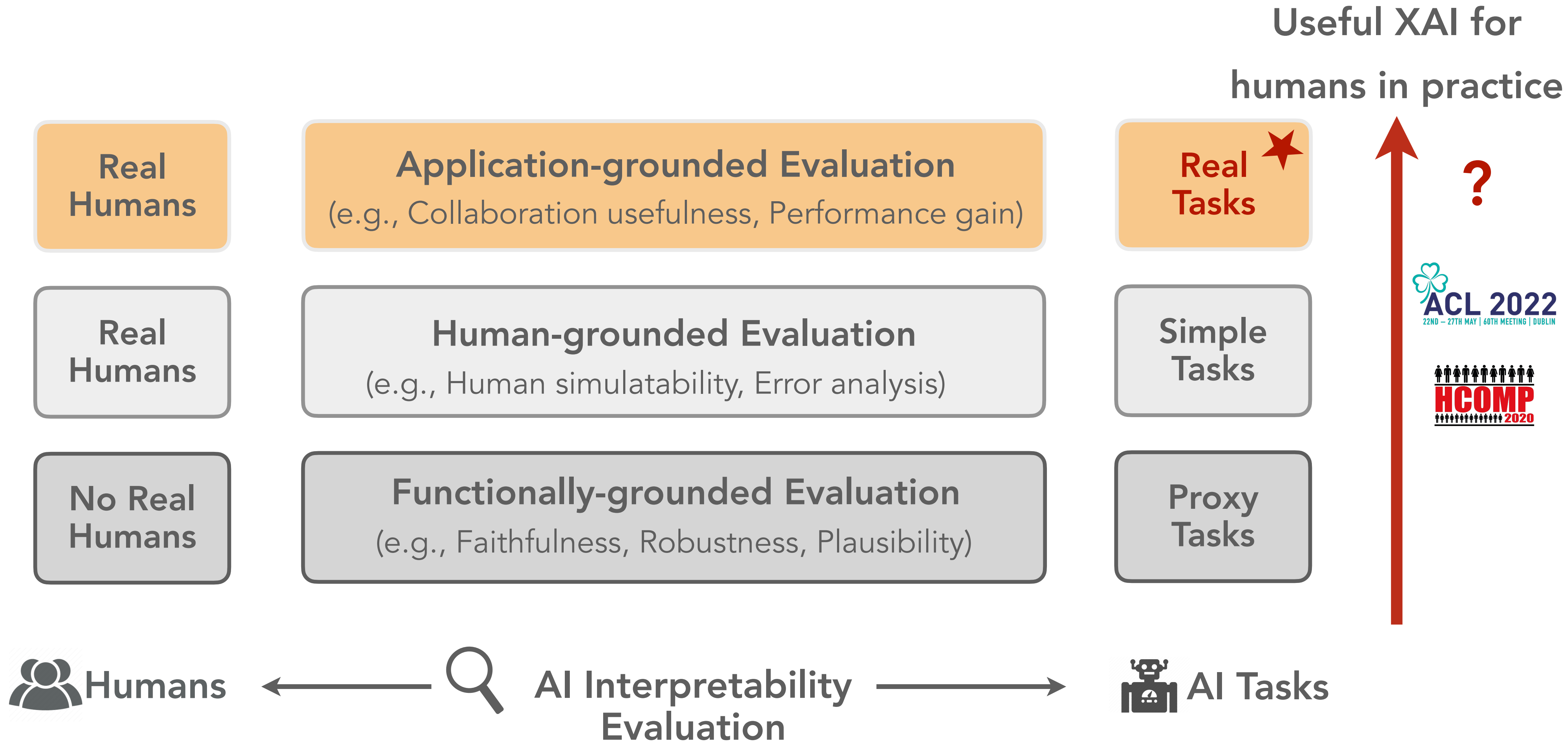
Solution: Conversational XAI

- Showing ONE specific explanation might NOT meet diverse XAI user needs.
- Showing MANY explanations at one time may lead to cognitive overload for humans



Human-centered Conversational XAI empowers humans to interactively inquire the specific explanation with minimal cognitive load.

What **Task** should we apply **Conversational XAI**?



Real Task: AI-assisted scientific writing by humans



Scientific Writing Support

Which **conference** are you most likely to **submit this paper abstract** to:

CHI (Human-Computer Interaction Domain) ▾

Select an abstract **example** to try:

Select an abstract example ▾

Or Edit **your paper abstract**:

Normal B I U Link Quote Code Image List Bulleted List

Click to Submit Your Writing

Click below buttons to **switch the model's** prediction on each sentence.

Writing Structure Model Writing Style Model

A good paper abstract should describe comprehensive research aspects, this model (i.e., a SciBERT-based) classifies each sentence into one of the five aspect labels.

Background Purpose Method Finding/Contribution Other

Reasons:

- The complex cognitive process of writing can potentially **elicit more XAIs**;
- Effectively observe **XAI usefulness** by checking human **writing improvement**;
- Common real tasks in graduate study, **easy to find real users**;

How to design the system?

Goal: *useful conversational XAI* for AI-assisted scientific writing

Human Linguistics

A Formative Study

A Preliminary System

- Predicting with 2 AI writing models
- Listing 8 XAI buttons

Participants

- 3 females, 4 males
- diverse background

Procedure

- Semi-Wizard-of-Oz (WoZ) process
- Think aloud during process

Four Design Principles for Conversational XAI

Scientific Writing Support

Which conference are you most likely to submit this paper abstract to:

CHI (Human-Computer Interaction Domain)

Select an abstract example to try:

Select an abstract example

Or Edit your paper abstract:

Normal B I U ...

While various AI explanation (XAI) methods have been designed to gain insights into AI systems , it is still hard for users to acquire the information they need .

Prior work suggested using chatbots to dynamically cater to human needs , but little has been explored about how conversational AI should be designed .

S3 aspect-purpose

In this paper , we examine the Conversational XAI potential in the context of scientific writing .

Informed by human linguistics and formative studies , we identify four design principles of Conversational XAI : address various user questions (' multifaceted ') , provide details on-demand (' controllability ') , proactively suggest and accept follow-up questions (' mix-initiative ' and ' context-aware drill-down ') .

We instantiate them into an interactive prototype , CONVXAI , which allows writers to interact with various explanations through a chatbot interface .

Through 14 user studies , we show that CONVXAI is promising for users to think through and address their diverse questions .

We conclude by discussing future conversational XAI use patterns and implications .

Click to Submit Your Writing

Click below buttons to switch the model's prediction on each sentence.

Writing Structure Model Writing Style Model

A good paper abstract should describe comprehensive research aspects, this model (i.e., a SciBERT-based) classifies each sentence into one of the five aspect labels.

Background Purpose Method Finding/Contribution Other

Predict with 2 AI writing models

Conversational AI Explanation (XAI) Assistant

Conversational Explainable AI (XAI) Assistant

Do you want some explanations on the above review?

YES No

I just selected Button [YES]

To better understand your review, you can ask me further questions about:

- **Basic Information and Statistics** of the data and model related to your selected conference, such as:

Data Statistics Model Description

Quality Score Range Aspect Distribution

- **Explanations for Each Sentence Prediction.** You can select (by double click) the specific sentence, then click the question buttons or directly type your questions below.

