

# Towards Useful AI Interpretability for Humans via Interactive AI Explanations

Hua Shen

 @huashen218  huashen218@psu.edu

*Penn State University*

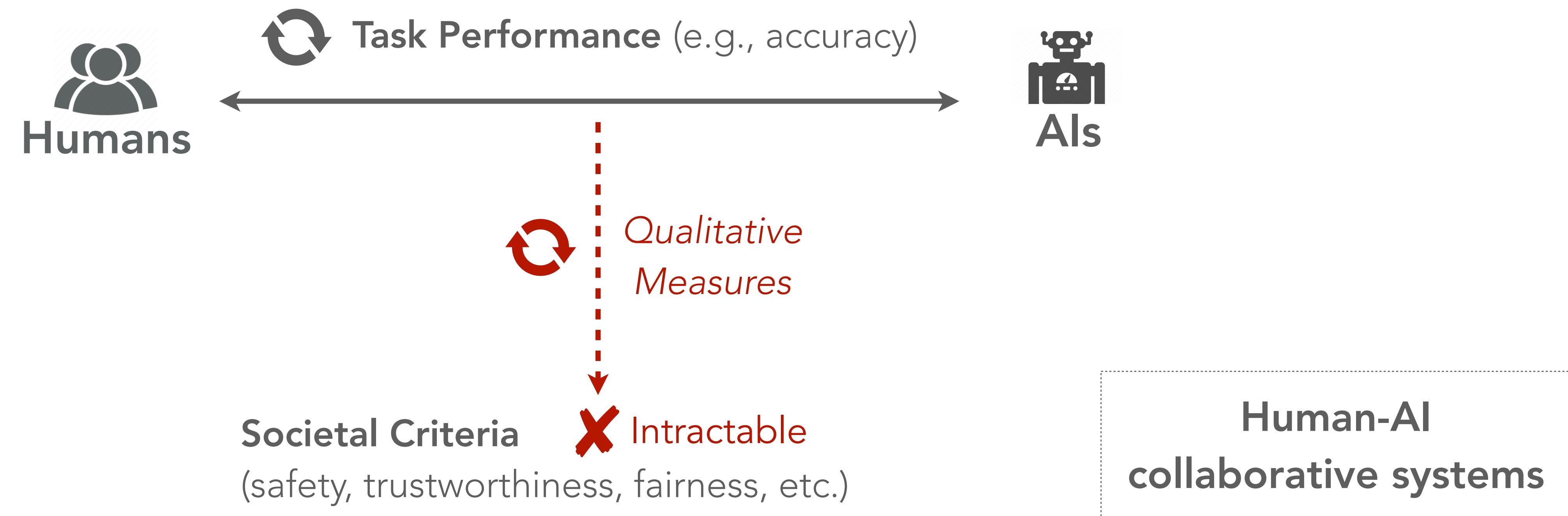
# Background & Motivation



★ Useful XAI  
for  
Humans

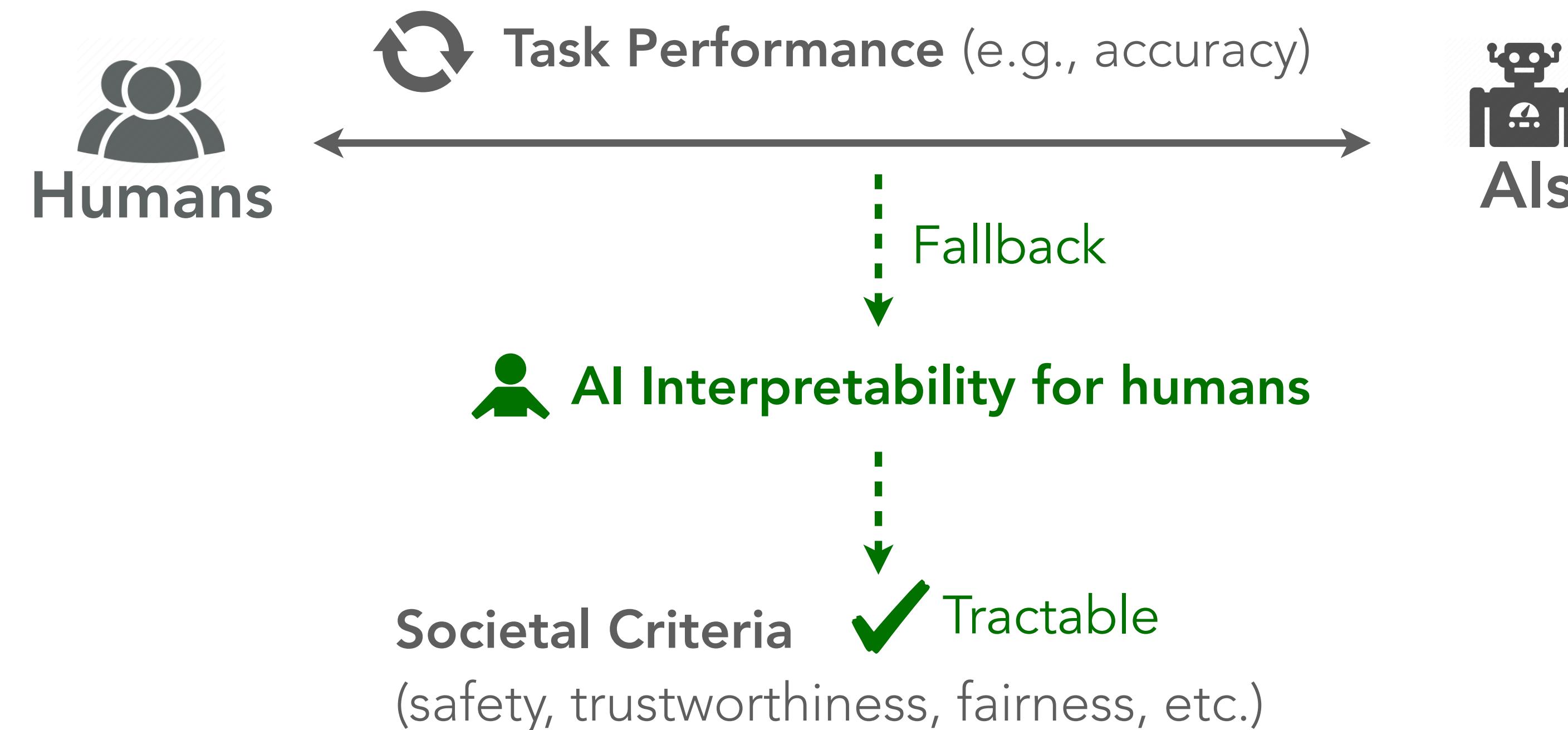
# Why do we need AI interpretability?

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.).

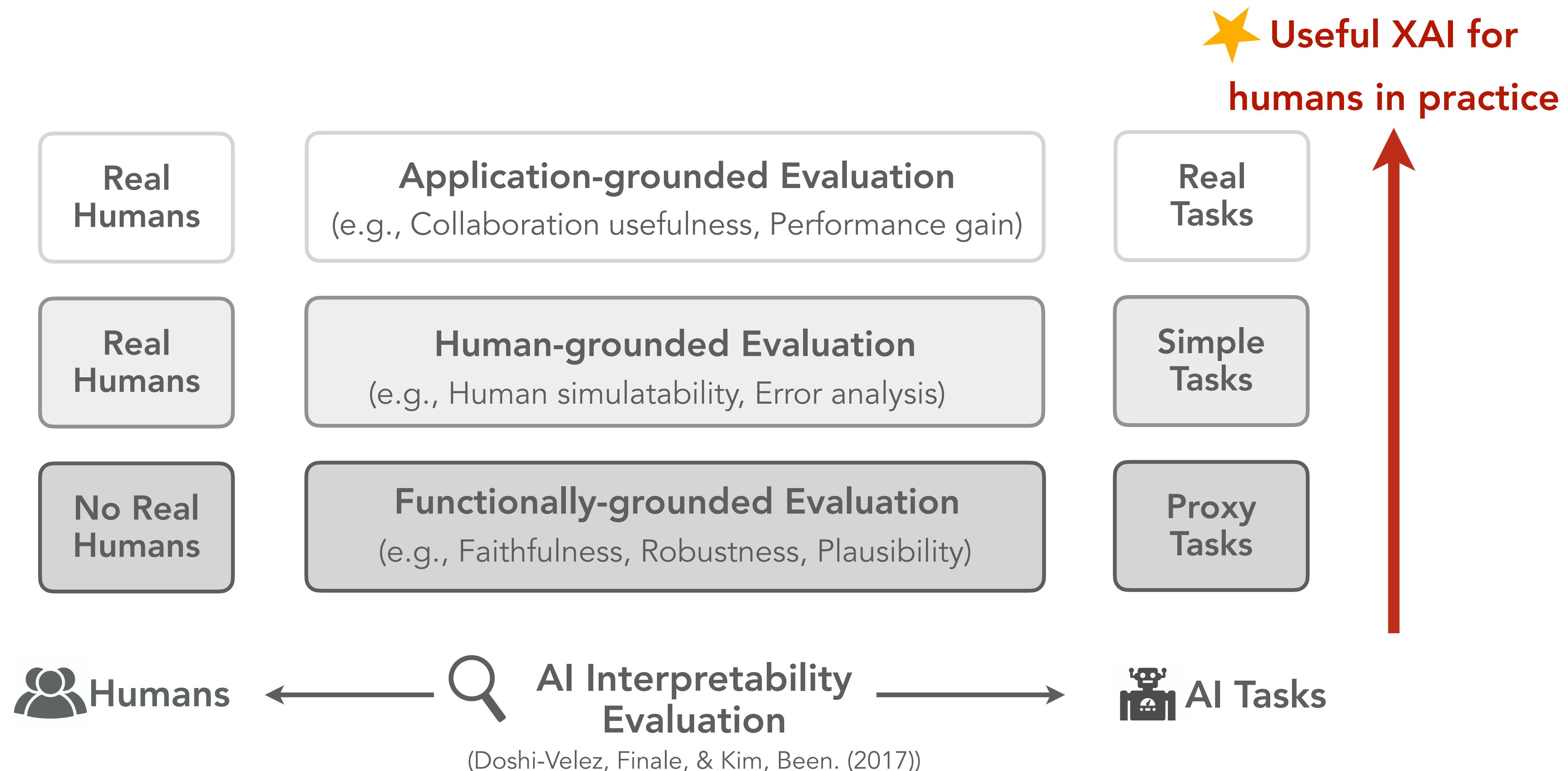


# The usefulness of XAI for humans is crucial

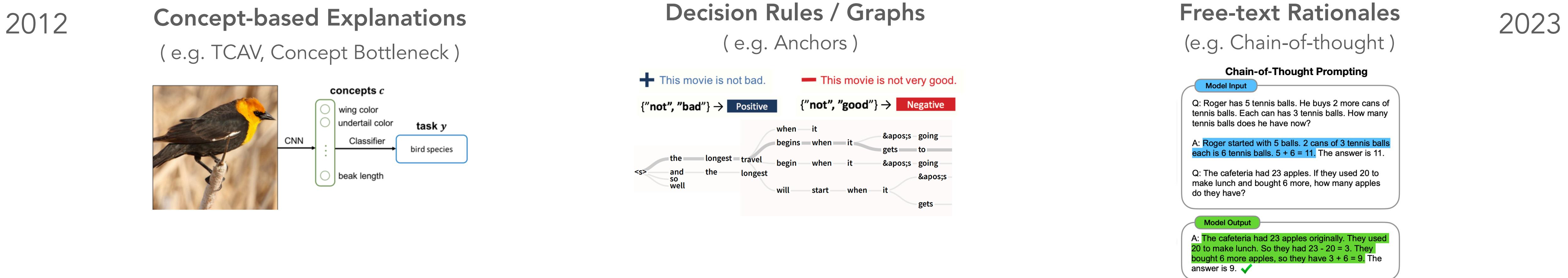
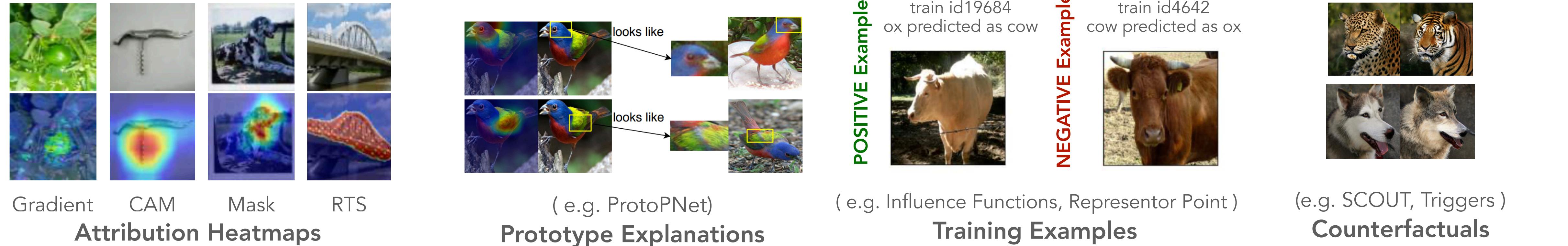
"AI interpretability is a **fallback** to be **used by humans** to **gauge the AI model reasoning** and **assess the societal measurements**"



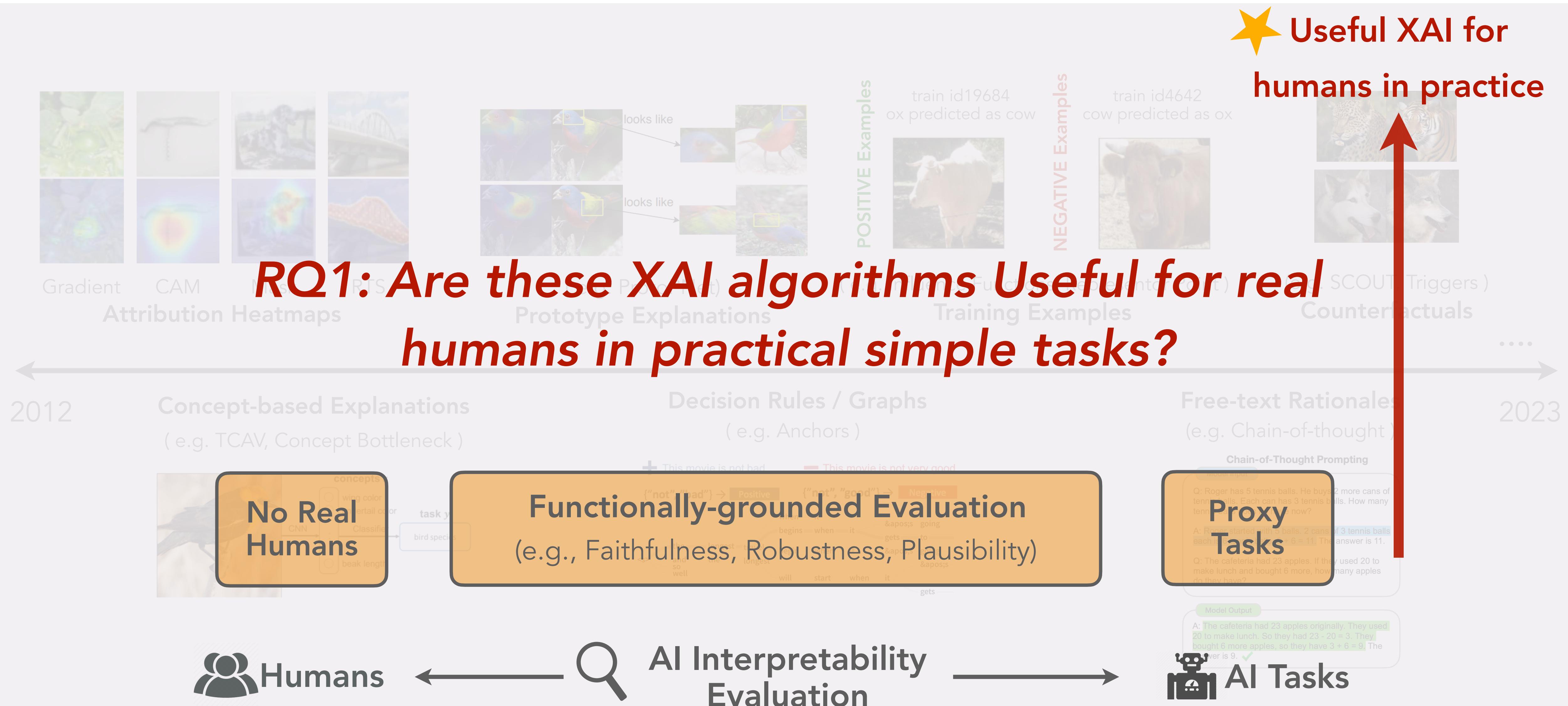
# Evaluation of XAI usefulness



# Trajectory of XAI algorithms



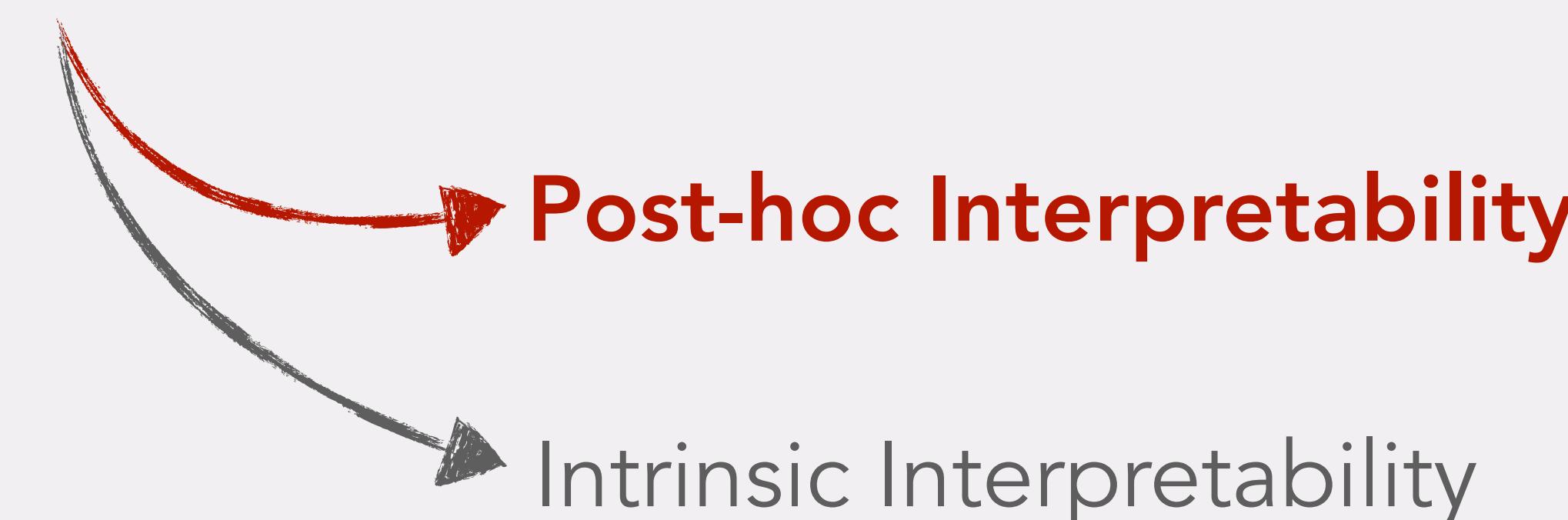
# Under-Explored: human evaluation of XAI usefulness



## Background & Motivation



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



# 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:

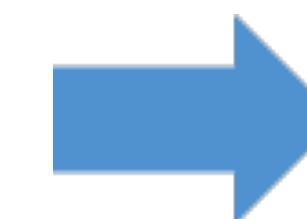
1



2

Correct Label:  
**Movie Theater**

**Input Image**



4

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

- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web



**Multiple Choice Question**



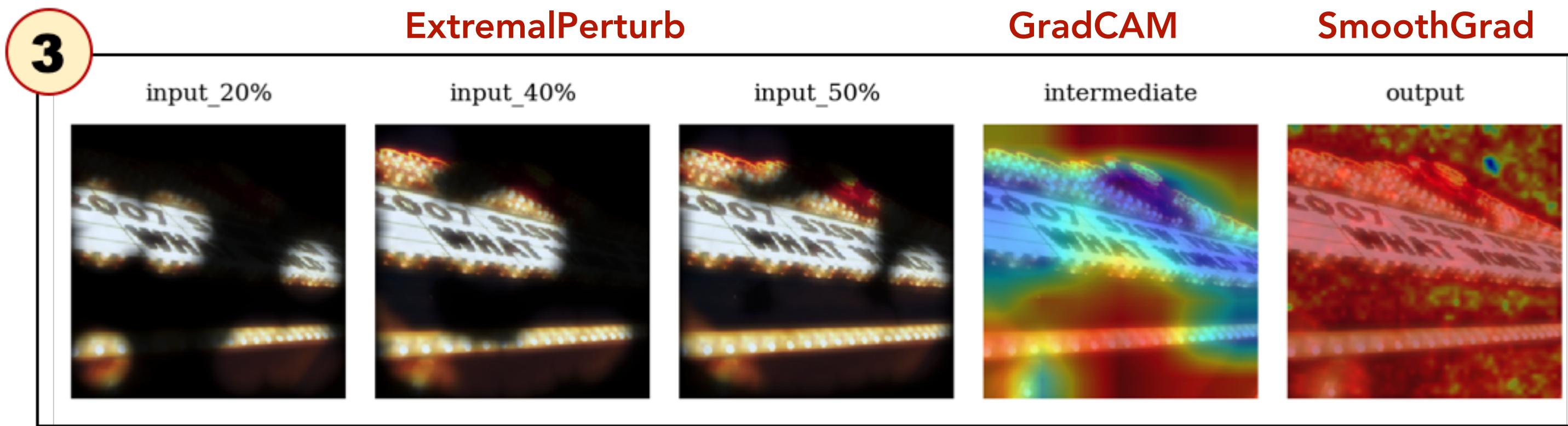
# What AI explanations are used?