Model Confidence Similar Examples

Important Words Counterfactual Explanation

You can ask below XAI questions for the selected sentence: @

Send

List 8 XAI buttons for human choice

Four Design Principles for useful conversational XAI

P1

Multifaceted XAI

Contain multiple XAI types that explain AI from various aspects

P2

Mixed-Initiative

Proactively send users XAI tutorials or hints to teach them "how to use XAIs"

P3

Context-aware Drill-down

Maintain the conversation history to generate responses with user needs

P4

Controllability

Enable humans to customize XAI with personalized needs

ConvXAI Demo:

Which conference are you most likely to submit this paper abstract to:

CHI (Human-Computer Interaction Domain) v

Select an abstract example to try:

Select an abstract example v

Or Edit your paper abstract:

Normal : B I S U 🔗 ” ⏪ 📄 ☰

While various AI explanation (XAI) methods have been designed to gain insights into AI systems , it is still hard for users to acquire the information they need .

Prior work suggested using chatbots to dynamically cater to human needs , but little has been explored about how conversational AI should be designed .

S3:aspect=purpose
In this paper , we examine the Conversational XAI potential in the context of scientific writing .

Informed by human linguistics and formative studies , we identify four design principles of Conversational XAI : address various user questions (' multifaceted ') , provide details on-demand (' controllability ') , proactively suggest and accept follow-up questions (' mix-initiative ' and ' context-aware drill-down ') .

We instantiate them into an interactive prototype , CONVXAI , which allows writers to interact with various explanations through a chatbot interface .

Through 13 user studies , we show that 9 out of 13 participants preferred CONVXAI over the static interface baseline SelectXAI .

CONVXAI is promising to help users think through and address their diverse questions .

We are also aware of the limitation of CONVXAI , such as a steeper learning curve than baseline .

We conclude by discussing implications and challenges of conversational XAI systems .

Click to Submit Your Writing

Click below buttons to **switch the model's prediction** on each sentence.

Writing Structure Model

Writing Style Model

A good paper abstract should describe comprehensive research aspects, this model (i.e., a SciBERT-based) classifies each sentence into one of the five aspect labels.

Background
Purpose
Method
Finding/Contribution
Other

Conversational Explainable AI (XAI) Assistant

To improve, you can check the most important words resulting in the prediction and further check how to revise input into another label . See XAI questions below:


Label Distribution

Prediction Confidence

Similar Published Sentences

Which words are most important for this prediction?

How can I revise the sentence to get a different label?



How are the structure labels distributed?

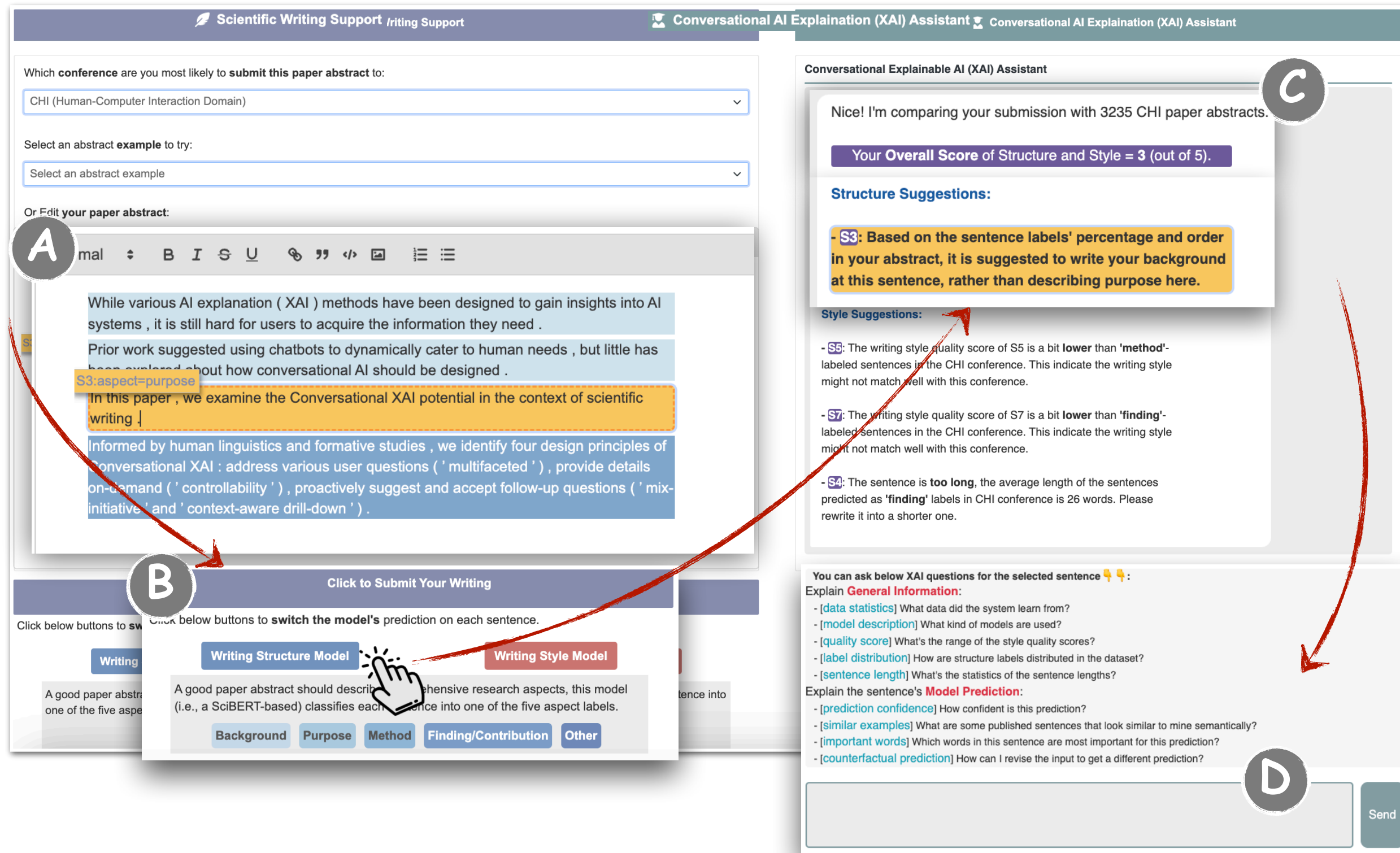
We use the Research Aspects Model to generate **aspect sequences** of all 9935 paper abstracts. Then we cluster these sequences into **five patterns** as below. We compare your writing with these patterns for review.

Types	Patterns
Pattern1	'background' (42.9%) -> 'purpose' (14.3%) -> 'finding' (42.9%)
Pattern2	'background' (22.2%) -> 'purpose' (11.2%) -> 'method' (33.3%) -> 'finding' (33.3%)
Pattern3	'background' (33.3%) -> 'purpose' (16.7%) -> 'method' (16.7%) -> 'finding' (33.3%)
Pattern4	'background' (33.3%) -> 'method' (16.7%) -> 'finding' (50%)
Pattern5	'background' (20%) -> 'finding' (6.7%) -> 'background' (13.3%) -> 'purpose' (6.7%) -> 'background' (13.3%) -> 'finding' (6.7%) -> 'method' (6.7%) -> 'finding' (26.7%)

You can ask below XAI questions for the selected sentence: ⓘ

In this paper , we examine the Conversational XAI potential in the context of scientific writing .