The model **misidentified** this image:



**2** Correct Label:  
**Movie Theater**

**Input Image**



**Machine-Generated Interpretations (Int)**

**Widely-adopted Saliency Maps  
as AI Explanations**

4

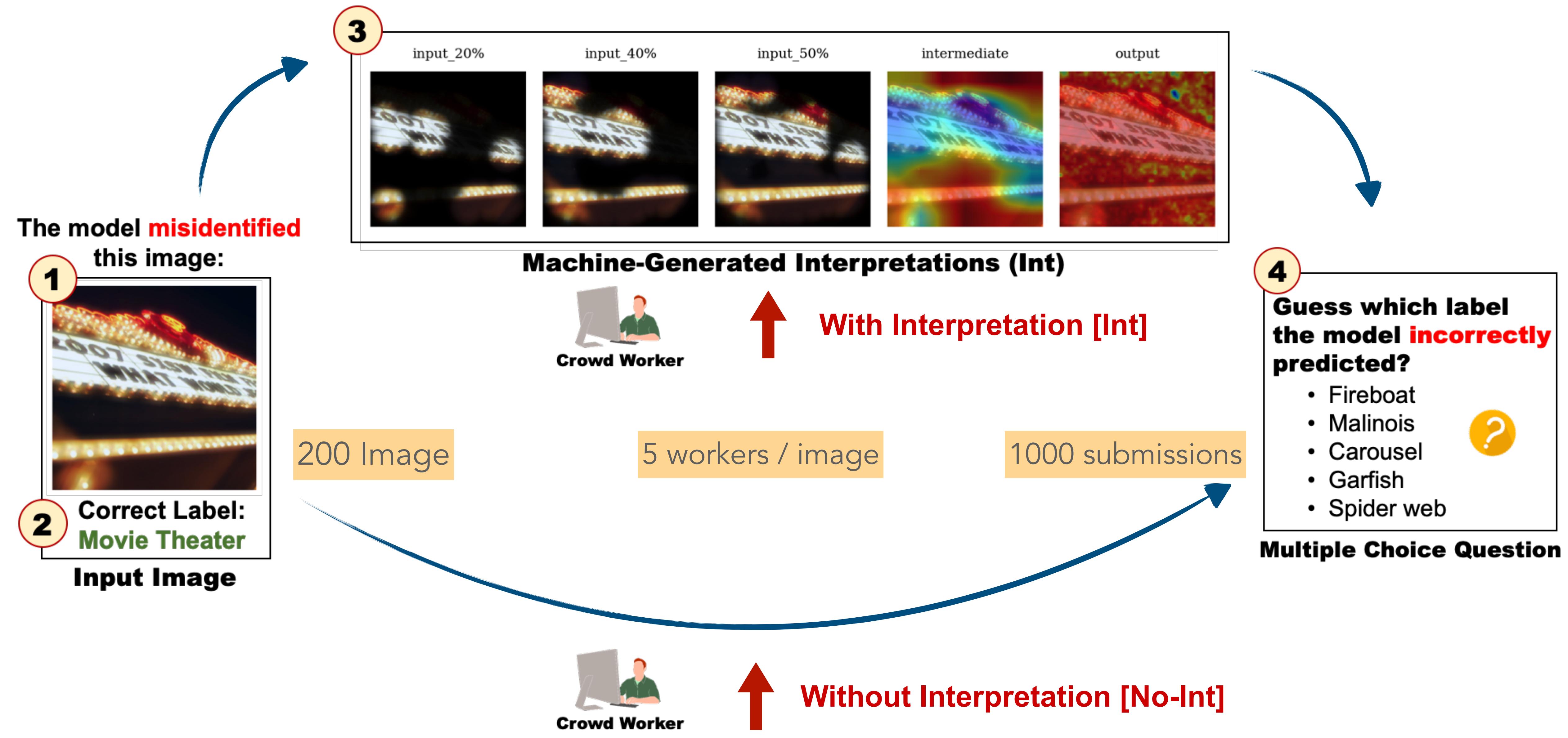
**Guess which label  
the model **incorrectly**  
predicted?**

- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web



**Multiple Choice Question**

# Design of Human Study



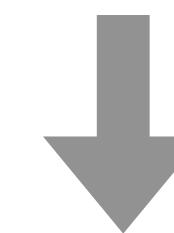
# Results

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

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

## Conclusion

[No-interpretation] condition > [Interpretation] condition



(statistically significant)

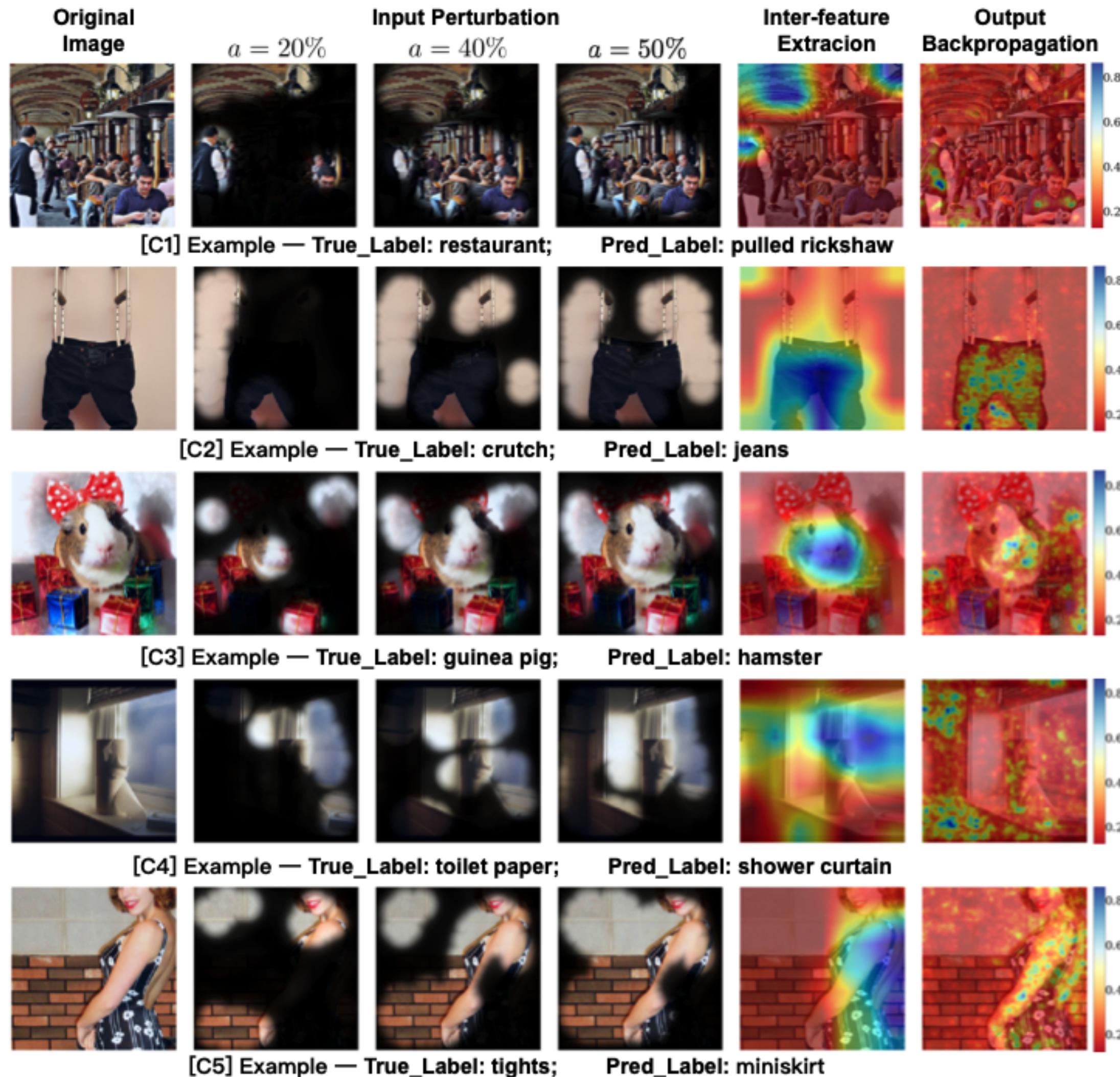
10% Accuracy Drop

# 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%.

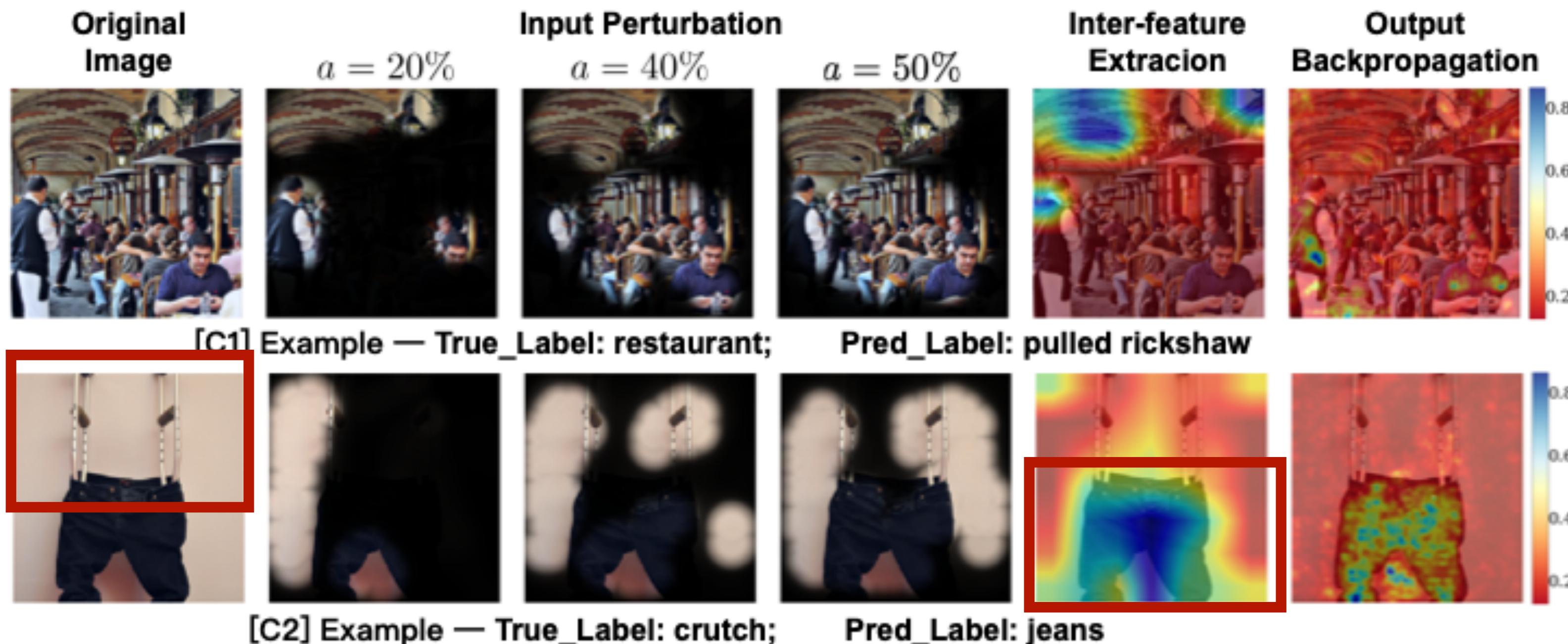
WHY?

# 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

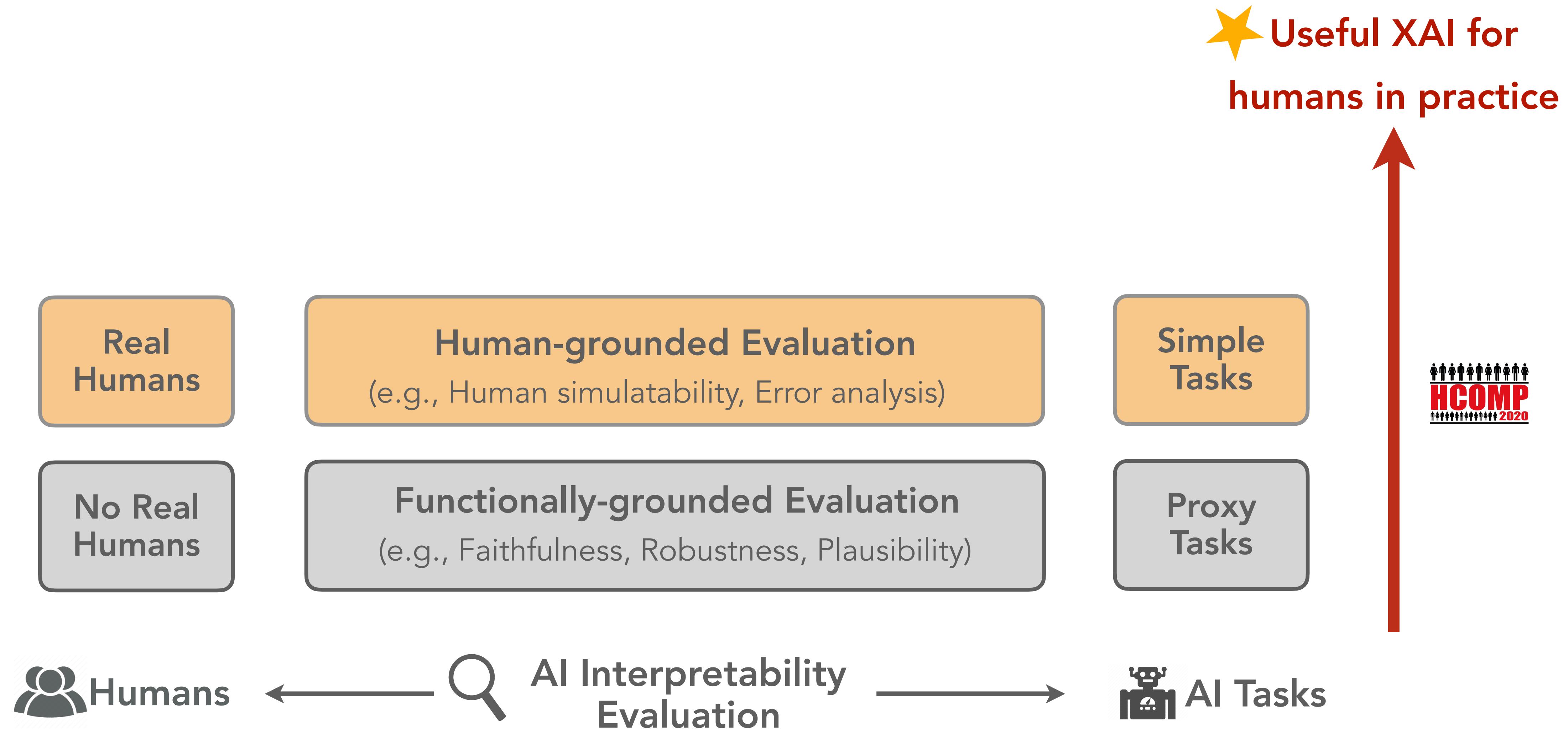
Table: Average Human on Inferring Model Misclassification (non-overlap users).

- **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



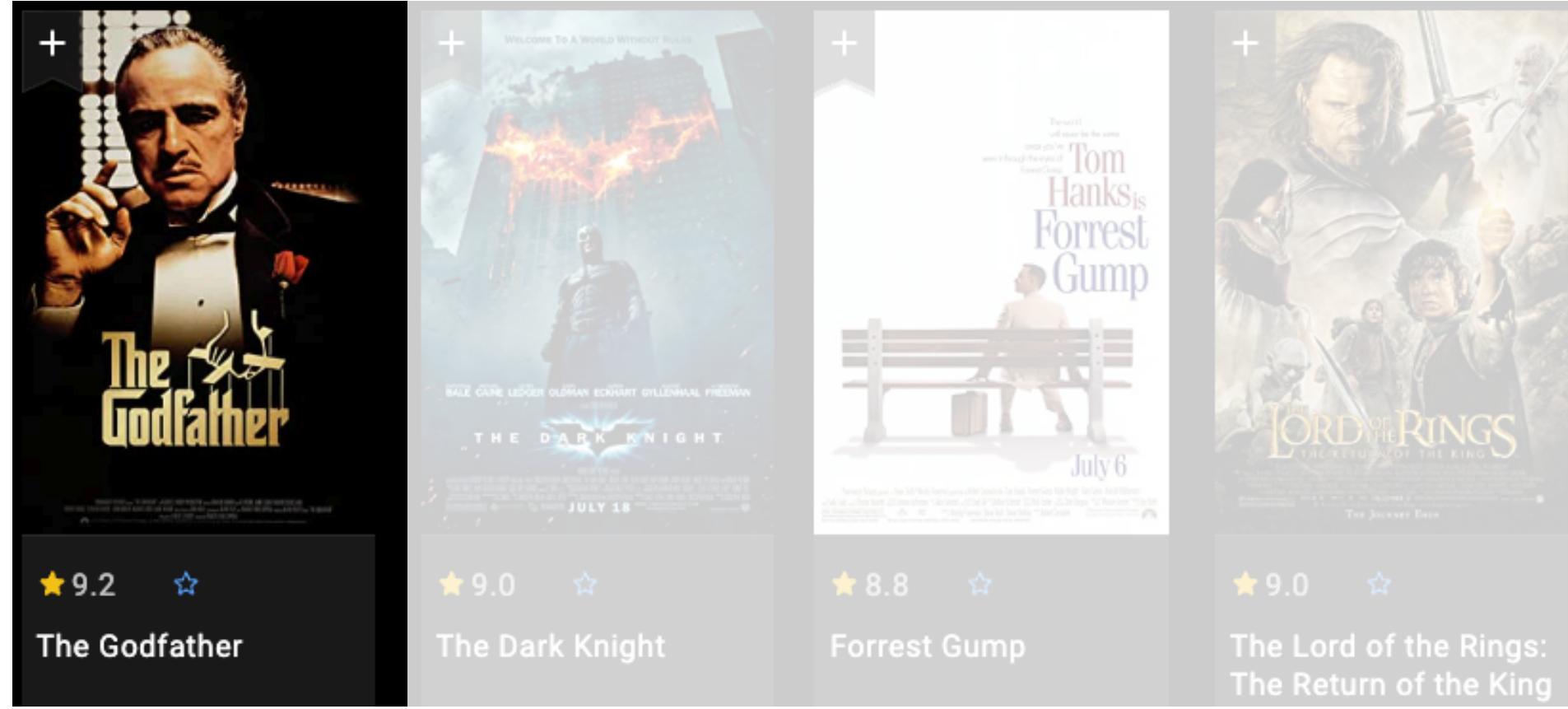
Wenbo Guo



Kenneth Huang

# Rationales in Text Classification Tasks

## Sentiment Analysis For Movie Reviews



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

Explanations

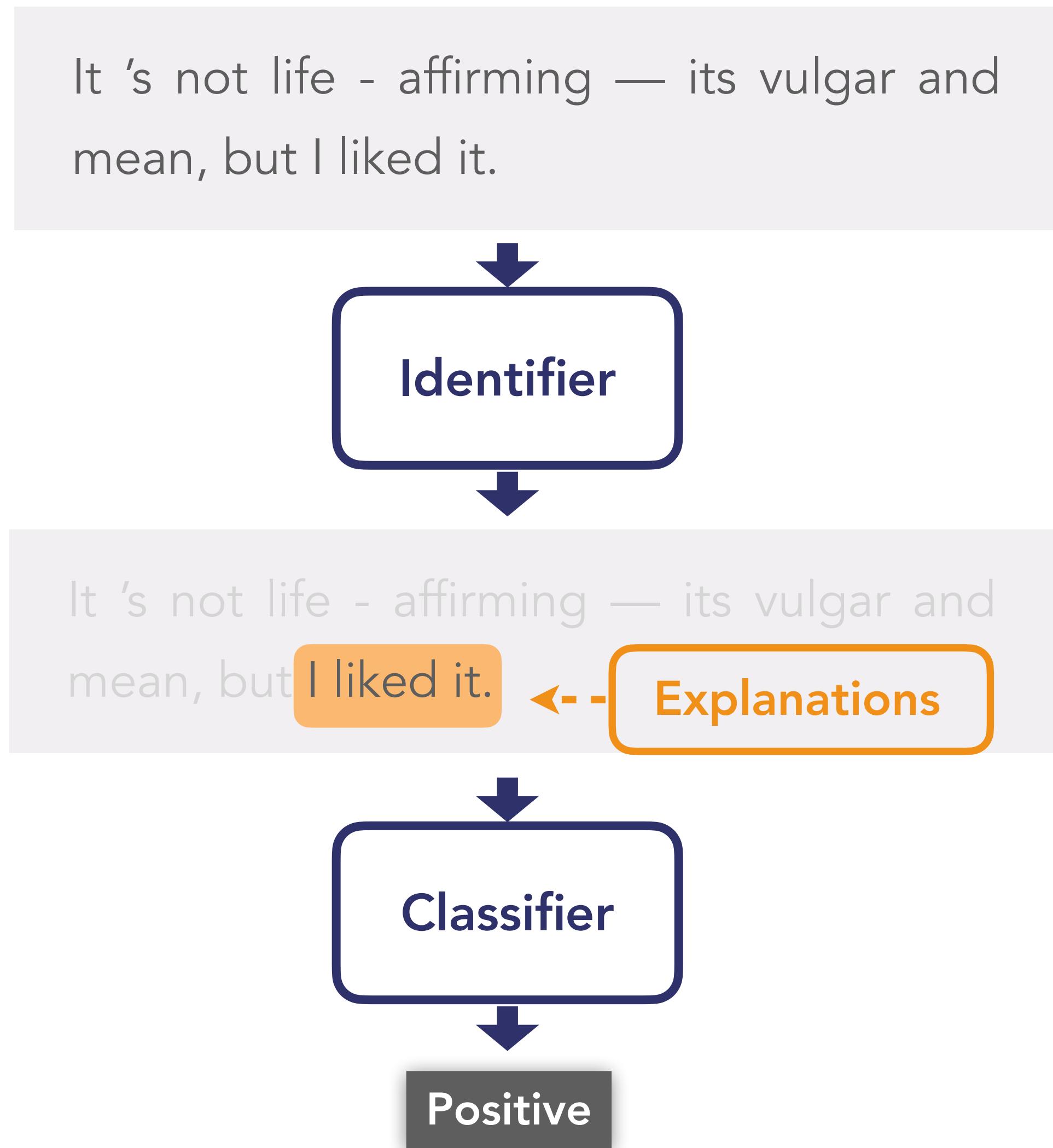
Positive

Negative

### Explanations:

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

# Self-Explaining Models



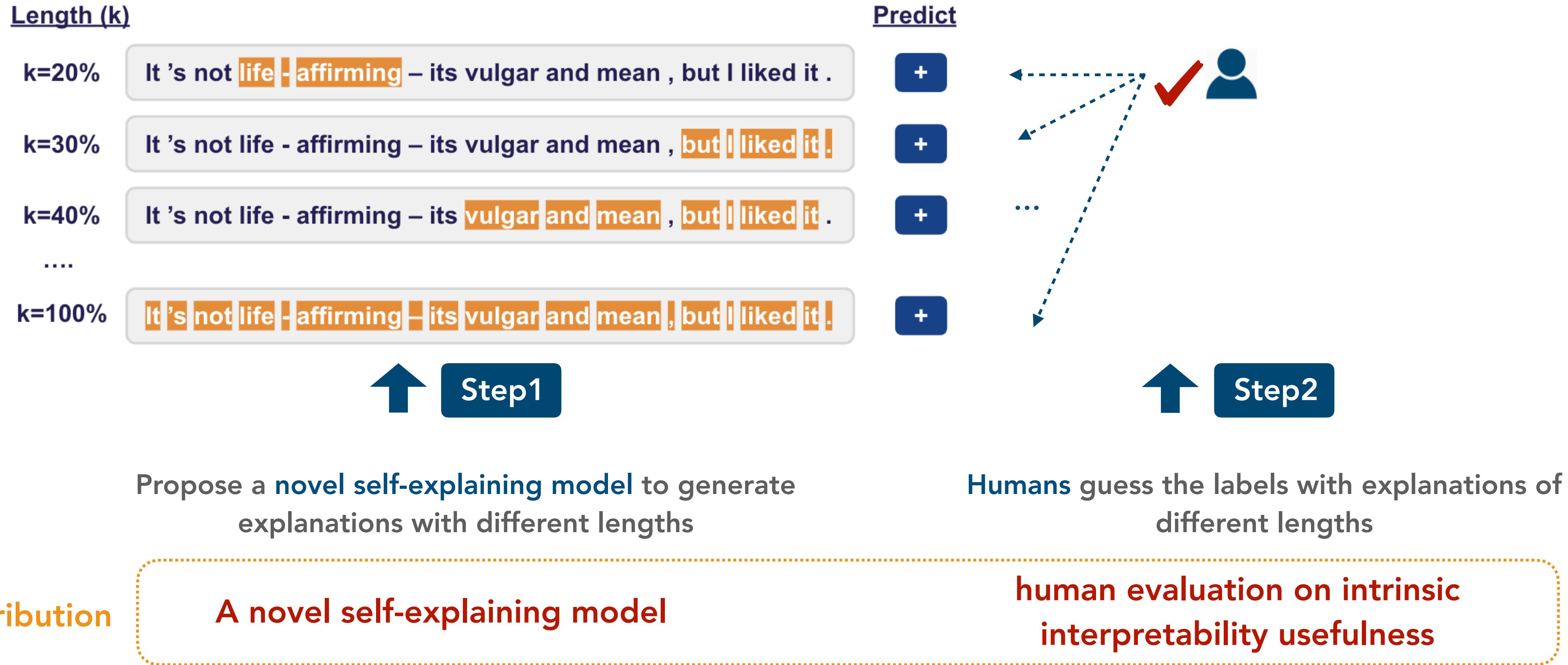
Prior Study  
**Shorter Explanations are Better.**

? Yet to be **validated by human studies!**

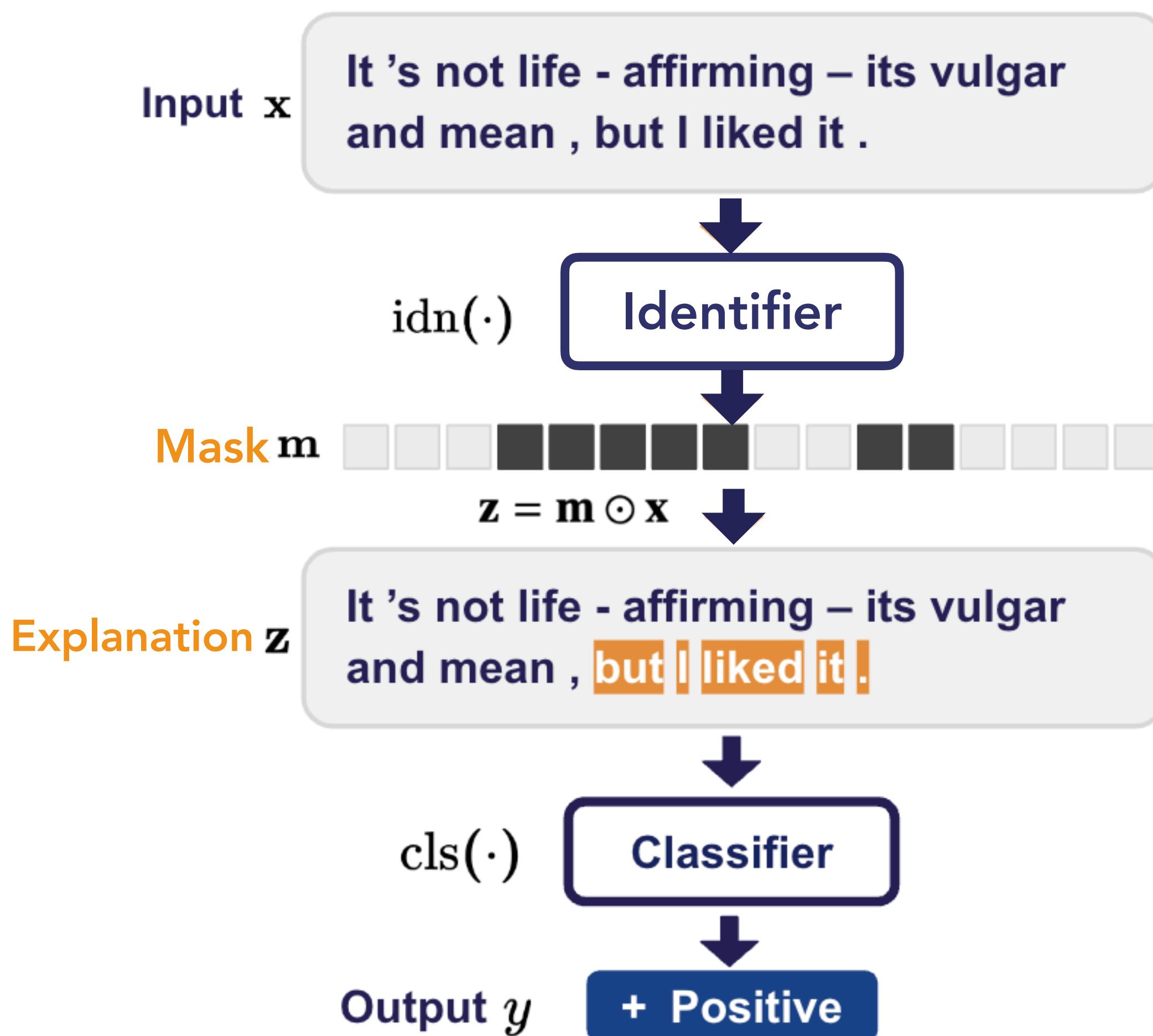
# **Are *Shortest Explanations* the *Best for Human Understanding?***