Send



Scientific Writing Support / Writing Support

Which conference are you most likely to submit this paper abstract to:
CHI (Human-Computer Interaction Domain)

Select an abstract example to try:
Select an abstract example

Or Edit your paper abstract:

A While various AI explanation (XAI) methods have been designed to gain insights into AI systems , it is still hard for users to acquire the information they need .
Prior work suggested using chatbots to dynamically cater to human needs , but little has been explored about how conversational AI should be designed .
S3.aspect=purpose
In this paper , we examine the Conversational XAI potential in the context of scientific writing .
Informed by human linguistics and formative studies , we identify four design principles of conversational XAI : address various user questions (' multifaceted ') , provide details on-demand (' controllability ') , proactively suggest and accept follow-up questions (' mix-initiative ' and ' context-aware drill-down ') .

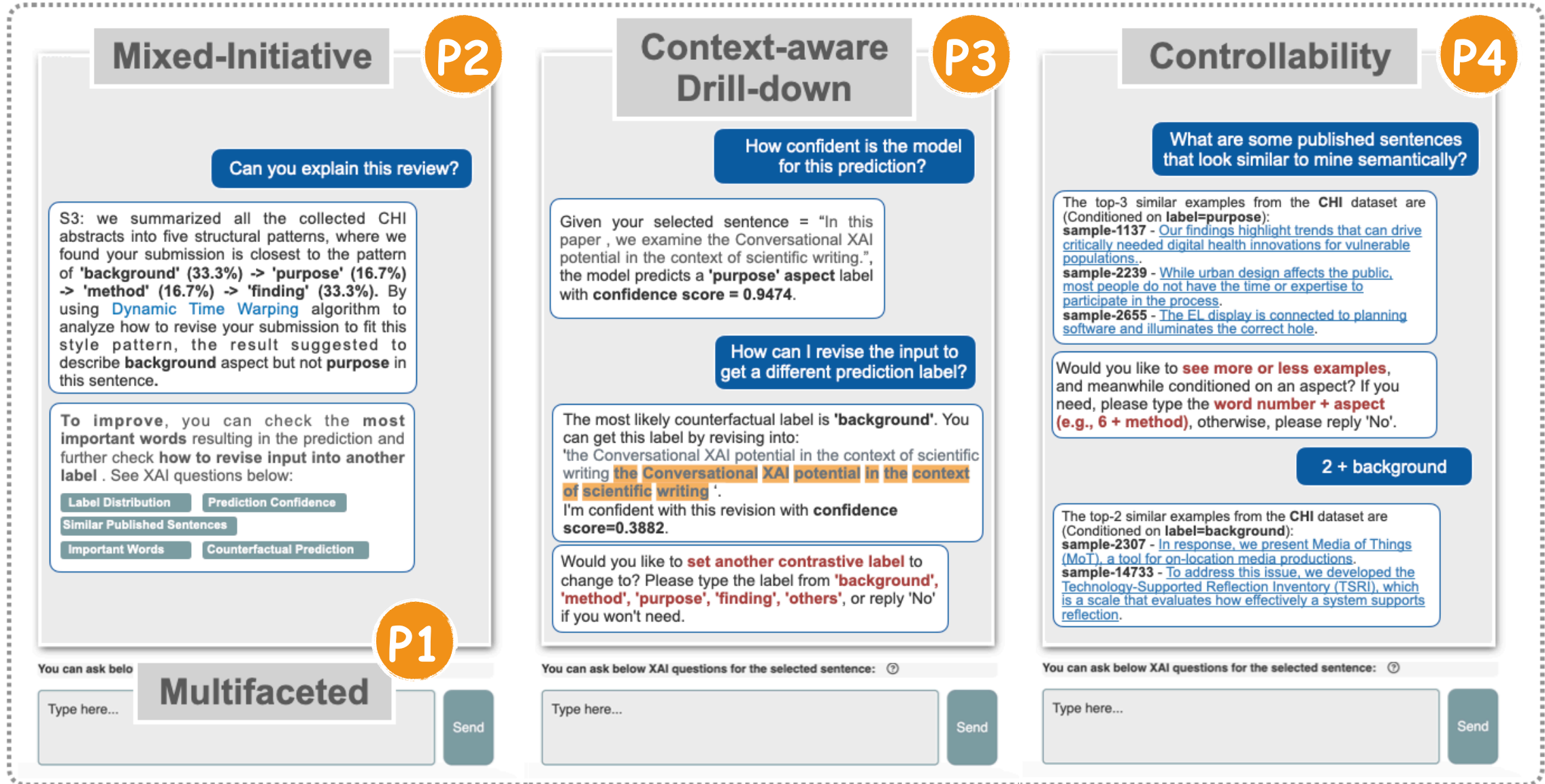
B Click to Submit Your Writing
Click below buttons to switch the model's prediction on each sentence.
Writing Structure Model Writing Style Model
Background Purpose Method Finding/Contribution Other

C Conversational Explainable AI (XAI) Assistant
Nice! I'm comparing your submission with 3235 CHI paper abstracts.
Your Overall Score of Structure and Style = 3 (out of 5).
Structure Suggestions:
- **S3:** Based on the sentence labels' percentage and order in your abstract, it is suggested to write your background at this sentence, rather than describing purpose here.
Style Suggestions:
- **S5:** The writing style quality score of S5 is a bit lower than 'method'-labeled sentences in the CHI conference. This indicate the writing style might not match well with this conference.
- **S7:** The writing style quality score of S7 is a bit lower than 'finding'-labeled sentences in the CHI conference. This indicate the writing style might not match well with this conference.
- **S4:** The sentence is too long, the average length of the sentences predicted as 'finding' labels in CHI conference is 26 words. Please rewrite it into a shorter one.

D You can ask below XAI questions for the selected sentence 🗨️:
Explain **General Information**:
- [data statistics] What data did the system learn from?
- [model description] What kind of models are used?
- [quality score] What's the range of the style quality scores?
- [label distribution] How are structure labels distributed in the dataset?
- [sentence length] What's the statistics of the sentence lengths?
Explain the sentence's **Model Prediction**:
- [prediction confidence] How confident is this prediction?
- [similar examples] What are some published sentences that look similar to mine semantically?
- [important words] Which words in this sentence are most important for this prediction?
- [counterfactual prediction] How can I revise the input to get a different prediction?