# Overview of Study Design

**Goal:** the impact of explanation lengths on human understanding:



# LimitedInk model generates rationals with different length



## Optimization Objective

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

1. Gumbel-Softmax Sampling

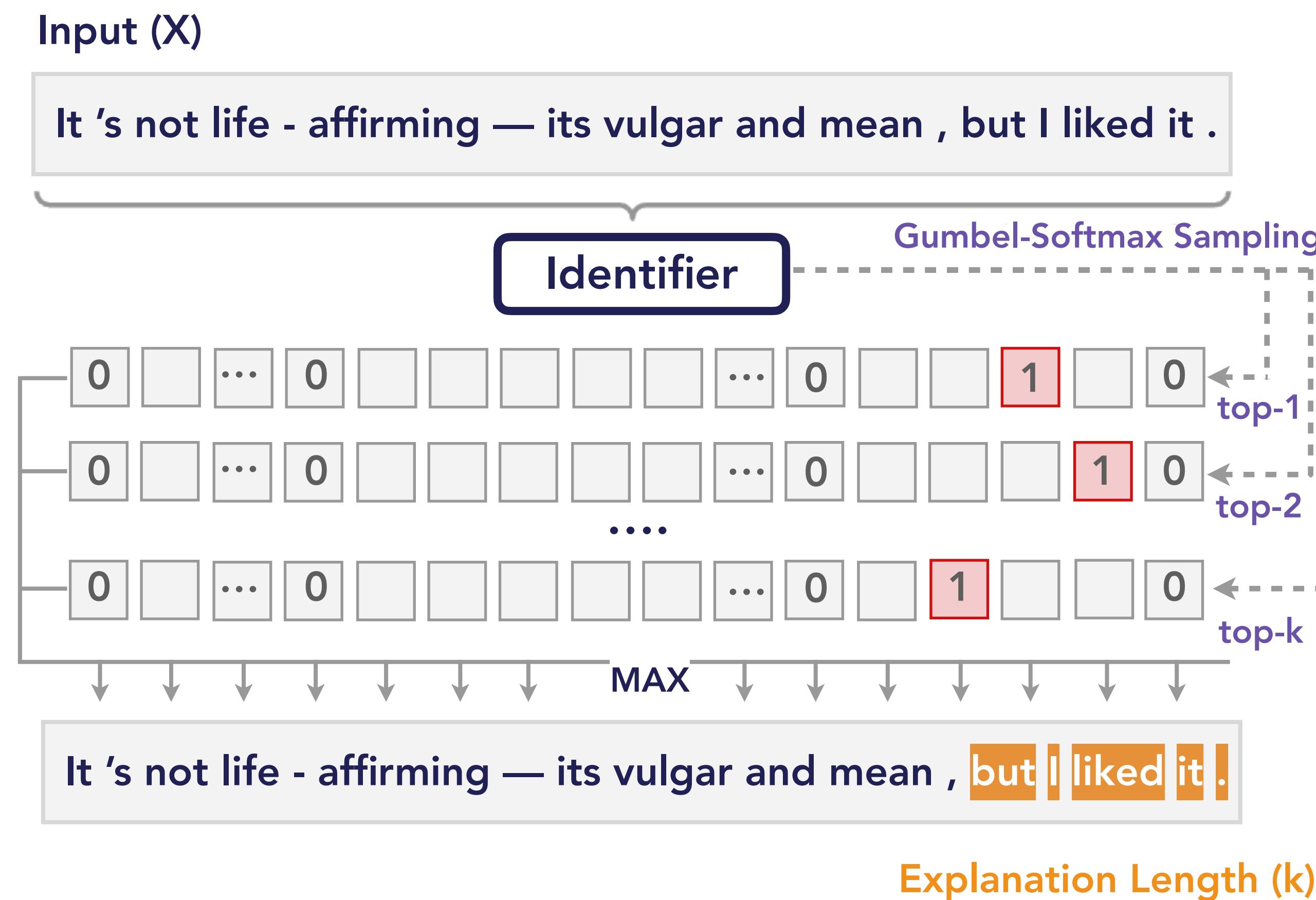
2. Vector and Sort Regularization



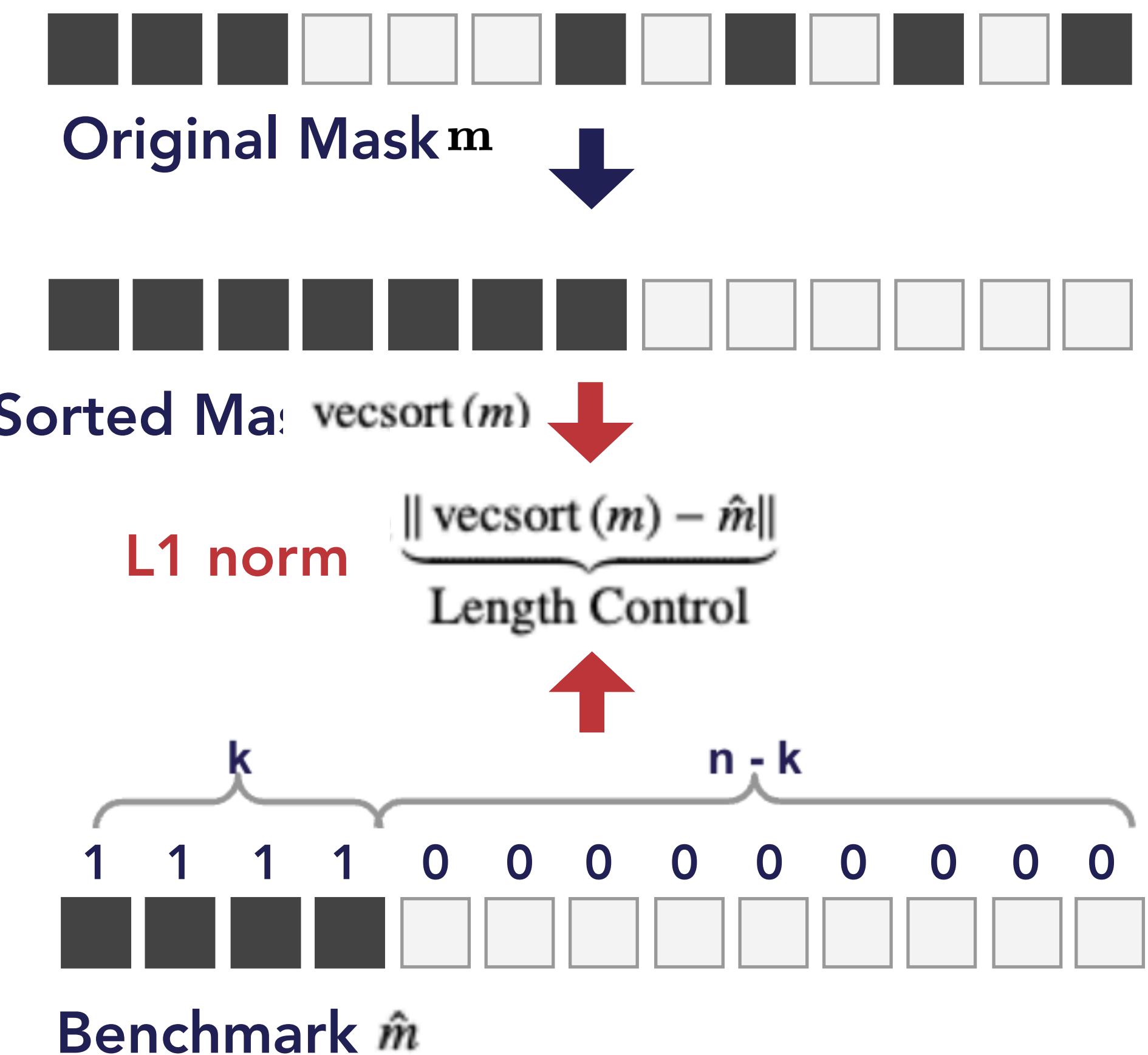
Control Different  
Explanation Length

# How to control explanation length in LimitedInk

# 1. Gumbel-Softmax Sampling



## 2. Vector and Sort Regularization



# Can LimitedInk perform well on classification?

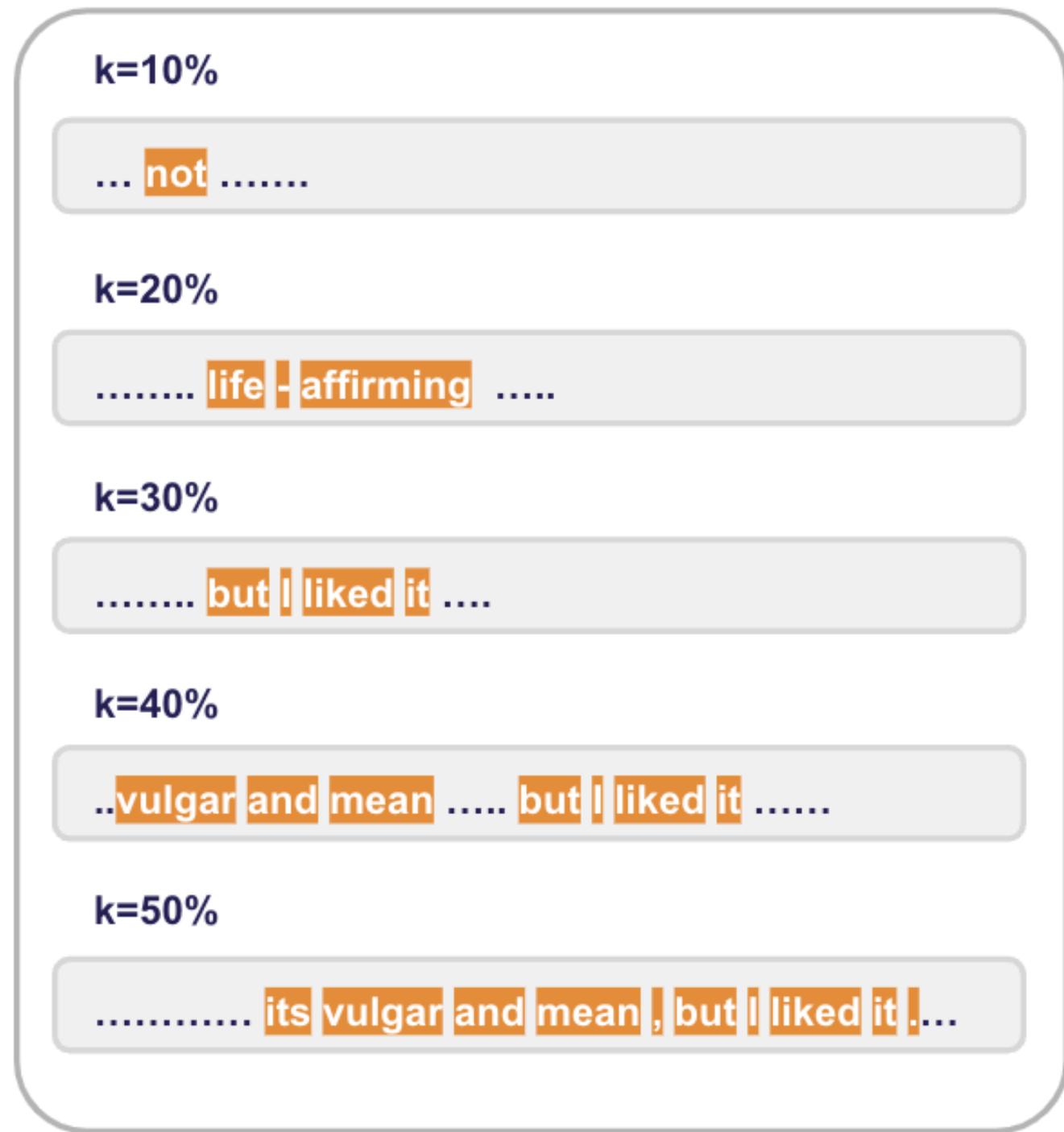
- End-task classification: Task, weighted average F1
- Human Plausibility with annotated dataset: Precision, Recall, Token-level F1

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	<b>.41</b>	.22	.83	.35	.52	.42
Sparse-IB	.84	.21	.42	.28	.46	<b>.17</b>	.15	.15	.43	.04	.21	.07	.62	.20	.33	.25	.85	<b>.37</b>	.50	<b>.43</b>
LIMITEDINK	<b>.90</b>	<b>.26</b>	<b>.50</b>	<b>.34</b>	<b>.56</b>	.13	<b>.17</b>	<b>.15</b>	<b>.50</b>	<b>.04</b>	<b>.27</b>	<b>.07</b>	<b>.67</b>	<b>.22</b>	.40	<b>.28</b>	<b>.90</b>	.28	<b>.67</b>	.39
Length Level		50%				30%				50%				50%				40%		

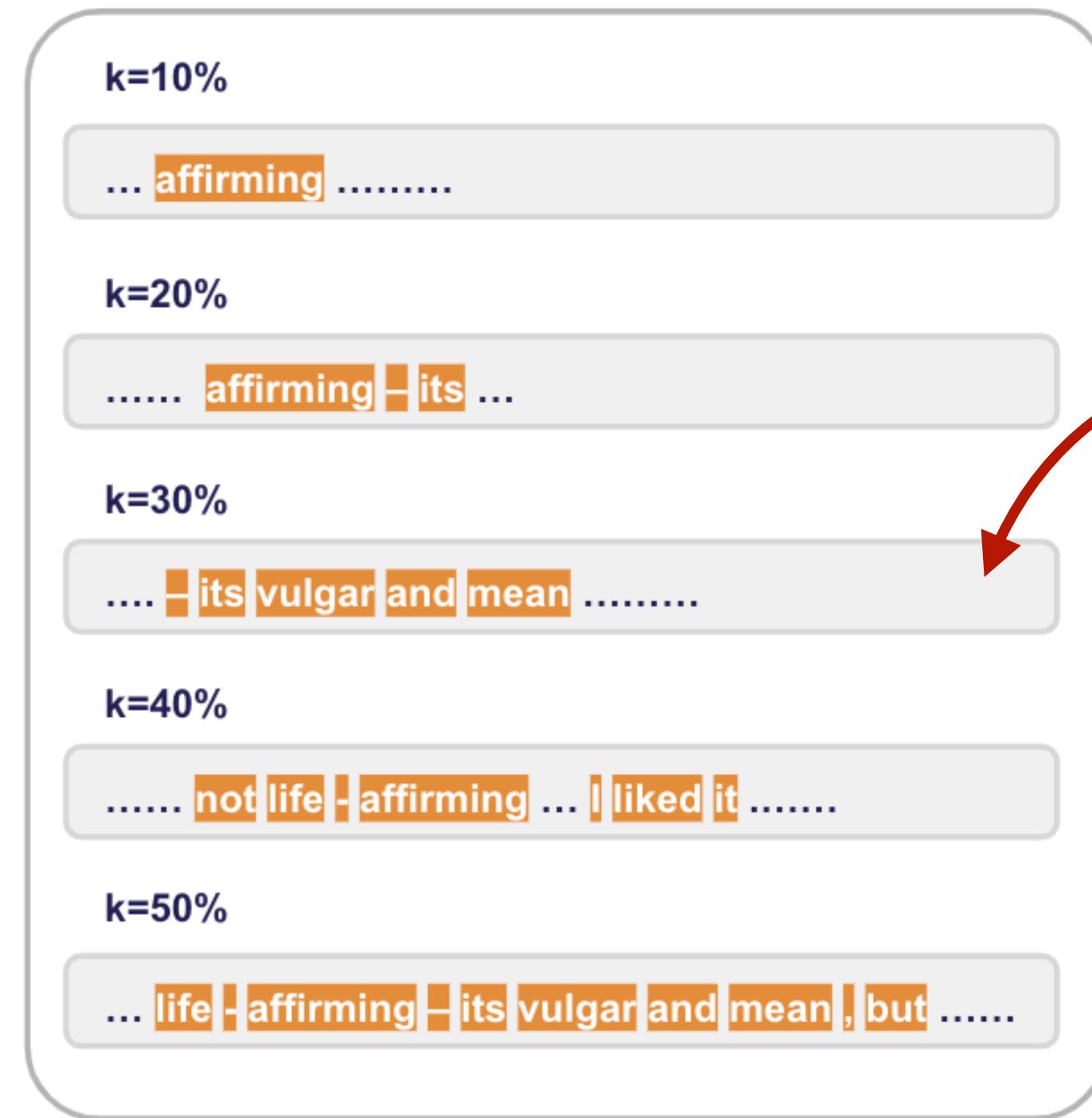
LimitedInk **performed compatible with three SOTA baselines** on the two common rationale metrics in five ERASER text classification benchmark datasets.

# Step2 - Human Study Setups

## LimitedInk Explanations



## Random text spans (similar length)



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

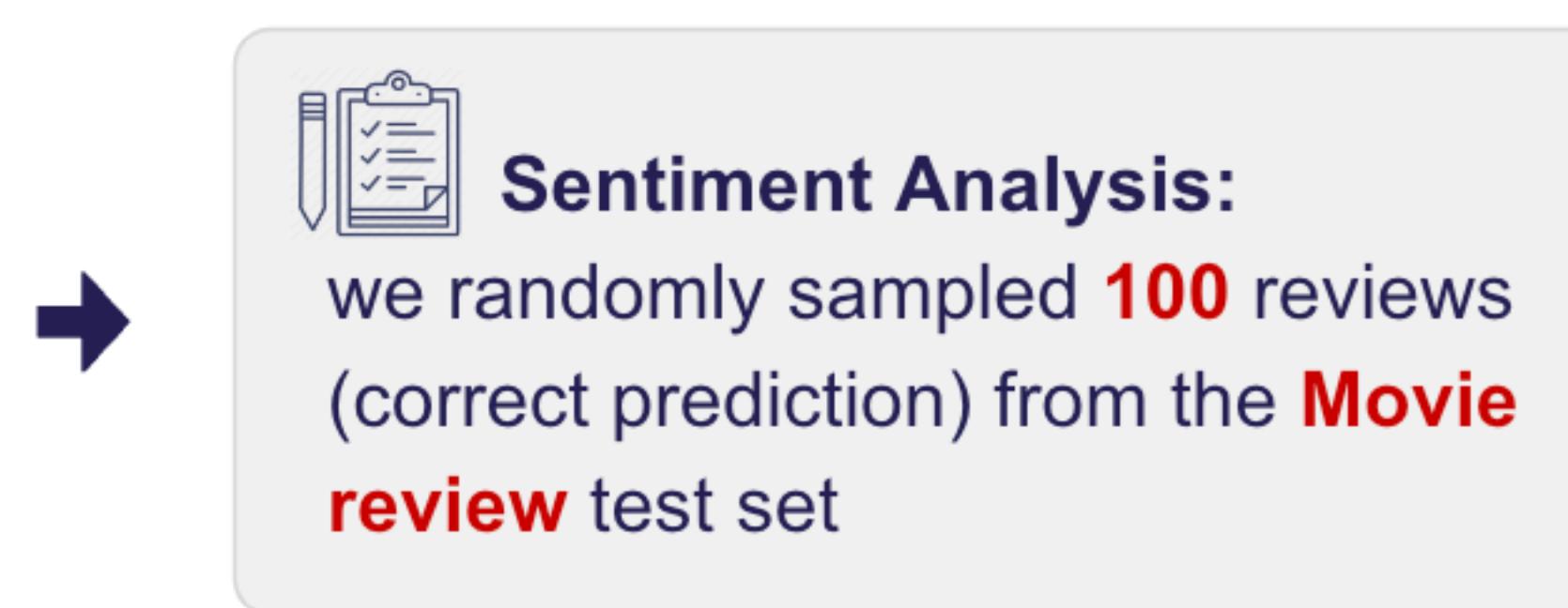
Question1: Is the movie review Positive or Negative? Please guess based on the parts of texts you see.

It's an Empty Input (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.

Submit



→ **prediction**

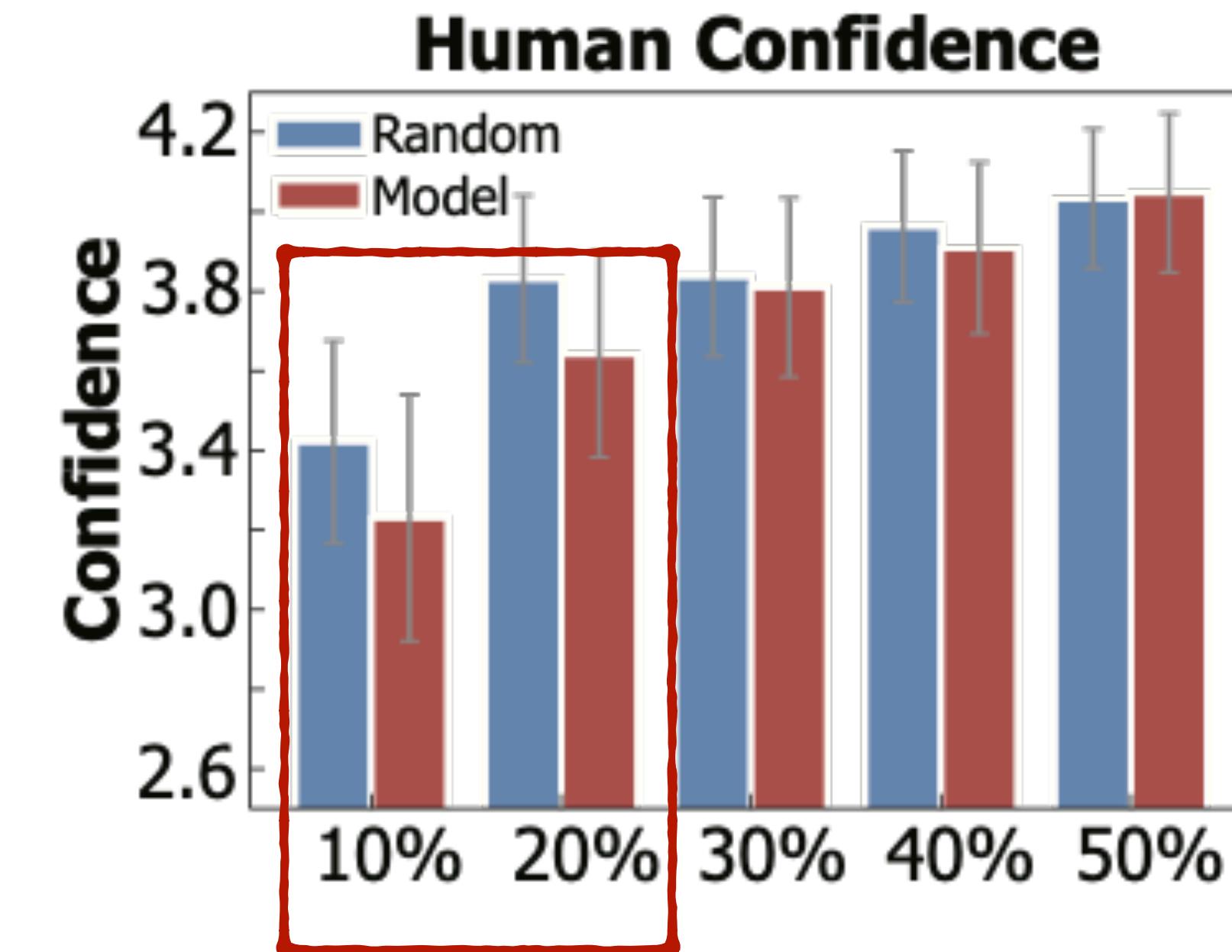
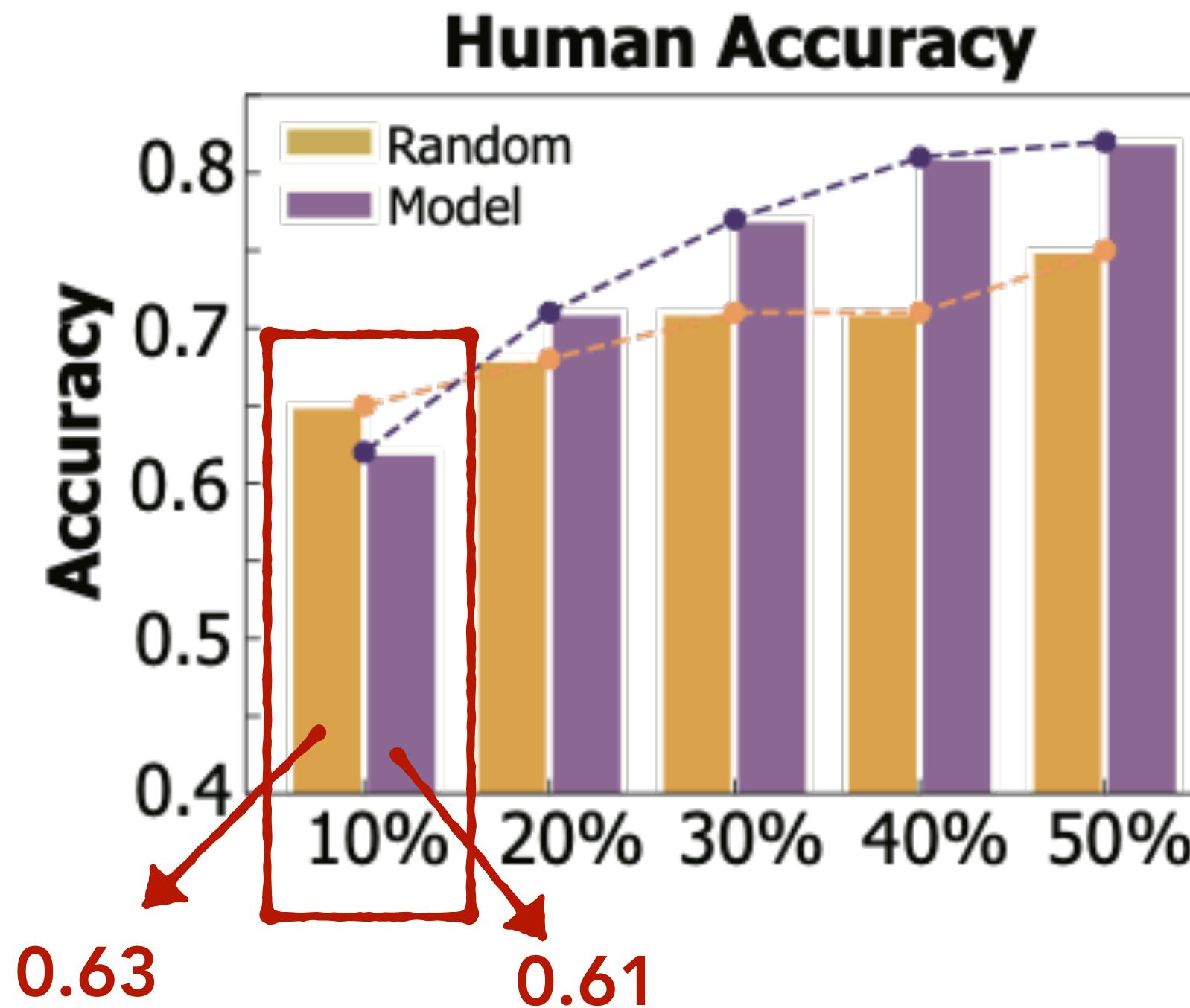


**MTurk  
Workers**

→ **confidence**

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.

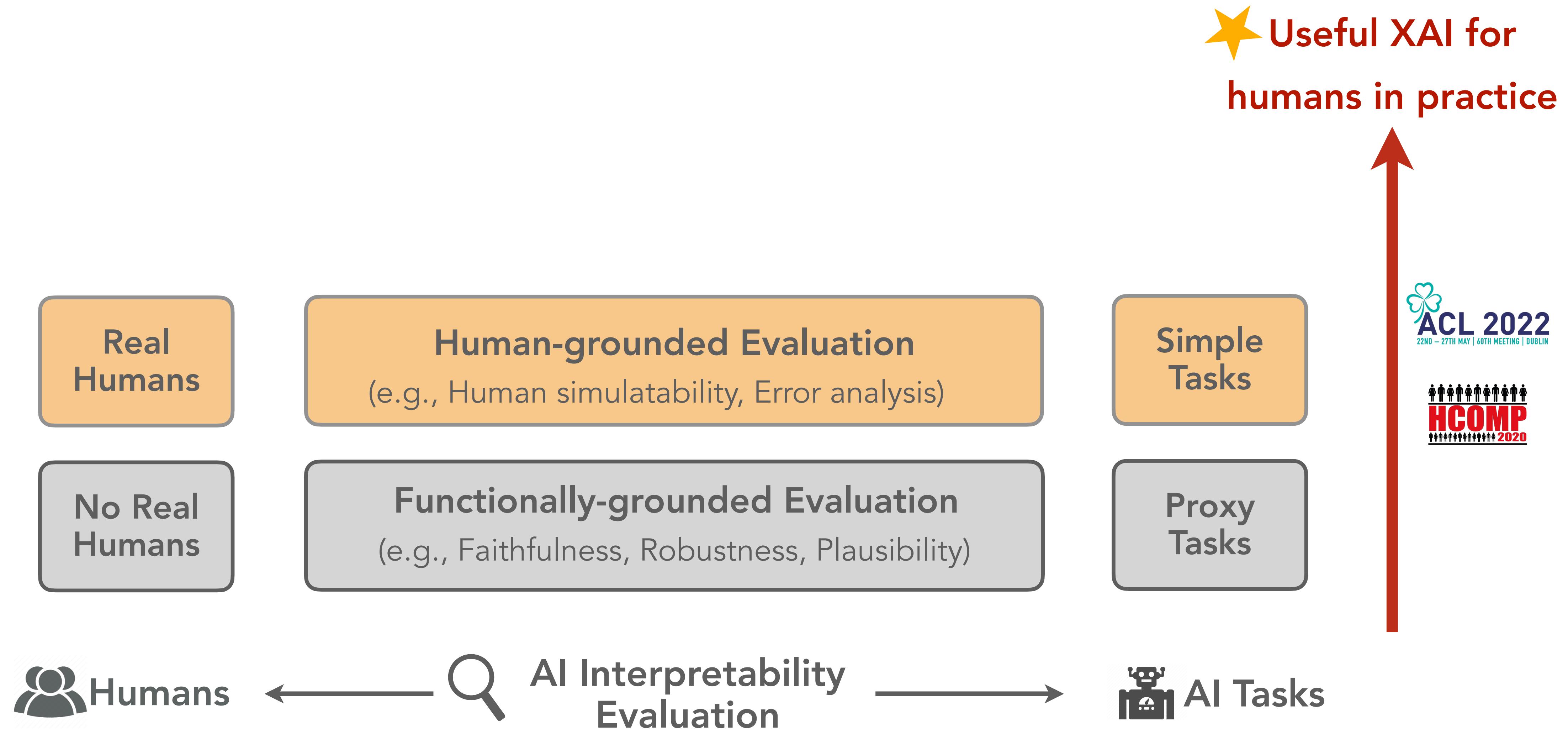
The **shortest rationales** are **NOT always the best** for human understanding.

# Take-away Message

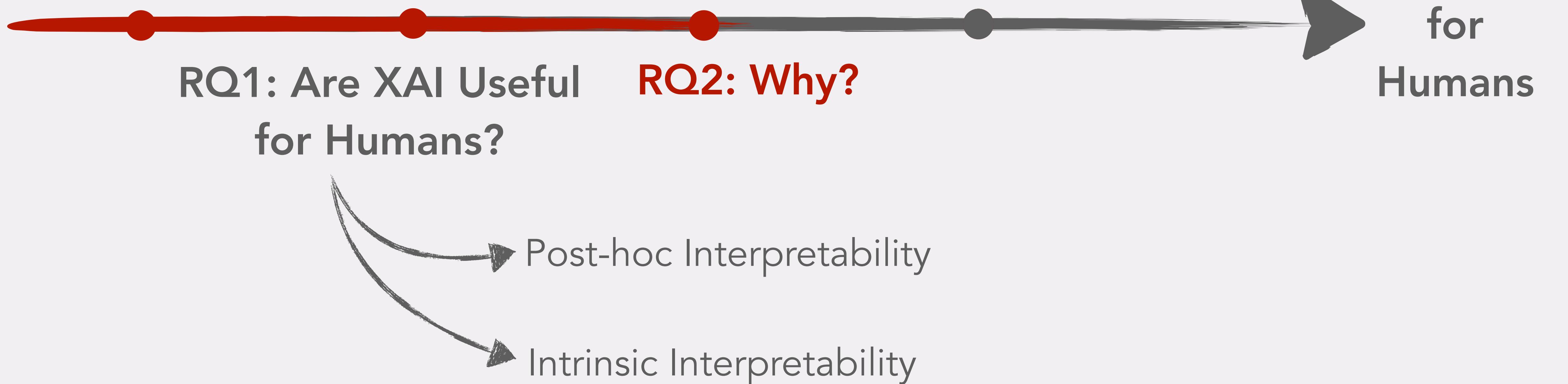
**Shortest explanations** are largely **NOT the best** for humans to simulate model predictions.

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

# Evaluation of XAI usefulness



## Background & Motivation



# Explaining the Road Not Taken



**Hua Shen**

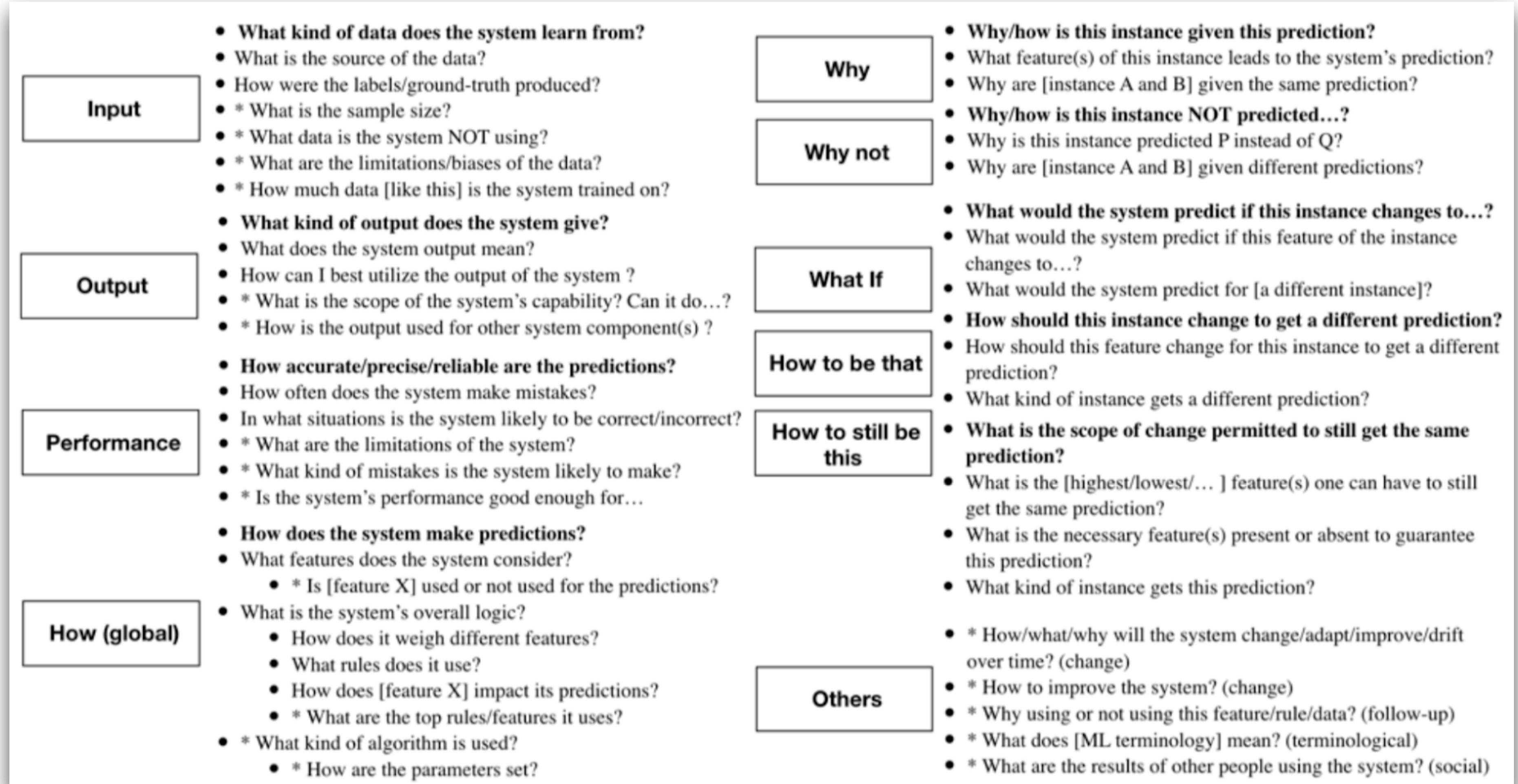


**Ting-Hao 'Kenneth' Huang**

# What are the real-world user needs for XAI?

## XAI Question Bank

43 XAI User Questions



**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	ID	Title	Year	Venue	Paper URL	ID	Title	Year	Venue	Paper URL
1	"Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>	74	FIND: Fix-in-the-Loop Debugging Deep Text Classifiers	2020	EMNLP	<a href="https://www.aclweb.org/anthology/2020.emnlp-1.147.pdf">https://www.aclweb.org/anthology/2020.emnlp-1.147.pdf</a>	147	Multimodal language analysis in the wild: Cmu-mosei dataset and Interpretation	2018	ACL	<a href="https://www.aclweb.org/anthology/P18-1208.pdf">https://www.aclweb.org/anthology/P18-1208.pdf</a>
2	A causal framework for explaining the predictions of black-box sequence-to-set	2017	EMNLP	<a href="https://arxiv.org/pdf/1707.01943.pdf">https://arxiv.org/pdf/1707.01943.pdf</a>	75	Fine-grained analysis of sentence embeddings using auxiliary prediction task	2017	ICLR	<a href="https://arxiv.org/pdf/1608.04207.pdf">https://arxiv.org/pdf/1608.04207.pdf</a>	148	Multimodal Routing: Improving Local and Global Interpretability of Multimodal	2020	EMNLP	<a href="https://arxiv.org/pdf/2004.14198.pdf">https://arxiv.org/pdf/2004.14198.pdf</a>
3	A Diagnostic Study of Explainability Techniques for Text Classification	2020	EMNLP	<a href="https://arxiv.org/pdf/1909.13295.pdf">https://arxiv.org/pdf/1909.13295.pdf</a>	76	Generating Fact Checking Explanations	2020	ACL	<a href="https://arxiv.org/pdf/2004.05773.pdf">https://arxiv.org/pdf/2004.05773.pdf</a>	149	Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels	2020	EMNLP	<a href="https://arxiv.org/pdf/2010.07526.pdf">https://arxiv.org/pdf/2010.07526.pdf</a>
4	A Learning-based English Math Word Problem Solver with Understanding, R	2016	COLING	<a href="https://www.aclweb.org/anthology/C16-2032.pdf">https://www.aclweb.org/anthology/C16-2032.pdf</a>	77	Generating question-relevant captions for visual question answering	2019	ACL	<a href="https://arxiv.org/pdf/1908.00513.pdf">https://arxiv.org/pdf/1908.00513.pdf</a>	150	Neural vector conceptualization for word vector space interpretation	2019	NAACL	<a href="https://www.aclweb.org/anthology/N19-2003.pdf">https://www.aclweb.org/anthology/N19-2003.pdf</a>
5	A paper in Interlogy: What we know about how text works	2020	TACL	<a href="https://arxiv.org/pdf/2004.12327.pdf">https://arxiv.org/pdf/2004.12327.pdf</a>	78	Generating Token-Level Explanations for Natural Language Inference	2019	ACL	<a href="https://arxiv.org/pdf/1904.10717.pdf">https://arxiv.org/pdf/1904.10717.pdf</a>	151	No Explainability without Accountability: An Empirical Study of Explanations	2020	CHI	<a href="https://homes.cs.washington.edu/~weidipapers/">https://homes.cs.washington.edu/~weidipapers/</a>
6	A Shared Attention Mechanism for Interpretation of Neural Automatic Post-E	2018	ACL	<a href="https://www.aclweb.org/anthology/W18-2702.pdf">https://www.aclweb.org/anthology/W18-2702.pdf</a>	79	GEval: Tool for Debugging NLP Datasets and Models	2019	BlackboxNLP	<a href="https://www.aclweb.org/anthology/W19-4826.pdf">https://www.aclweb.org/anthology/W19-4826.pdf</a>	152	Obtaining Faithful Interpretations from Compositional Neural Networks	2020	ACL	<a href="https://www.aclweb.org/anthology/2020.acl-main.pdf">https://www.aclweb.org/anthology/2020.acl-main.pdf</a>
7	A structured probe for finding syntax in word representations	2019	NAACL	<a href="https://www.aclweb.org/anthology/N19-1419.pdf">https://www.aclweb.org/anthology/N19-1419.pdf</a>	80	Global model interpretation via recursive partitioning	2018	DSS	<a href="https://arxiv.org/pdf/1802.04253.pdf">https://arxiv.org/pdf/1802.04253.pdf</a>	153	Open Sesame: Getting Inside BERT's Linguistic Knowledge	2019	BlackboxNLP	<a href="https://www.aclweb.org/anthology/W19-4825.pdf">https://www.aclweb.org/anthology/W19-4825.pdf</a>
8	A Survey of the State of Explainable AI for Natural Language Processing	2020	AAACL-JCNLP	<a href="https://arxiv.org/pdf/2010.00711.pdf">https://arxiv.org/pdf/2010.00711.pdf</a>	81	GLUoME: Generalized and Contextualized Story Explanations	2019	EMNLP	<a href="https://arxiv.org/pdf/2009.07758.pdf">https://arxiv.org/pdf/2009.07758.pdf</a>	154	OpenDialKG: Explainable Conversational Reasoning with Attention-based W	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1081.pdf">https://www.aclweb.org/anthology/P19-1081.pdf</a>
9	AllenNLP interpret: A framework for explaining predictions of nlp models	2019	EMNLP	<a href="https://www.aclweb.org/anthology/D19-3002.pdf">https://www.aclweb.org/anthology/D19-3002.pdf</a>	82	Guiding the Training of Semantics: Interpretable Video Captioning via POS Tag	2019	EMNLP	<a href="https://arxiv.org/pdf/1705.07405.pdf">https://arxiv.org/pdf/1705.07405.pdf</a>	155	Pathologies of Neural Models Make Interpretations Difficult	2018	EMNLP	<a href="https://arxiv.org/pdf/1804.07781.pdf">https://arxiv.org/pdf/1804.07781.pdf</a>
10	An Information Bottleneck Approach for Controlling Conciseness in Rationales	2020	EMNLP	<a href="https://arxiv.org/pdf/2005.00652.pdf">https://arxiv.org/pdf/2005.00652.pdf</a>	83	HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-3023.pdf">https://www.aclweb.org/anthology/P19-3023.pdf</a>	156	Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BI	2020	ACL	<a href="https://arxiv.org/pdf/2004.14786.pdf">https://arxiv.org/pdf/2004.14786.pdf</a>
11	An Interpretive Knowledge Transfer Model for Knowledge Base Completion	2017	ACL	<a href="https://www.aclweb.org/anthology/P17-1068.pdf">https://www.aclweb.org/anthology/P17-1068.pdf</a>	84	HolotQA: A Dataset for Divisible, Explainable Multi-hop Question Answering	2018	EMNLP	<a href="https://www.aclweb.org/anthology/W18-1259.pdf">https://www.aclweb.org/anthology/W18-1259.pdf</a>	157	Predicting and Interpreting Embeddings for out-of-vocabulary words in down towns	2018	BlackboxNLP	<a href="https://www.aclweb.org/anthology/W18-5439.pdf">https://www.aclweb.org/anthology/W18-5439.pdf</a>
12	An Interpretable Reasoning Network for Multi-Relation Question Answering	2018	COLING	<a href="https://www.aclweb.org/anthology/C18-1171.pdf">https://www.aclweb.org/anthology/C18-1171.pdf</a>	85	How contextual are contextualized word representations? Comparing the gap	2019	EMNLP-JCNLP	<a href="https://arxiv.org/pdf/1809.00512.pdf">https://arxiv.org/pdf/1809.00512.pdf</a>	158	Principles of Explanatory Debugging to Personalize Interactive Machine Lear	2015	IJCAI	<a href="https://openaccess.city.ac.uk/381970x.pdf">https://openaccess.city.ac.uk/381970x.pdf</a>
13	Analyzing the potential of seq-to-seq models for incremental interpretation	2018	BlackboxNLP	<a href="https://www.aclweb.org/anthology/W18-5419.pdf">https://www.aclweb.org/anthology/W18-5419.pdf</a>	86	How do Decisions Emerge across Layers of Neural Models? Interpretation w	2020	EMNLP	<a href="https://arxiv.org/pdf/2004.12242.pdf">https://arxiv.org/pdf/2004.12242.pdf</a>	159	Probing Emergent Semantics in Predictive Agents via Question Answering	2020	ArXiv	<a href="https://arxiv.org/pdf/2006.01016.pdf">https://arxiv.org/pdf/2006.01016.pdf</a>
14	Analysis methods in neural language processing: A survey	2019	TACL	<a href="https://arxiv.org/pdf/1812.08951.pdf">https://arxiv.org/pdf/1812.08951.pdf</a>	87	How Important is a Neuron	2019	ICLR	<a href="https://arxiv.org/pdf/1805.12233.pdf">https://arxiv.org/pdf/1805.12233.pdf</a>	160	Probing for semantic evidence of composition by means of simple classificat	2016	EVSR-W	<a href="https://www.aclweb.org/anthology/W16-2524.pdf">https://www.aclweb.org/anthology/W16-2524.pdf</a>
15	Analytical methods for interpretable ultradense word embeddings	2019	EMNLP	<a href="https://www.aclweb.org/anthology/D19-1111.pdf">https://www.aclweb.org/anthology/D19-1111.pdf</a>	88	How much should you ask? On the question structure in Q&A systems	2018	BlackboxNLP	<a href="https://arxiv.org/pdf/1809.03734.pdf">https://arxiv.org/pdf/1809.03734.pdf</a>	161	Probing Neural Dialog Models for Conversational Understanding	2020	ACL-NLP4ConvAI	<a href="https://arxiv.org/pdf/2008.08331.pdf">https://arxiv.org/pdf/2008.08331.pdf</a>
16	Analyzing the Structure of Attention in a Transformer Language Model	2019	BlackboxNLP	<a href="https://www.aclweb.org/anthology/W19-4808.pdf">https://www.aclweb.org/anthology/W19-4808.pdf</a>	89	How Useful Are the Machine-Generated Interpretations to Generic Users? A	2020	HCOMP	<a href="https://arxiv.org/abs/2018.11721">https://arxiv.org/abs/2018.11721</a>	162	Rationale Induction by Rationale Generation: Learning to Solve and Explain A	2017	ACL	<a href="https://www.aclweb.org/anthology/P17-1015.pdf">https://www.aclweb.org/anthology/P17-1015.pdf</a>
17	Anchors: High-Precision Model-Agnostic Explanations	2018	AAAI	<a href="https://homes.cs.washington.edu/~marcotra/aaai/">https://homes.cs.washington.edu/~marcotra/aaai/</a>	90	Human Attention in Visual Question Answering: Do Humans and Deep Models	2016	EMNLP	<a href="https://www.aclweb.org/anthology/D16-1082.pdf">https://www.aclweb.org/anthology/D16-1082.pdf</a>	163	PROVER: Proof and Generation for Interpretable Reasoning over Rules	2020	EMNLP	<a href="https://www.aclweb.org/anthology/2020.emnlp-r.pdf">https://www.aclweb.org/anthology/2020.emnlp-r.pdf</a>
18	Are sliders heads ready to be beaten?	2019	NeuIPS	<a href="https://arxiv.org/pdf/1905.10650.pdf">https://arxiv.org/pdf/1905.10650.pdf</a>	91	Human Attention Maps for Text Classification: Do Humans and Neural Netwo	2020	ACL	<a href="https://www.aclweb.org/anthology/2020.acl-main.pdf">https://www.aclweb.org/anthology/2020.acl-main.pdf</a>	164	Quick and (not so) dirty: Unsupervised Selection of Justification Sentences	2020	EMNLP	<a href="https://www.aclweb.org/anthology/W19-1260.pdf">https://www.aclweb.org/anthology/W19-1260.pdf</a>
19	Assessing social and interactional biases in contextualized word representa	2019	NeuIPS	<a href="https://arxiv.org/pdf/1911.01485.pdf">https://arxiv.org/pdf/1911.01485.pdf</a>	92	Human-grounded Evaluations of Explanation Methods for Text Classification	2019	EMNLP-JCNLP	<a href="https://www.aclweb.org/anthology/D19-1523.pdf">https://www.aclweb.org/anthology/D19-1523.pdf</a>	165	Quint: Interpretable question answering over knowledge bases.	2017	EMNLP	<a href="https://www.aclweb.org/anthology/D17-2011.pdf">https://www.aclweb.org/anthology/D17-2011.pdf</a>
20	Attention interpretability across nlp tasks	2019	ArXiv	<a href="https://arxiv.org/pdf/1909.11216.pdf">https://arxiv.org/pdf/1909.11216.pdf</a>	93	Identification, Interpretability, and Bayesian word embeddings	2019	NAACL	<a href="https://www.aclweb.org/anthology/W19-2102.pdf">https://www.aclweb.org/anthology/W19-2102.pdf</a>	166	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://www.aclweb.org/anthology/P16-1031.pdf">https://www.aclweb.org/anthology/P16-1031.pdf</a>
21	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>	94	Identifying and Controlling Important Neurons in Neural Machine Translation	2019	ICLR	<a href="https://arxiv.org/pdf/1811.01157.pdf">https://arxiv.org/pdf/1811.01157.pdf</a>	167	Rethinking Cooperative Rationalization: Introspective Selection and Comple	2018	EMNLP	<a href="https://arxiv.org/pdf/1810.13294.pdf">https://arxiv.org/pdf/1810.13294.pdf</a>
22	Attention is not an Explanation	2019	EMNLP	<a href="https://www.aclweb.org/anthology/D19-1002.pdf">https://www.aclweb.org/anthology/D19-1002.pdf</a>	95	Imparting Interpretability to Word Embeddings while Preserving Semantic Str	2018	TASLP	<a href="https://arxiv.org/pdf/1807.07279.pdf">https://arxiv.org/pdf/1807.07279.pdf</a>	168	Saliency-driven word alignment interpretation for neural machine transl	2019	ACL	<a href="https://www.aclweb.org/anthology/W19-5201.pdf">https://www.aclweb.org/anthology/W19-5201.pdf</a>
23	AttentionMeSH: Simple, Effective and Interpretable Automatic MeSH Indexer	2018	BioASQ	<a href="https://www.aclweb.org/anthology/W18-5306.pdf">https://www.aclweb.org/anthology/W18-5306.pdf</a>	96	Improving Abstractive Document Summarization with Saliency Information Mo	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1205.pdf">https://www.aclweb.org/anthology/P19-1205.pdf</a>	169	Self-Assembling Modular Networks for Interpretable Multi-Hop Reasoning	2019	EMNLP	<a href="https://www.aclweb.org/anthology/D19-1455.pdf">https://www.aclweb.org/anthology/D19-1455.pdf</a>
24	Auditing deep learning processes through kernel-based explanatory models	2019	EMNLP-JCNLP	<a href="https://www.aclweb.org/anthology/D19-1415.pdf">https://www.aclweb.org/anthology/D19-1415.pdf</a>	97	Interpretable emoji prediction via label-wise attention LSTMs	2018	EMNLP	<a href="https://www.aclweb.org/anthology/W18-1508.pdf">https://www.aclweb.org/anthology/W18-1508.pdf</a>	170	Self-Critical Reasoning for Robust Visual Question Answering	2019	NeuIPS	<a href="https://arxiv.org/pdf/1905.09998.pdf">https://arxiv.org/pdf/1905.09998.pdf</a>
25	Automatic rule extraction from long short term memory networks	2017	ICLR	<a href="https://arxiv.org/pdf/1702.02540.pdf">https://arxiv.org/pdf/1702.02540.pdf</a>	98	Interpretable Entity Representations through Large-Scale Typing	2020	EMNLP	<a href="https://arxiv.org/pdf/2004.03147.pdf">https://arxiv.org/pdf/2004.03147.pdf</a>	171	Self-Explaining Structures Improve NLP Models	2020	ArXiv	<a href="https://arxiv.org/pdf/2012.01786.pdf">https://arxiv.org/pdf/2012.01786.pdf</a>
26	BERT Rediscovered the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/P19-1452.pdf">https://www.aclweb.org/anthology/P19-1452.pdf</a>	99	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	<a href="https://arxiv.org/pdf/2011.06854.pdf">https://arxiv.org/pdf/2011.06854.pdf</a>	172	Seq2seq-viz: A visual debugging tool for sequence-to-sequence models	2018	TVG	<a href="https://www.aclweb.org/anthology/D18-1304.pdf">https://www.aclweb.org/anthology/D18-1304.pdf</a>
27	Beyond Word Importance: Contextual Data	2019	ArXiv	<a href="https://arxiv.org/pdf/1902.08792.pdf">https://arxiv.org/pdf/1902.08792.pdf</a>						173	Seq2seq-viz: A visual debugging tool for sequence-to-sequence models	2018	TVG	<a href="https://www.aclweb.org/anthology/D18-1304.pdf">https://www.aclweb.org/anthology/D18-1304.pdf</a>
28	Captum: A unified and generic model interpreter	2019	ArXiv	<a href="https://arxiv.org/pdf/1801.09041.pdf">https://arxiv.org/pdf/1801.0904</a>										