Keynotes of User Interaction

Implementing Four Design Principles



Mixed-Initiative P2
Can you explain this review?
S3: we summarized all the collected CHI abstracts into five structural patterns, where we found your submission is closest to the pattern of 'background' (33.3%) -> 'purpose' (16.7%) -> 'method' (16.7%) -> 'finding' (33.3%). By using Dynamic Time Warping algorithm to analyze how to revise your submission to fit this style pattern, the result suggested to describe background aspect but not purpose in this sentence.
To improve, you can check the most important words resulting in the prediction and further check how to revise input into another label . See XAI questions below:
Label Distribution Prediction Confidence
Similar Published Sentences
Important Words Counterfactual Prediction

Context-aware Drill-down P3
How confident is the model for this prediction?
Given your selected sentence = "In this paper , we examine the Conversational XAI potential in the context of scientific writing.", the model predicts a 'purpose' aspect label with confidence score = 0.9474.
How can I revise the input to get a different prediction label?
The most likely counterfactual label is 'background'. You can get this label by revising into: 'the Conversational XAI potential in the context of scientific writing the Conversational XAI potential in the context of scientific writing'.
I'm confident with this revision with confidence score=0.3882.
Would you like to set another contrastive label to change to? Please type the label from 'background', 'method', 'purpose', 'finding', 'others', or reply 'No' if you won't need.

Controllability P4
What are some published sentences that look similar to mine semantically?
The top-3 similar examples from the CHI dataset are (Conditioned on label=purpose):
sample-1137 - Our findings highlight trends that can drive critically needed digital health innovations for vulnerable populations.
sample-2239 - While urban design affects the public, most people do not have the time or expertise to participate in the process.
sample-2655 - The EL display is connected to planning software and illuminates the correct hole.
Would you like to see more or less examples, and meanwhile conditioned on an aspect? If you need, please type the word number + aspect (e.g., 6 + method), otherwise, please reply 'No'.
2 + background
The top-2 similar examples from the CHI dataset are (Conditioned on label=background):
sample-2307 - In response, we present Media of Things (MoT), a tool for on-location media productions.
sample-14733 - To address this issue, we developed the Technology-Supported Reflection Inventory (TSRI), which is a scale that evaluates how effectively a system supports reflection.

Multifaceted P1
Type here... Send

Technical Challenges & Contributions

Challenges:

- No unified approach to incorporate various XAI types into one interface
- No existing XAI approaches to parse interactive user needs and generate customized XAIs

Technical Challenges & Contributions

Challenges:

- No unified approach to incorporate various XAI types into one interface
- No existing XAI approaches to parse interactive user needs and generate customized XAIs

Technical Contribution

- A Unified conversational XAI API for various XAI types that enable user to customize AI explanations.

```
In [3]: """Human-ConvXAI Interaction with the unified API"""  
  
### 10 Types of AI Explanation Questions  
user_xai_questions = {  
    "global-ask-data": "What data did the system learn from?",  
    "global-ask-model": "What kind of models are used?",  
    "global-ask-quality-score": "What's the range of the style quality scores?",  
    "global-ask-label-distribution": "How are structure labels distributed in the dataset?",  
    "global-ask-sent-length": "What's the statistics of the sentence lengths?",  
    "local-ask-model-confidence": "How confident is this prediction?",  
    "local-ask-xai-example": "What are some published sentences that look similar to mine semantically?",  
    "local-ask-feature-attribution": "Which words in this sentence are most important for this prediction?",  
    "local-ask-counterfactual": "How can I revise the input to get a different prediction?",  
    "others": "Who are you?"  
}
```

Global AI Explanations -- generating AI explanation for meta information

[XAI Type 1] - global AI explanation for describing dataset

```
In [4]: user_xai_question = user_xai_questions['global-ask-data']  
response = convxai_agent.explain(  
    user_xai_question,  
    ai_input,  
    ai_predict_output,  
    conference,  
    visualize=visualize  
)
```

what data did the system learn from?

Sure! We are comparing your writing with our collected **CHI Paper Abstract** dataset to generate the above review. The dataset includes **21643 sentences** in **3235 papers**.

[XAI Type 2] - global AI explanation for describing models

```
In [5]: user_xai_question = user_xai_questions['global-ask-model']  
response = convxai_agent.explain(  
    user_xai_question,  
    ai_input,  
    ai_predict_output,  
    conference,  
    visualize=visualize  
)
```

what kind of models are used?

Of course! The **Writing Structure Model** is a **SciBERT** based classifier finetuned with the **CODA-19** dataset. Also, the **Writing Style Model** is a **GPT-2** based generative model finetuned with **9935 abstracts** from **CHI**, **ACL** and **ICLR** papers (click the terms to view more).

[XAI Type 3] - global AI explanation for describing quality scores

Evaluate ConvXAI with real human studies

	 Task1	 Task2
Who is studies	 13 graduate researchers	 8 researchers
When	09/2022 (90min)	12/2022 (90min) (rejoin)
How it's studied	<ol style="list-style-type: none">Two think-aloud scientific writing tasks:<ul style="list-style-type: none">Within-Subjects Study: ConvXAI vs. BaselineImprove a paper's abstract;Paper domains: NLP, or HCI, or AIPost Survey - QuestionnairesSemi-structured Interviews	
What is the goal	<ul style="list-style-type: none">Effective system designSelf-perceived usefulness	<ul style="list-style-type: none">Writing output qualityTemporal change of user need

Baseline System (SelectXAI)

Within-Subjects Study Design

The image displays a composite interface for 'Scientific Writing Support' and an 'AI Explanation (XAI) Panel'. The left panel, titled 'Scientific Writing Support', includes a dropdown menu for selecting a conference (currently 'CHI (Human-Computer Interaction Domain)'), a dropdown for selecting an abstract example, and a rich text editor for editing the paper abstract. The right panel, titled 'AI Explanation (XAI) Panel', shows a 'Writing Feedback Summary' with an overall score of 3 out of 5 and a list of 'Structure Suggestions'. A 'Data Statistics' tooltip is visible over the 'Data Statistics' button in the XAI panel, providing context about the dataset used for the review. A 'Click to Submit Your Writing' dialog box is also present, showing options to switch between 'Writing Structure Model' and 'Writing Style Model', and a list of aspect labels: Background, Purpose, Method, Finding/Contribution, and Other. Hand icons and red arrows highlight specific elements like the 'Data Statistics' tooltip and the 'Sentence-wise Explanations' button.

Scientific Writing Support

Which **conference** are you most likely to **submit this paper abstract** to:

CHI (Human-Computer Interaction Domain)

Select an abstract **example** to try:

Select an abstract example

Or Edit your paper abstract:

Normal B I U Link Quote Code Image List

While various AI explanation (XAI) methods have been design systems , it is still hard for users to acquire the information they Prior work suggested using chatbots to dynamically cater to hu been explored about how conversational AI should be design

S3:aspect=purpose

In this paper , we examine the Conversational XAI potential in the context of scientific writing .

Informed by human linguistics and formative studies , we identify four design principles of Conversational XAI : address various user questions (' multifaceted ') , provide details on-demand (' controllability ') , proactively suggest and accept follow-up questions (' mix-initiative ' and ' context-aware drill-down ') .

AI Explanation (XAI) Panel

Writing Feedback Summary

Nice! I'm comparing your submission with 3235 CHI paper abstracts.

Your **Overall Score** of Structure and Style = 3 (out of 5).

Structure Suggestions:

St

Based on the sentence labels' percentage and order listed to write your background describing purpose here.

Data Statistics (What data did the system learn from?)

Sure! We are comparing your writing with our collected CHI Paper Abstract dataset to generate the above review. The dataset includes 21643 sentences in 3235 papers.