# Matching XAI Papers with XAI Question Bank?

43 User Questions

- |               |   |
|---------------|---|
| <b>Input</b>  | <ul style="list-style-type: none"><li>• What kind of data does the system learn from?</li><li>• What is the source of the data?</li><li>• How were the labels/ground-truth produced?</li><li>• * What is the sample size?</li><li>• * What data is the system NOT using?</li><li>• * What are the limitations/biases of the data?</li></ul> |
| ⋮             |   |
| <b>Others</b> | <ul style="list-style-type: none"><li>• * How/what/why will the system change/adapt/improve/drift over time? (change)</li><li>• * How to improve the system? (change)</li><li>• * Why using or not using this feature/rule/data? (follow-up)</li><li>• * What does [ML terminology] mean? (terminological)</li></ul>                        |



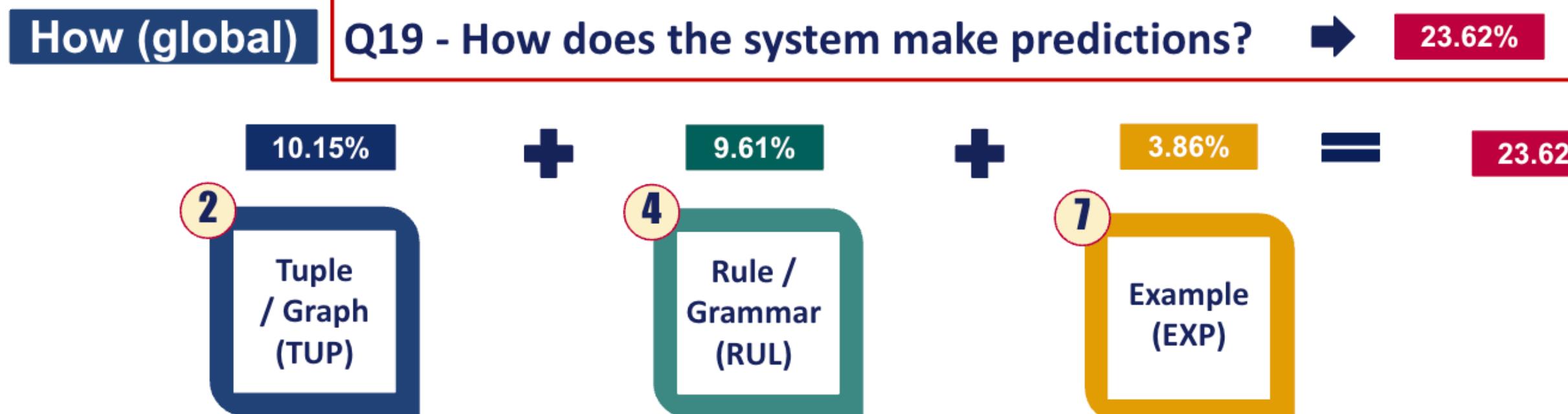
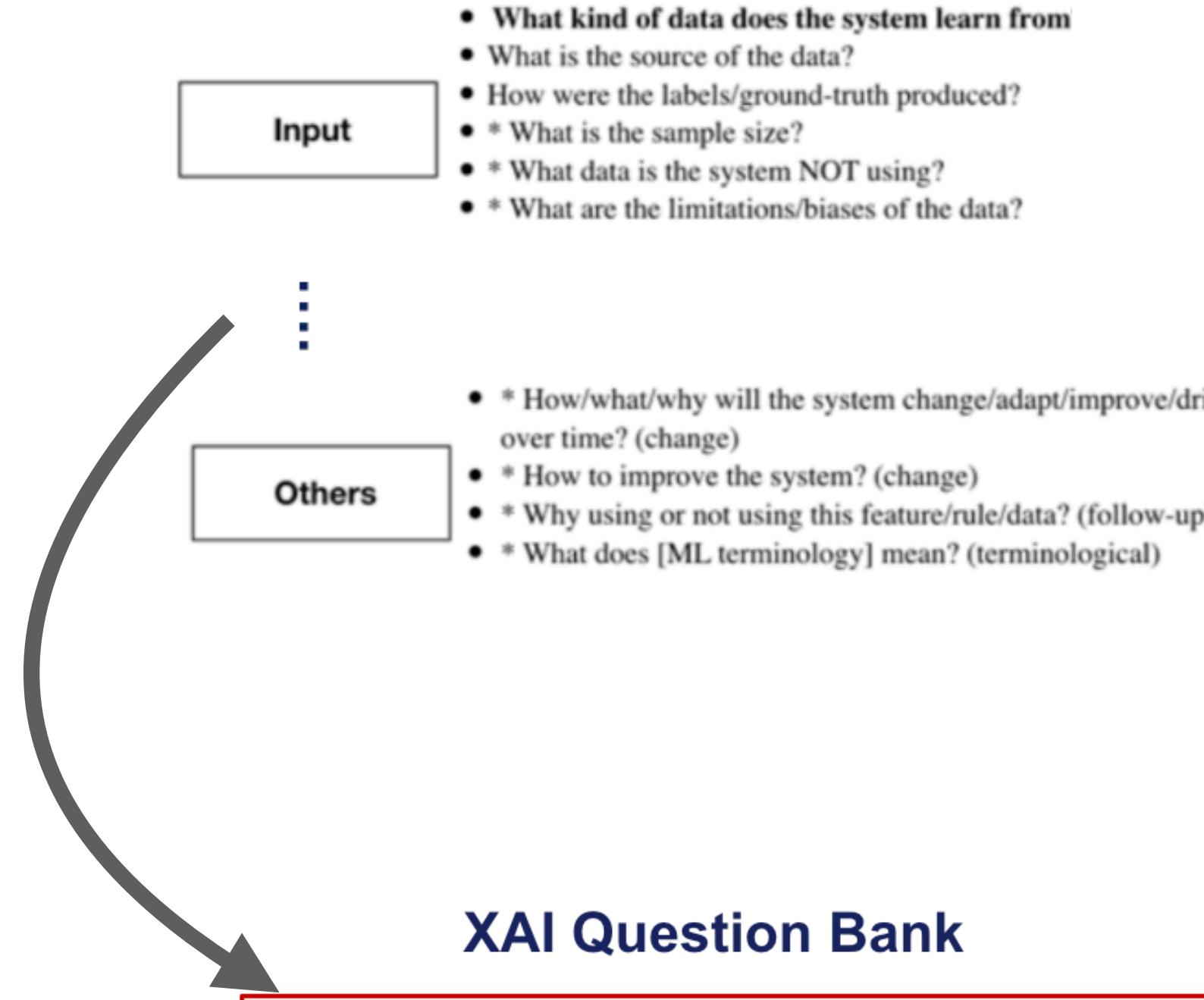
218 XAI Papers

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

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

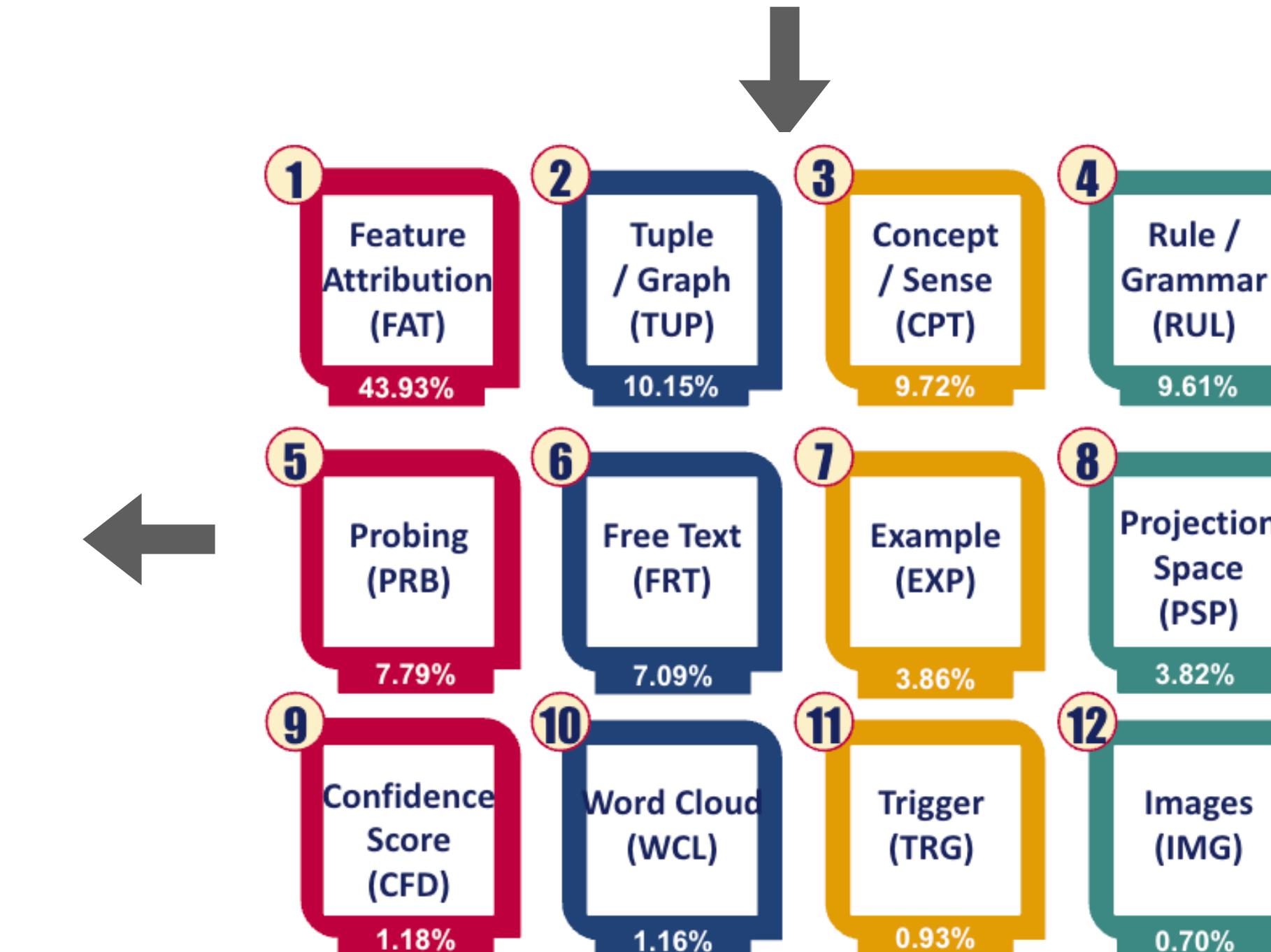
# Matching Each User Question with XAI Forms in NLP

43 User Questions

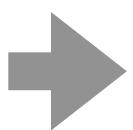
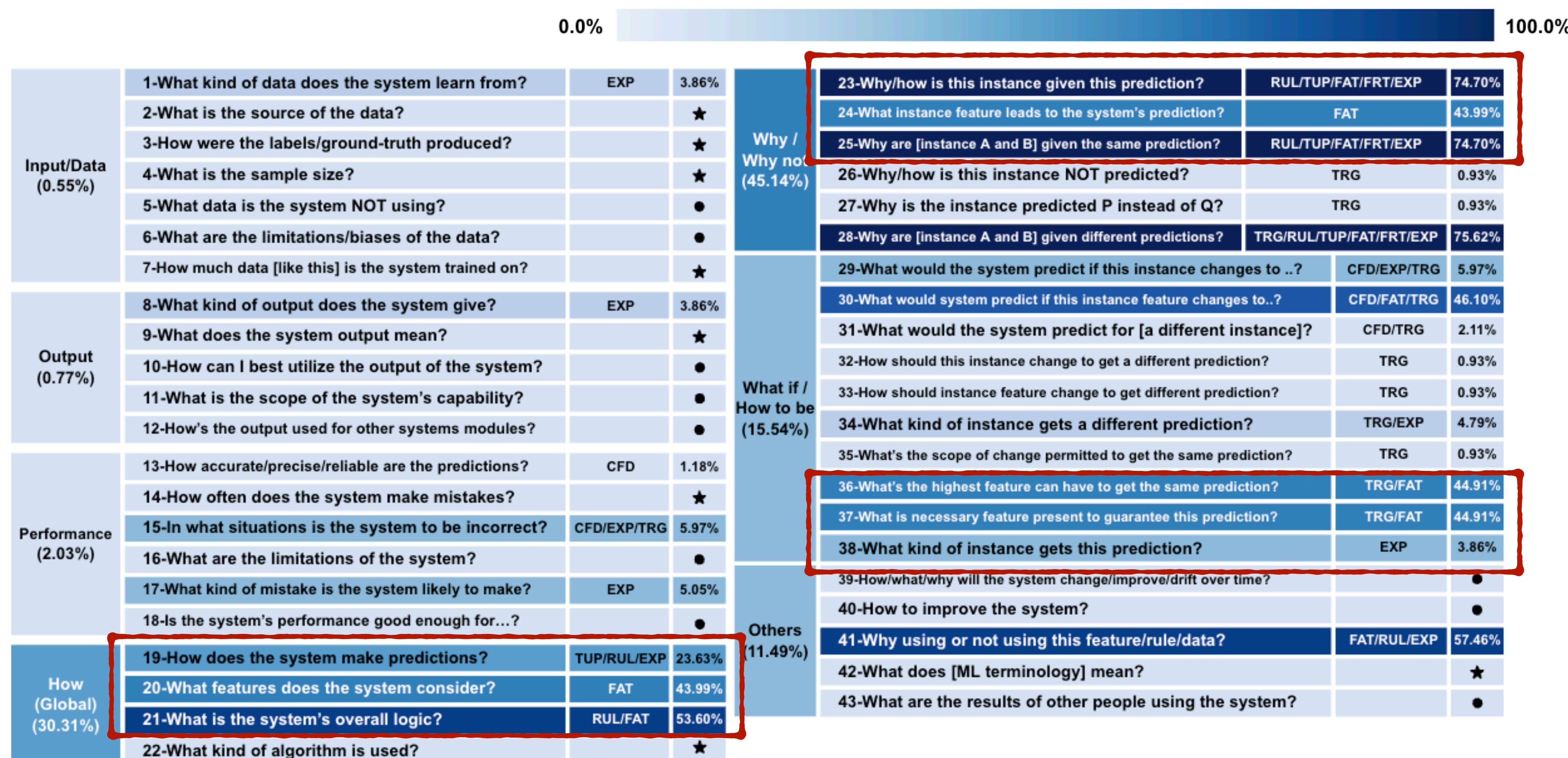


200+ XAI Papers

ID	Title	Year	Venue	Paper URL
1	"Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	<a href="https://arxiv.org/pdf/1602.04938.pdf">https://arxiv.org/pdf/1602.04938.pdf</a>
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	<a href="https://www.aclweb.org/anthology/N16-1001.pdf">https://www.aclweb.org/anthology/N16-1001.pdf</a>
3	Rationalizing Neural Predictions	2016	EMNLP	<a href="https://people.csail.mit.edu/taolei/paper/RNP.pdf">https://people.csail.mit.edu/taolei/paper/RNP.pdf</a>
4	BERT RedisCOVERS the Classical NLP Pipeline	2019	ACL	<a href="https://www.aclweb.org/anthology/C19-1001.pdf">https://www.aclweb.org/anthology/C19-1001.pdf</a>
5	Attention is not Explanation	2019	NAACL	<a href="https://arxiv.org/pdf/1902.10186.pdf">https://arxiv.org/pdf/1902.10186.pdf</a>
...	...	...	...	...
214	How much should you ask? On the question structure in QA systems	2018	BlackboxNLP	<a href="https://arxiv.org/pdf/1809.03734.pdf">https://arxiv.org/pdf/1809.03734.pdf</a>
215	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	<a href="https://arxiv.org/pdf/2011.06854.pdf">https://arxiv.org/pdf/2011.06854.pdf</a>
216	A Survey of the State of Explainable AI for Natural Language Processing	2020	AACL-IJCNLP	<a href="https://arxiv.org/pdf/2010.00711.pdf">https://arxiv.org/pdf/2010.00711.pdf</a>
217	Explaining Simple Natural Language Inference	2019	ACL	<a href="https://www.aclweb.org/anthology/C19-1002.pdf">https://www.aclweb.org/anthology/C19-1002.pdf</a>
218	Understanding Neural Abstractive Summarization Models via Uncertainty	2020	EMNLP	<a href="https://arxiv.org/pdf/2010.07882.pdf">https://arxiv.org/pdf/2010.07882.pdf</a>

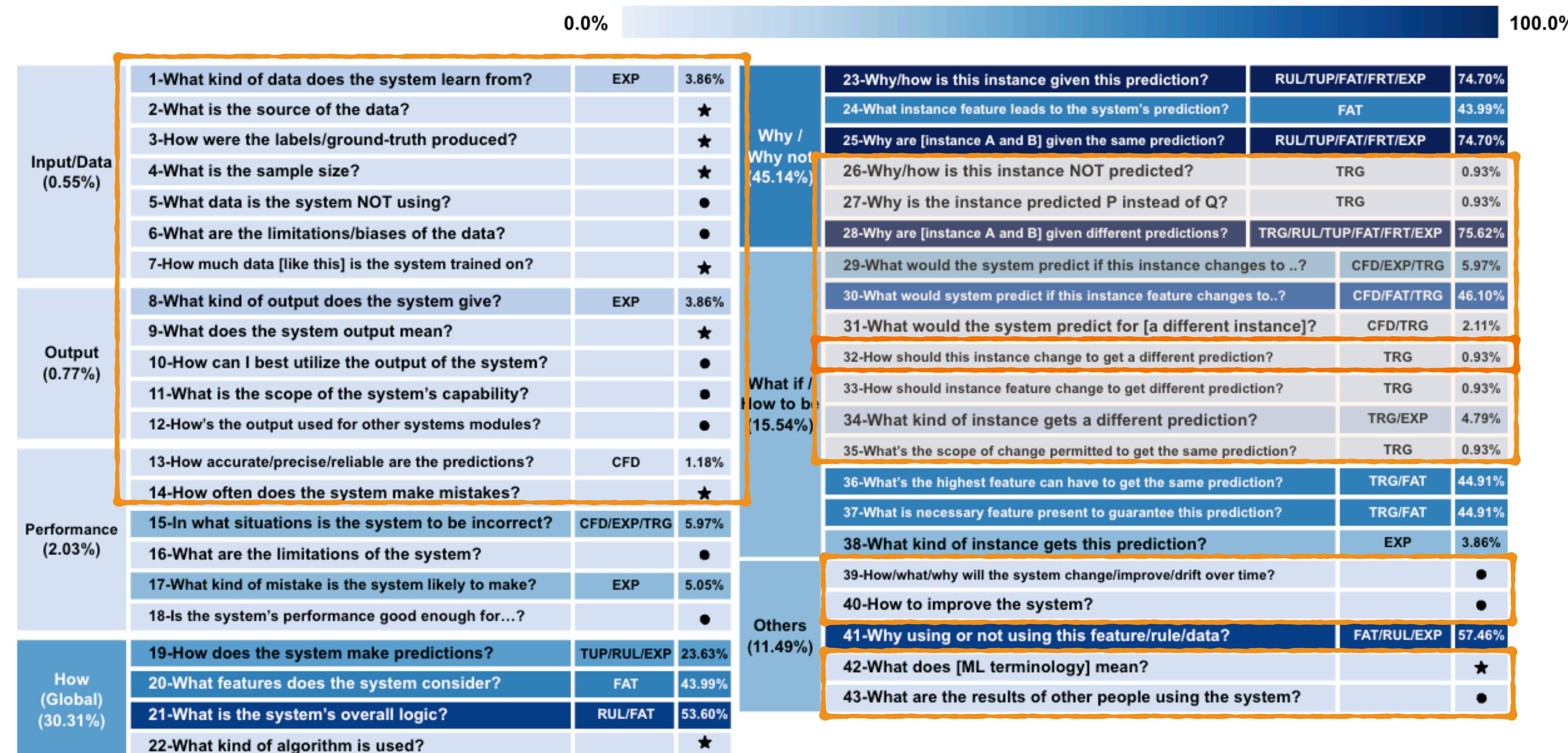


# Results: Heatmap for XAI Question Bank



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

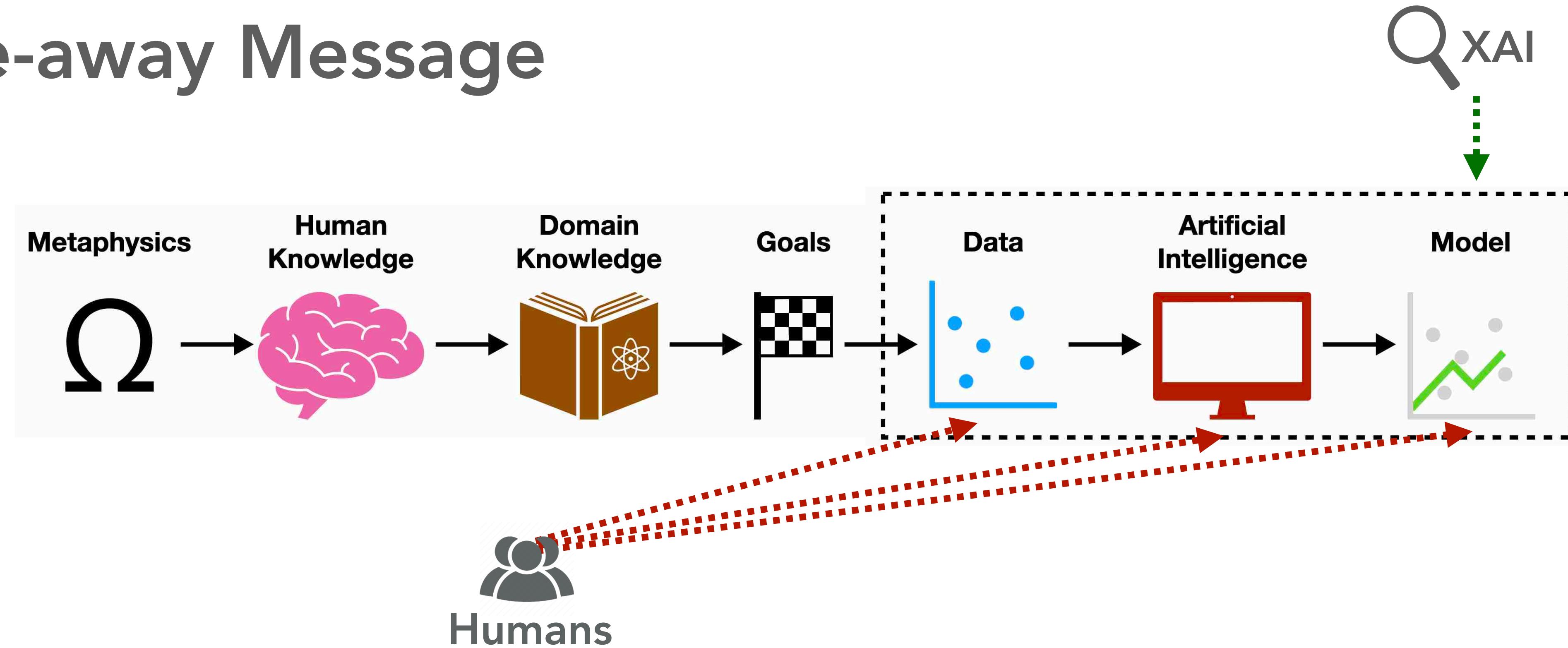
# Results: Heatmap for XAI Question Bank



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

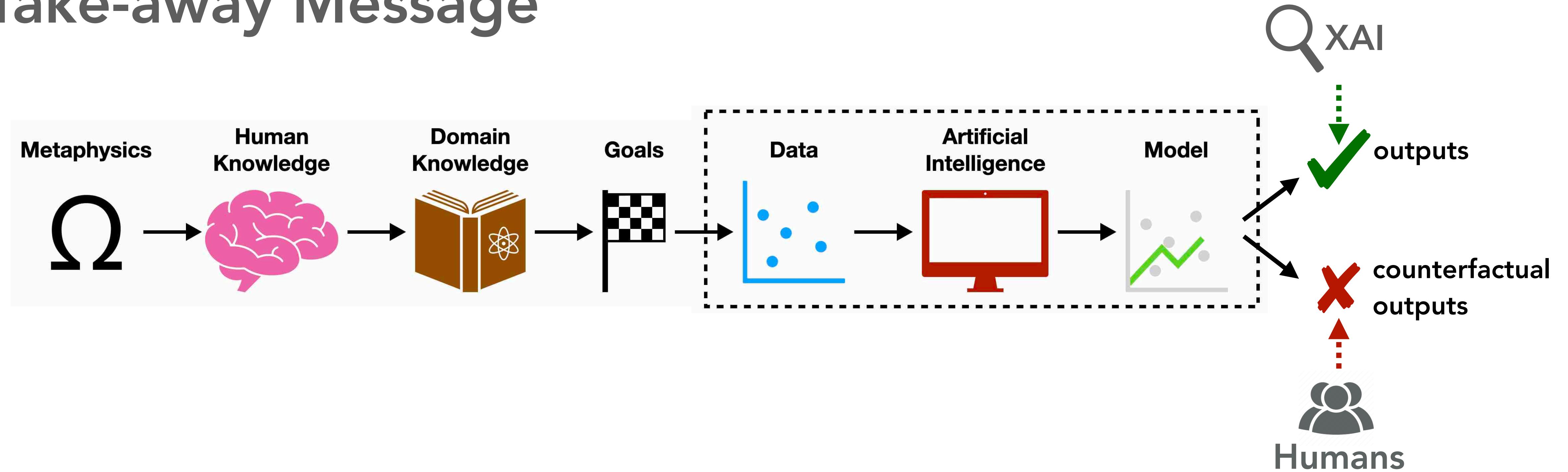
→ XAI largely ignored: diverse information across AI development process (data, model, deployment, etc.)

# Take-away Message



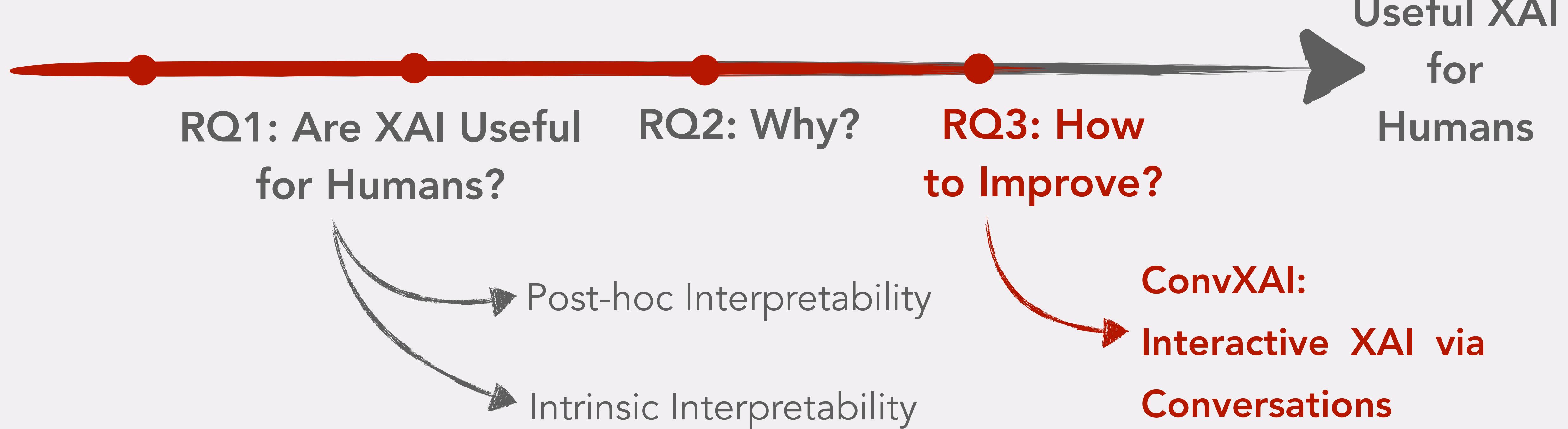
**Users demand diverse XAI questions to understand AI models across AI development lifecycle, whereas existing XAI methods commonly answer only ONE XAI question.**

# Take-away Message



Users are widely interested in **what AI systems cannot achieve** other than what AI already succeeded, indicating the necessity 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