Sentence-wise Explanations

Data Statistics (What data did the system learn from?)

Model Description (What kind of models are used?)

Quality Score (What's the range of the style quality scores?)

Aspect Distribution (How are the structure labels distributed?)

Sentence Length (What's the statistics of the sentence lengths?)

Prediction Confidence (How confident is the model for this prediction?)

Similar Examples (What are the most similar examples in the trainset?)

Important Words (Which words in this sentence are most important for this prediction?)

Counterfactual Inputs (How can I revise the input to get a different prediction label?)

Click to Submit Your Writing

Click below buttons to **switch the model's** prediction on each sentence.

Click below button

Writing Structure Model **Writing Style Model**

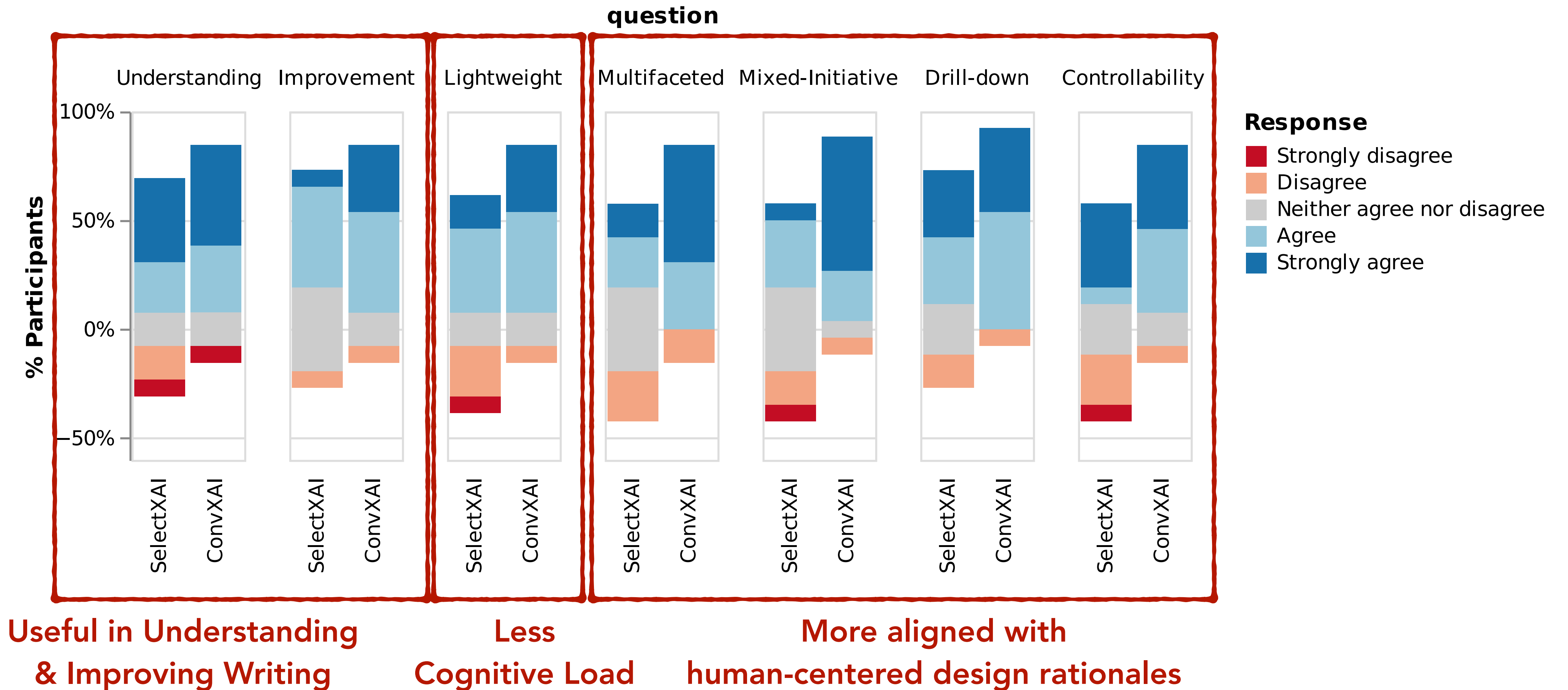
A good paper abstract should describe comprehensive research aspects, this model (i.e., a SciBERT-based) classifies each sentence into one of the five aspect labels.

Background **Purpose** **Method** **Finding/Contribution** **Other**

Background Purpose Method Finding/Contribution Other

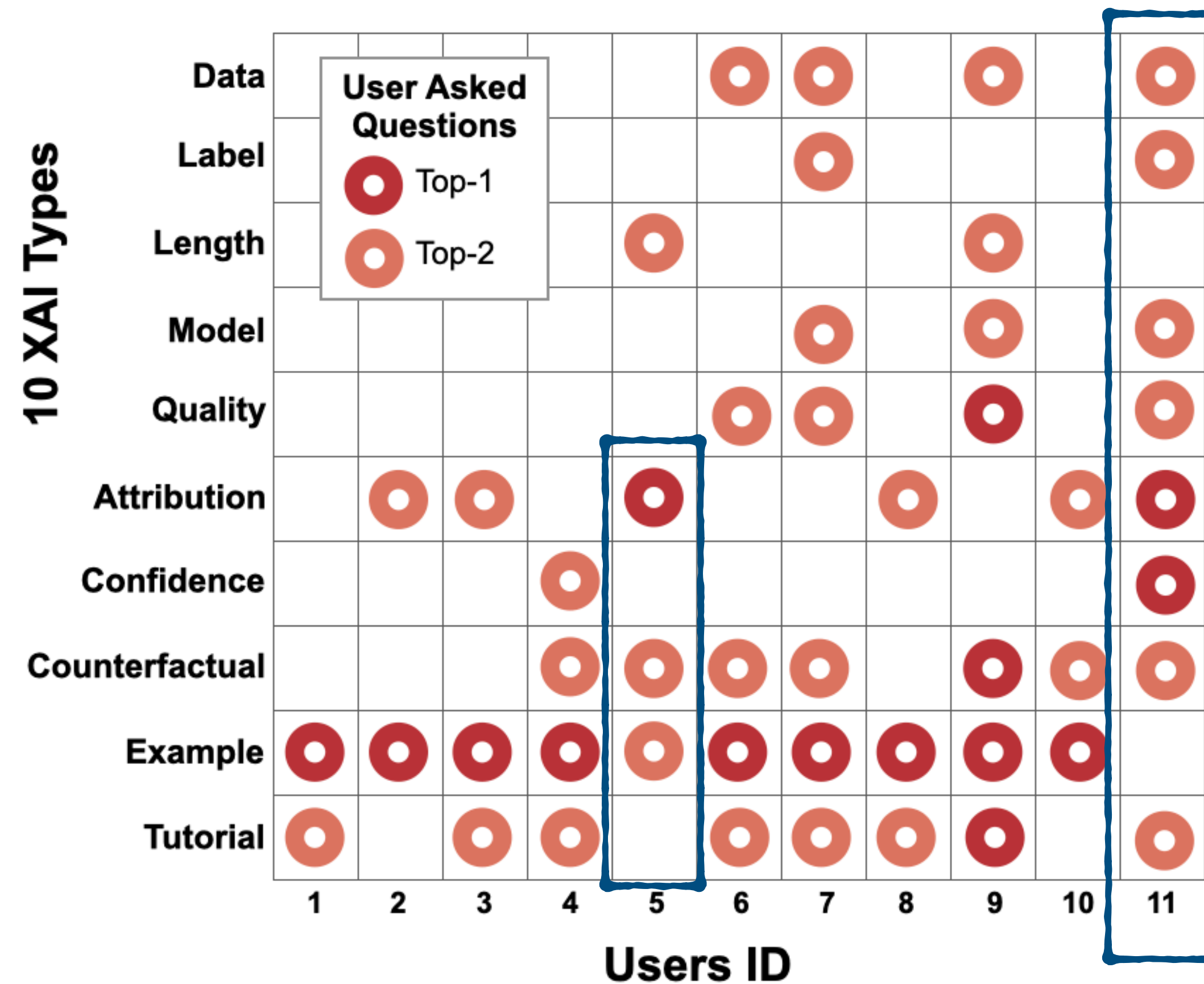
Survey results of human study in Task1

Finding#1: **ConvXAI is a useful approach** to help end users understand and collaborate with AI models.



Humans' XAI usage patterns in Task1

Finding#2: **Different users prefer to use different XAI formats** in the real-world tasks.



Usefulness evaluation in Task2

Finding#3: ConvXAI can improve **human-perceived usefulness**.

Condition ↑	Overall Writing	Writing Structure	Writing Quality
SelectXAI	3.25 (±1.035)	3.375 (±1.302)	3 (±1.195)
ConvXAI	4.25 (±1.389)	4.375 (±1.408)	4 (±1.414)
P	0.1248	0.1624	0.1489

Table1. Survey results of human-perceived usefulness rating.

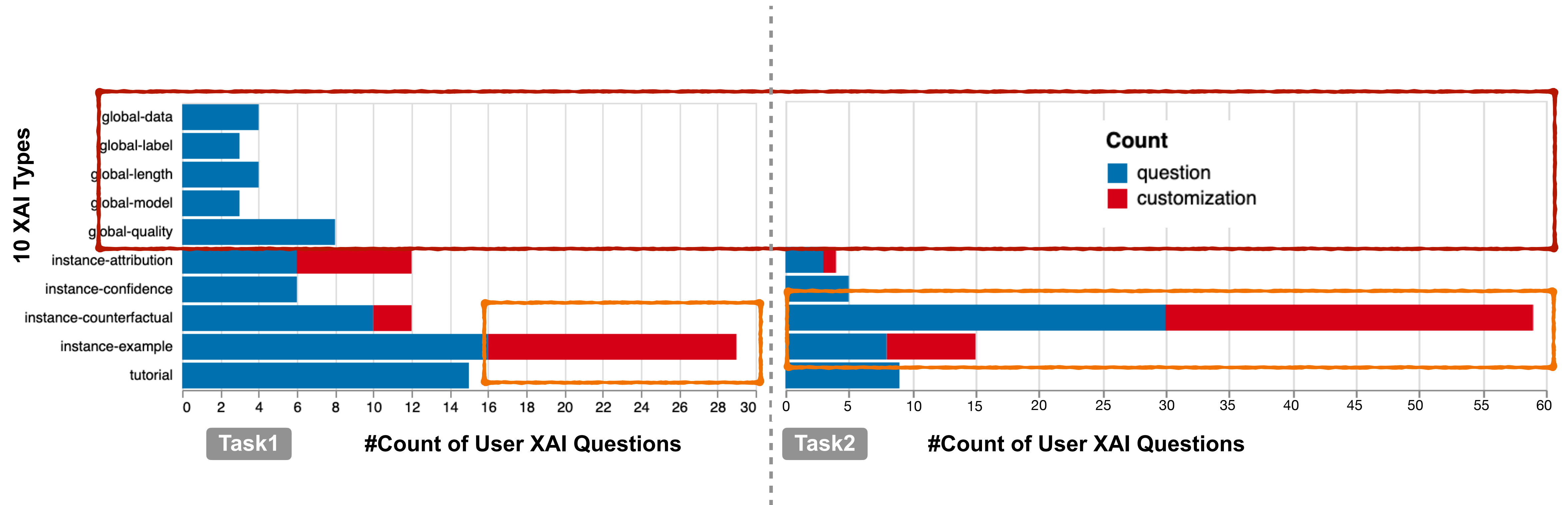
Condition ↑	Grammarly (1-100)		Model Quality (1-5)		Model Structure (1-5)		Human Quality (1-10)		Human Structure (1-10)	
	Original	Improved	Original	Improved	Original	Improved	Original	Improved	Original	Improved
SelectXAI	84.8 (±10.4)	85.1 (±5.52)	2.82 (0.75)	3.05 (0.64)	4.19 (0.37)	4.75 (0.38)	6.5 (1.69)	6.50 (1.30)	6.5 (1.07)	6.63 (1.19)
ConvXAI		86.6 (±6.50)		3.18 (0.71)		4.31 (0.46)		6.38 (0.93)		6.63 (1.19)
P	-	0.6264	-	0.6965	-	0.0560	-	0.8281	-	1.00

Table2. Objective scores of evaluating usefulness in task 2.

Finding#4: But ConvXAI **didn't always improve objective writing performance**

Task1 v.s. Task2: user needs changed along time

Finding#5: **Users XAI needs changed along time** and converged to instance-wise XAIs.



Finding#6: User-oriented **XAI Customization is important** in many XAI types.

Take-away Message

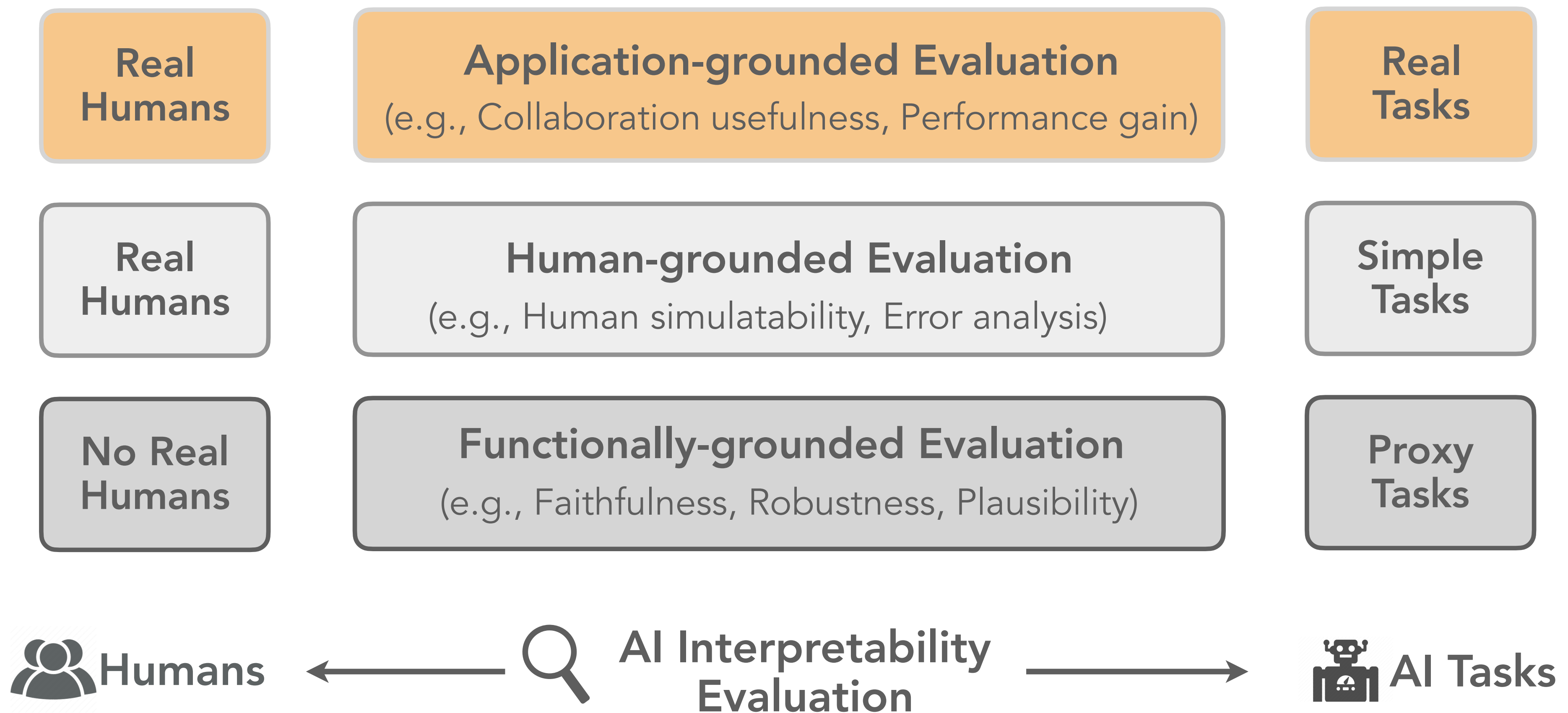
ConvXAI is a potentially useful human-centered XAI approach that empowers humans to interactively inquire heterogeneous AI Explanations via a simple conversation interface.

Key ingredients for useful XAI for humans in practice:

- A simple human interactive interface
- Multifaceted XAIs
- XAI customization
- Proactive XAI tutorials/hints for usage

Evaluation of XAI usefulness

★ Useful XAI for humans in practice



CSCW 2023 Demo

ACL 2022
22ND - 27TH MAY | 60TH MEETING | DUBLIN

HCOMP 2020



RQ1: Are XAI Useful
for Humans?

RQ2: Why?

RQ3: How
to Improve?

Useful XAI
for
Humans

Limitation

- In real world human-AI tasks, “how to **quantify** human’s **subjective** goal of **XAI usefulness**, and **align** it with **objective AI predictions**” is still **challenging**.

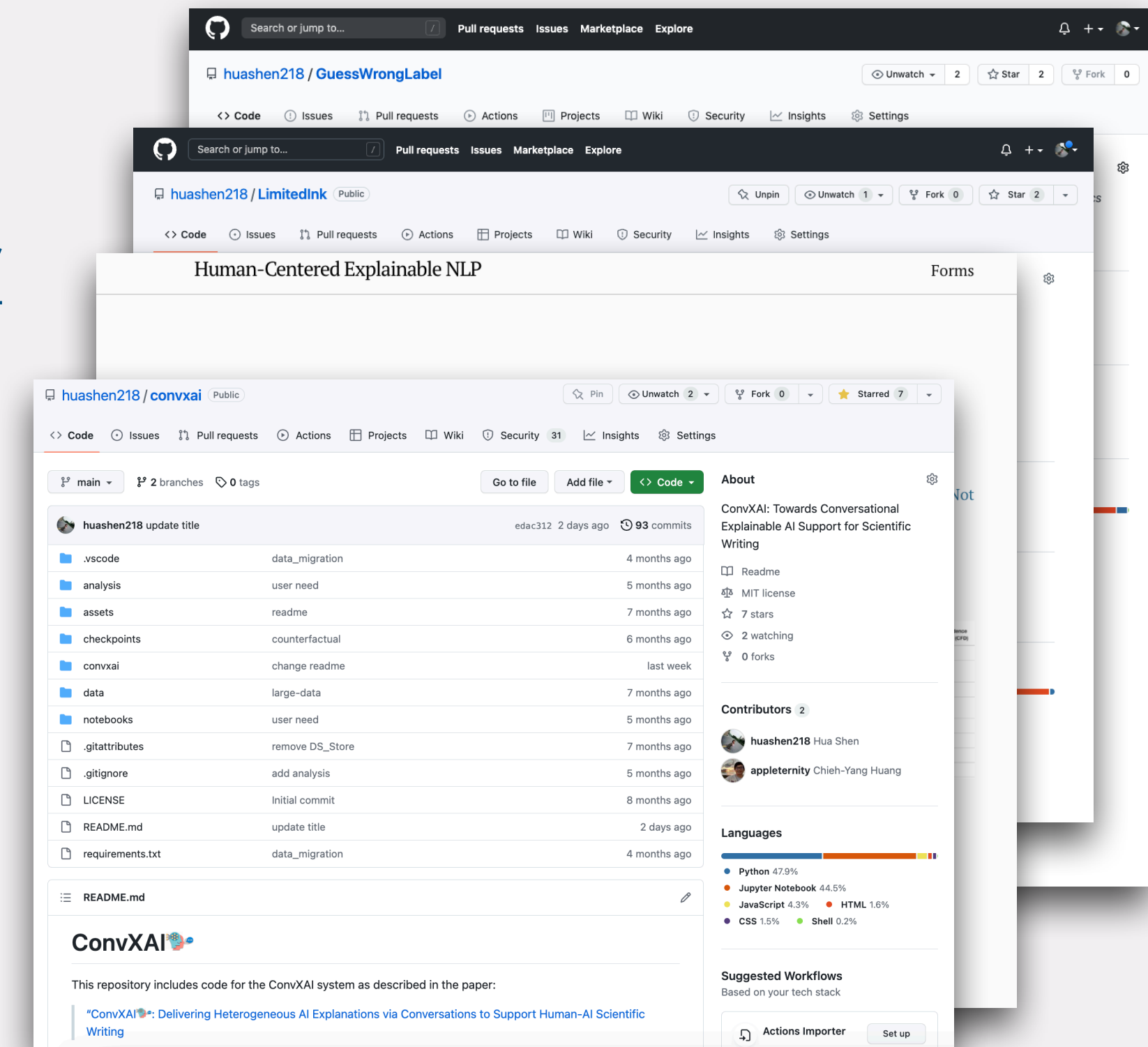
Future Work

- Develop the **benchmark** for evaluating **XAI usefulness metrics**;
- Generalize the **interactive XAI customization** for user need is crucial;
- Extend the Interactive AI Explanations to **more real-world tasks** (e.g., healthcare);







All projects are open-sourced on Github

- Human Evaluation for Model Errors: <https://github.com/huashen218/GuessWrongLabel>
- LimitedInk Model & Human Study codes: <https://github.com/huashen218/LimitedInk.git>
- 200+ Paper website & Annotations: <https://human-centered-exnlp.github.io/>
- ConvXAI System codes: <https://github.com/huashen218/convxai.git>
- The Unified XAI API: https://github.com/huashen218/convxai/blob/main/notebooks/convxai_universal_xai_api.ipynb



Other papers outside this talk (2020 - 2023)

- [1] [Hua Shen](#), Sherry Wu, Parachute: Evaluating Interactive Human-LM Co-writing Systems. CHI In2Writing Workshop, 2023. **Human-AI Eval Framework**
- [2] Sherry Wu, [Hua Shen](#), Daniel S Weld, Jeffrey Heer, Marco Tulio Ribeiro. ScatterShot: Interactive In-context Example Curation for Text Transformation. IUI, 2023. ( **Best paper award, Honorable Mention**)  **Human interaction/evaluation on AI systems**
- [3] [Hua Shen](#), Vicky Zayats, Johann C. Rocholl, Daniel D. Walker, Dirk Padfield. MultiTurnCleanup: A Benchmark for Multi-Turn Spoken Conversational Transcript Cleanup. Arxiv, 2023. (**Google AI Intern**) **Collect a Dataset with Crowdsourcing**
- [4] [Hua Shen](#)*, Yuguang Yang*, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, Andreas Stolcke. Improving Fairness in Speaker Verification via Group-adapted Fusion Network. ICASSP, 2022. (**Amazon AI Intern**) **Speech Fairness**
- [5] [Hua Shen](#)*, Adaku Uchendu*, Jooyoung Lee*, Thai Le, Kenneth Huang, Dongwon Lee.  Does Human Collaboration Enhance the Accuracy of Identifying Deepfake Texts?. Arxiv, 2023.
- [6] Shih-Hong Huang, Chieh-Yang Huang, Yuxin Deng, [Hua Shen](#), Szu-Chi Kuan, Kenneth Huang.  Too Slow to Be Useful? On Incorporating Humans in the Loop of Smart Speakers. HCOMP 2022 WiP/Demo.
- [7] Jiaqi Wang, [Hua Shen](#), Chacha Chen, Frank E. Ritter. Are Learners Satisfied with Their MOOC Experiences? Assessing and Improving Online Learners' Interactions. Asian CHI Symposium, 2021.
- [8] Ren Pang, [Hua Shen](#), Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex X. Liu, Ting Wang. The Tale of Evil Twins: Adversarial Inputs versus Poisoned Models. ACM CCS, 2020. **XAI robustness & security**
- [9] Xinyang Zhang, Ningfei Wang, [Hua Shen](#), Shouling Ji, Ting Wang. Interpretable Deep Learning under Fire. USENIX Security Symposium, 2020.

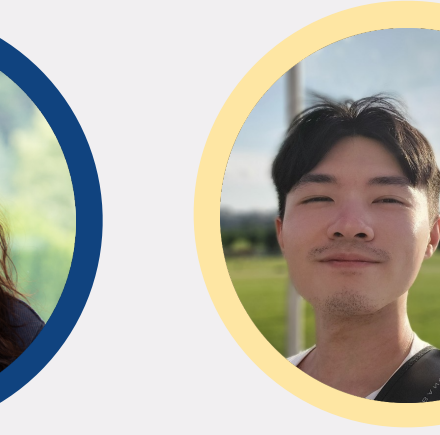
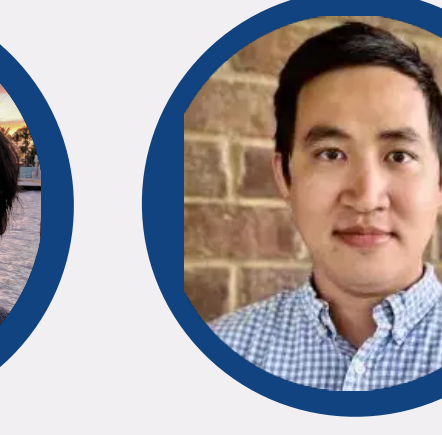
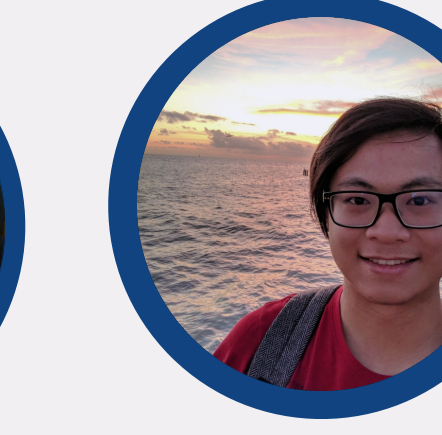
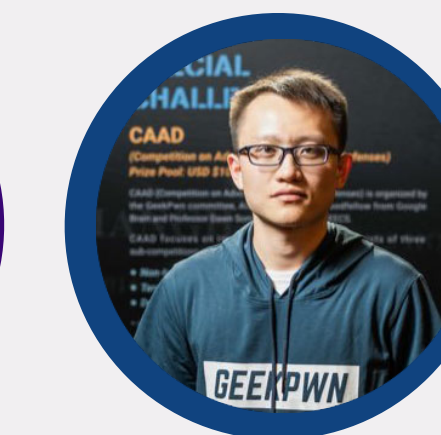
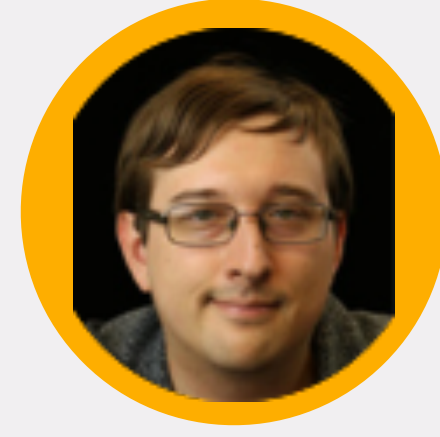
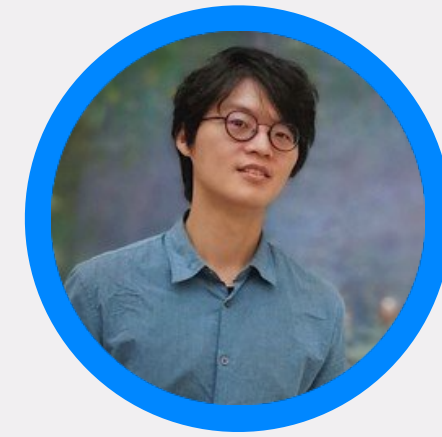
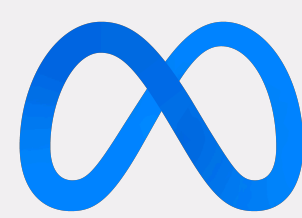
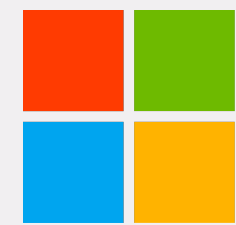


Carnegie Mellon University



Committee Members

Mentors & Collaborators



thank you!