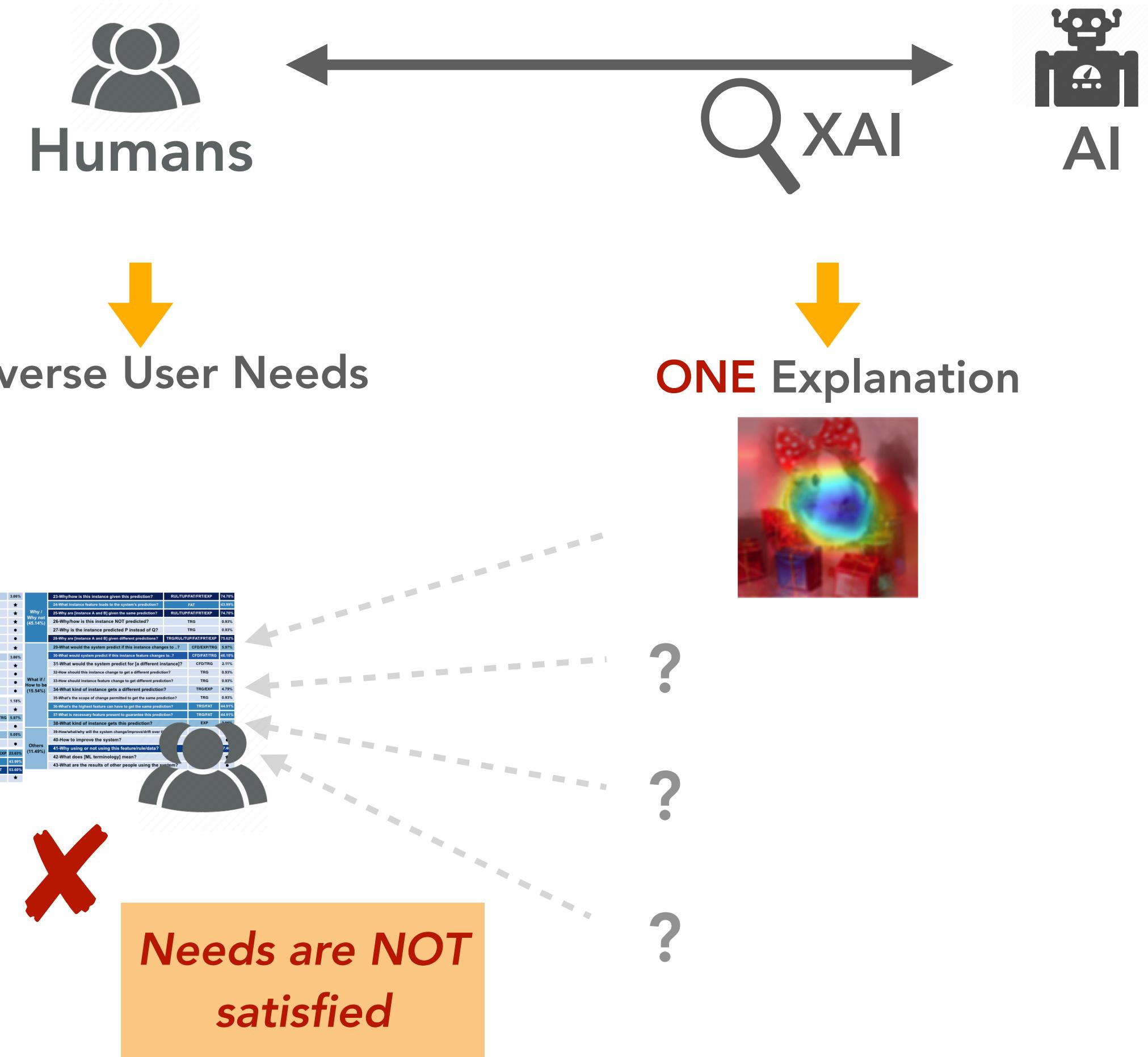


Sherry Wu

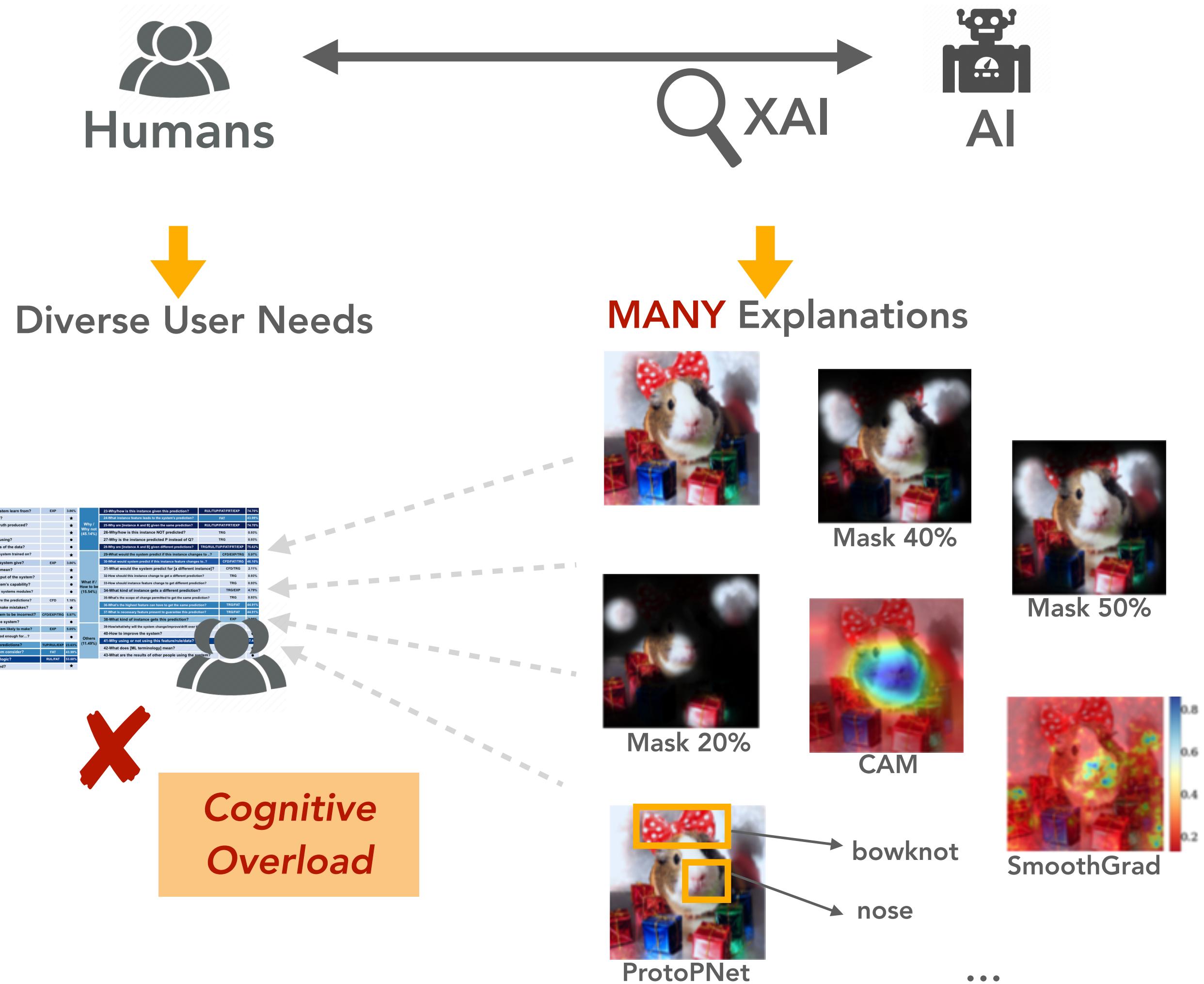


Kenneth Huang

# Challenges of Existing XAI

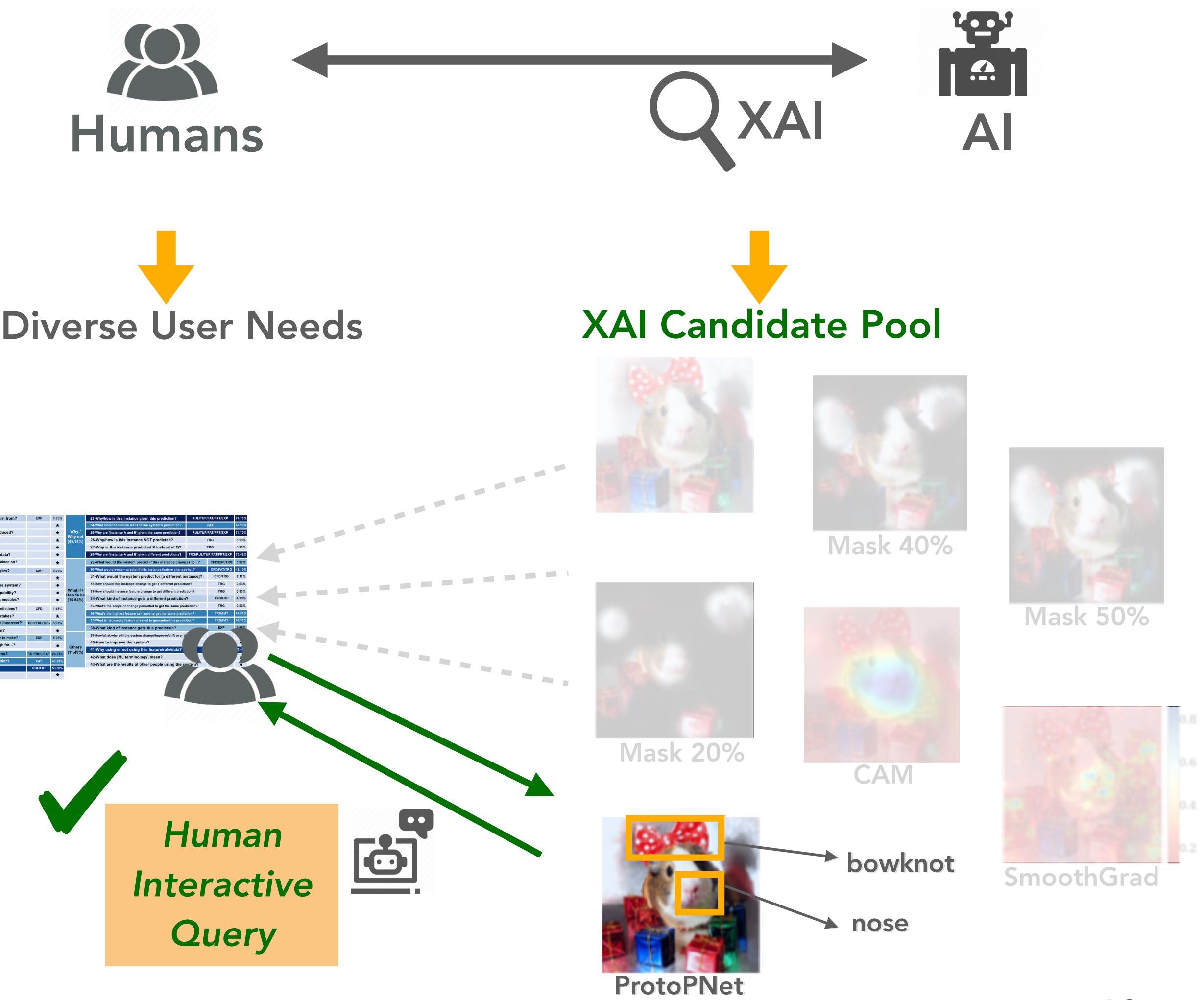


# Challenges of Existing XAI



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

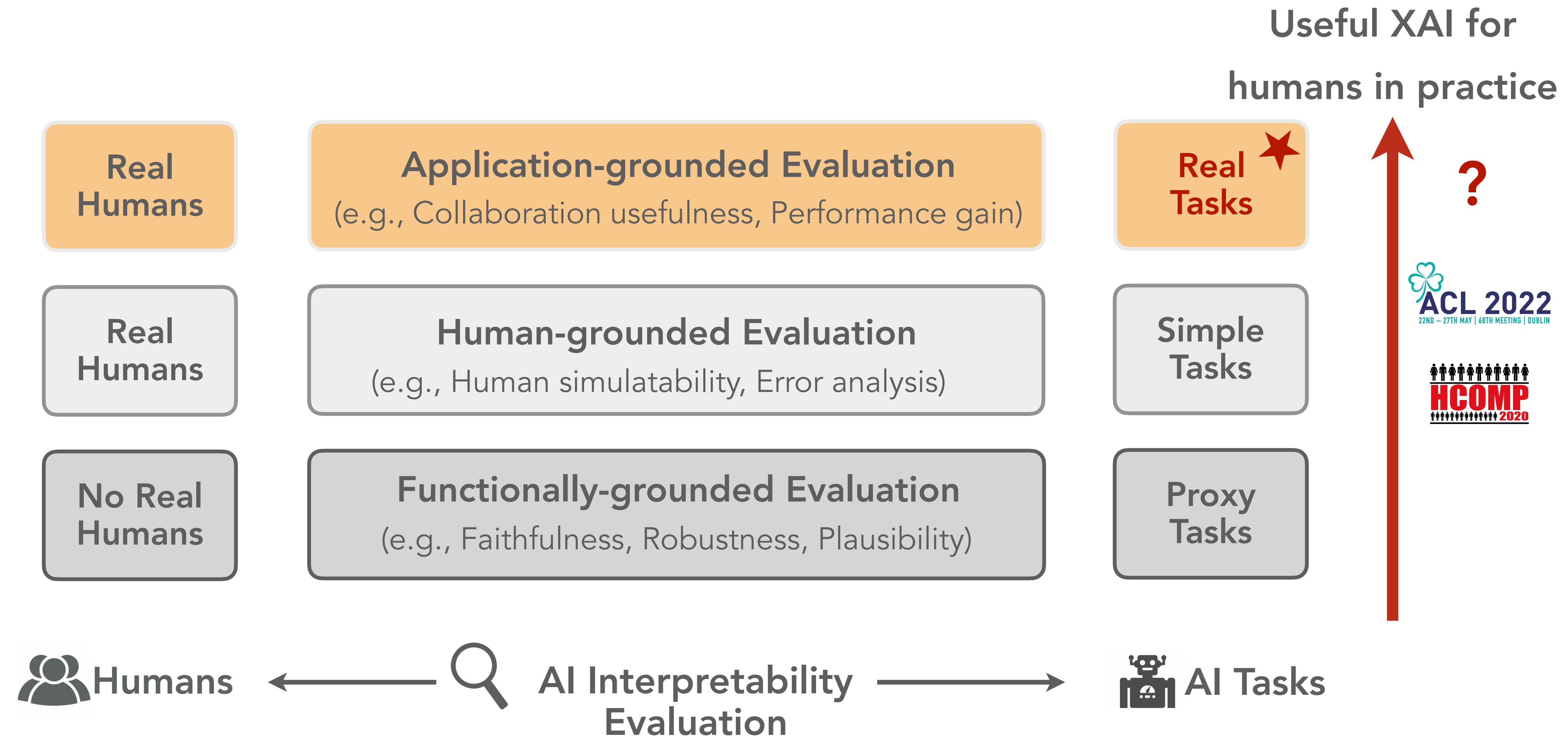
# Solution: Conversational XAI



- 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



The screenshot shows the "Scientific Writing Support" application interface:

- Top Bar:** "Scientific Writing Support" with a feather icon.
- Conference Selection:** "Which conference are you most likely to submit this paper abstract to?" dropdown menu showing "CHI (Human-Computer Interaction Domain)".
- Abstract Example Selection:** "Select an abstract example to try:" dropdown menu showing "Select an abstract example".
- Text Editor:** "Or Edit your paper abstract:" rich text editor toolbar with buttons for Normal, B, I, S, U, etc.
- Submit Button:** "Click to Submit Your Writing" button.
- Model Selection:** "Click below buttons to switch the model's prediction on each sentence." with "Writing Structure Model" (blue button) and "Writing Style Model" (red button).
- Aspect Labels:** A note: "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." followed by buttons for "Background", "Purpose", "Method", "Finding/Contribution", and "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

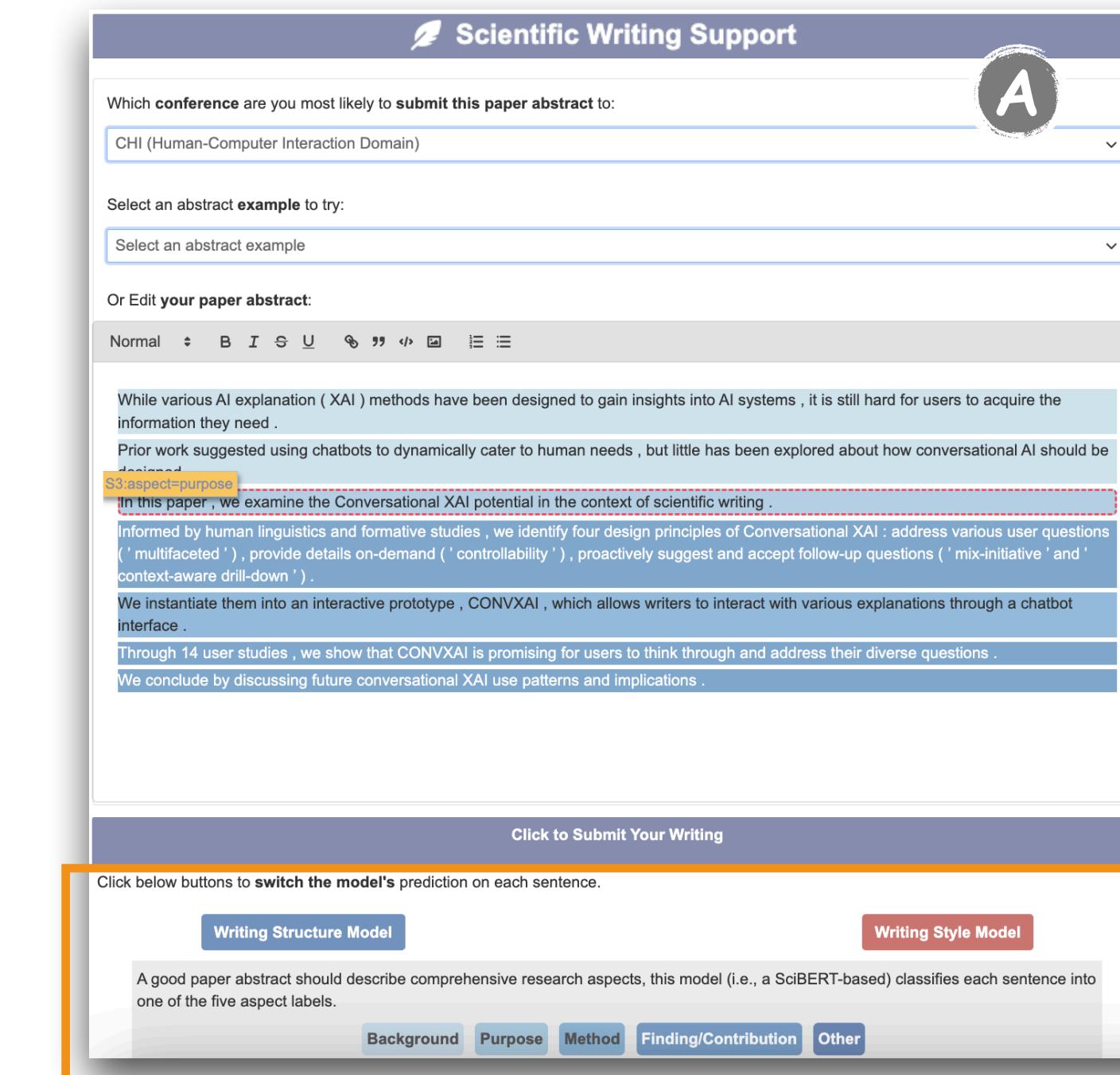
## Participants

- 3 females, 4 males
- diverse background

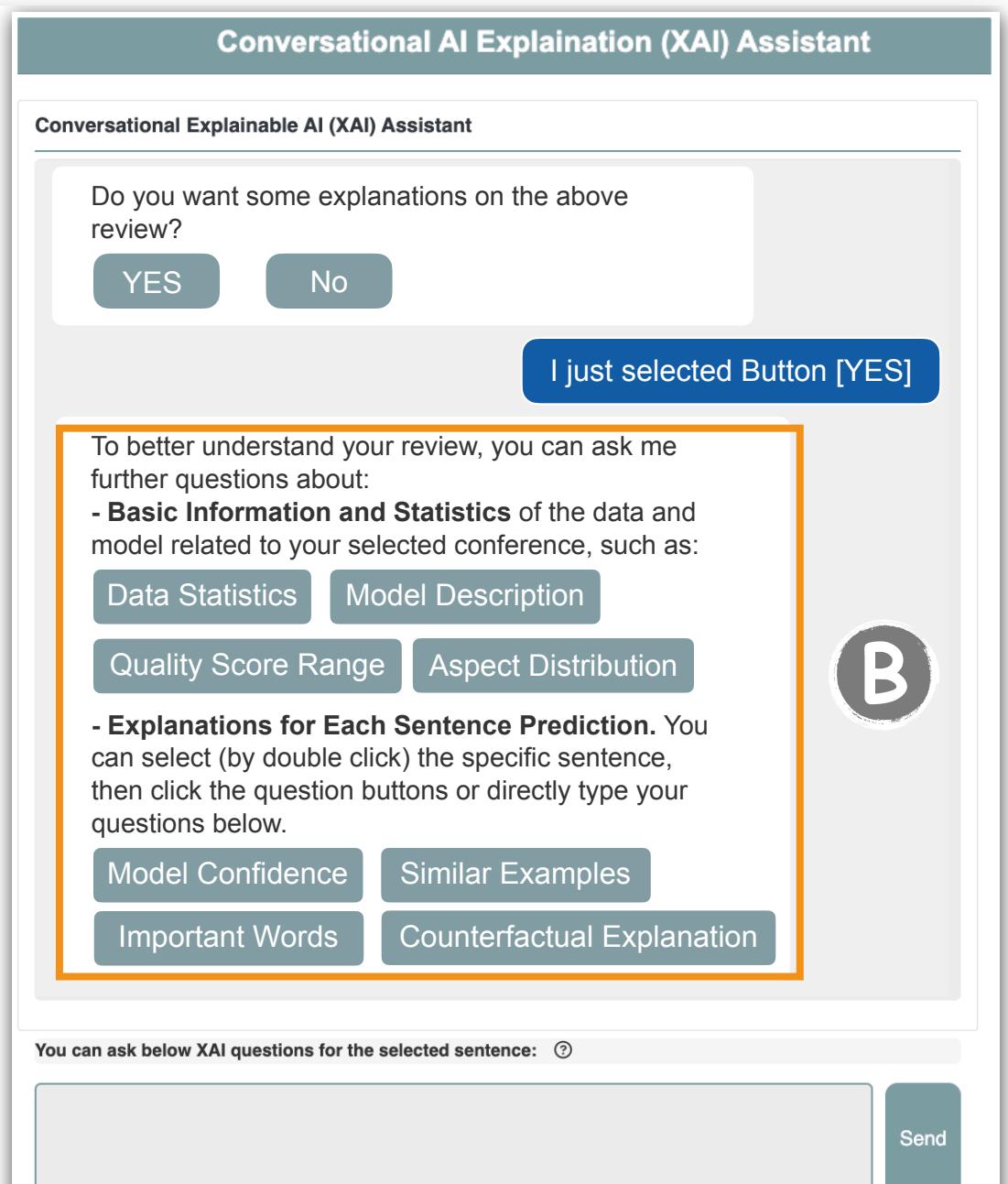
## Procedure

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

## Four Design Principles for Conversational XAI



Predict with 2 AI writing models



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)

Select an abstract example to try:

Select an abstract example

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 XAI should be designed .

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

**A** 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:

**B** 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 ' ).

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

**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.

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?

**D**

## Keynotes of User Interaction

## Implementing Four Design Principles

**P1** Multifaceted

Mixed-Initiative

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

**P2**

Context-aware Drill-down

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.

**P3**

Controllability

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'.

**P4**

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.

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

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

Who is studied	 <b>Task1</b> <b>13</b> graduate researchers	 <b>Task2</b> <b>8</b> researchers
	09/2022 (90min)	12/2022 (90min) ( <b>rejoin</b> )
When		
How it's studied	<ol style="list-style-type: none"><li>1. <b>Two</b> think-aloud <b>scientific writing tasks</b>:<ul style="list-style-type: none"><li>• Within-Subjects Study: ConvXAI vs. Baseline</li><li>• Improve a paper's abstract;</li><li>• Paper domains: NLP, or HCI, or AI</li></ul></li><li>2. Post <b>Survey</b> - Questionnaires</li><li>3. Semi-<b>structured Interviews</b></li></ol>	
What is the goal	<ul style="list-style-type: none"><li>• Effective system design</li><li>• Self-perceived usefulness</li></ul>	<ul style="list-style-type: none"><li>• Writing output quality</li><li>• Temporal change of user need</li></ul>

# Baseline System (SelectXAI)

## Within-Subjects Study Design

The image shows two panels of the Baseline System (SelectXAI) interface. On the left is the "Scientific Writing Support" panel (labeled A), and on the right is the "AI Explanation (XAI) Panel" (labeled B).

**Scientific Writing Support Panel (A):**

- Header: Scientific Writing Support
- Form fields:
  - 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:
- Text area:

While various AI explanation ( XAI ) methods have been designed to explain the predictions of machine learning systems , it is still hard for users to acquire the information they need to effectively use them . Prior work suggested using chatbots to dynamically cater to human needs . 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 ' ).
- Buttons:
  - Click to Submit Your Writing
  - Click below buttons to switch the model's prediction on each sentence.
    - Writing Structure Model
    - Writing Style Model
- Model labels:

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

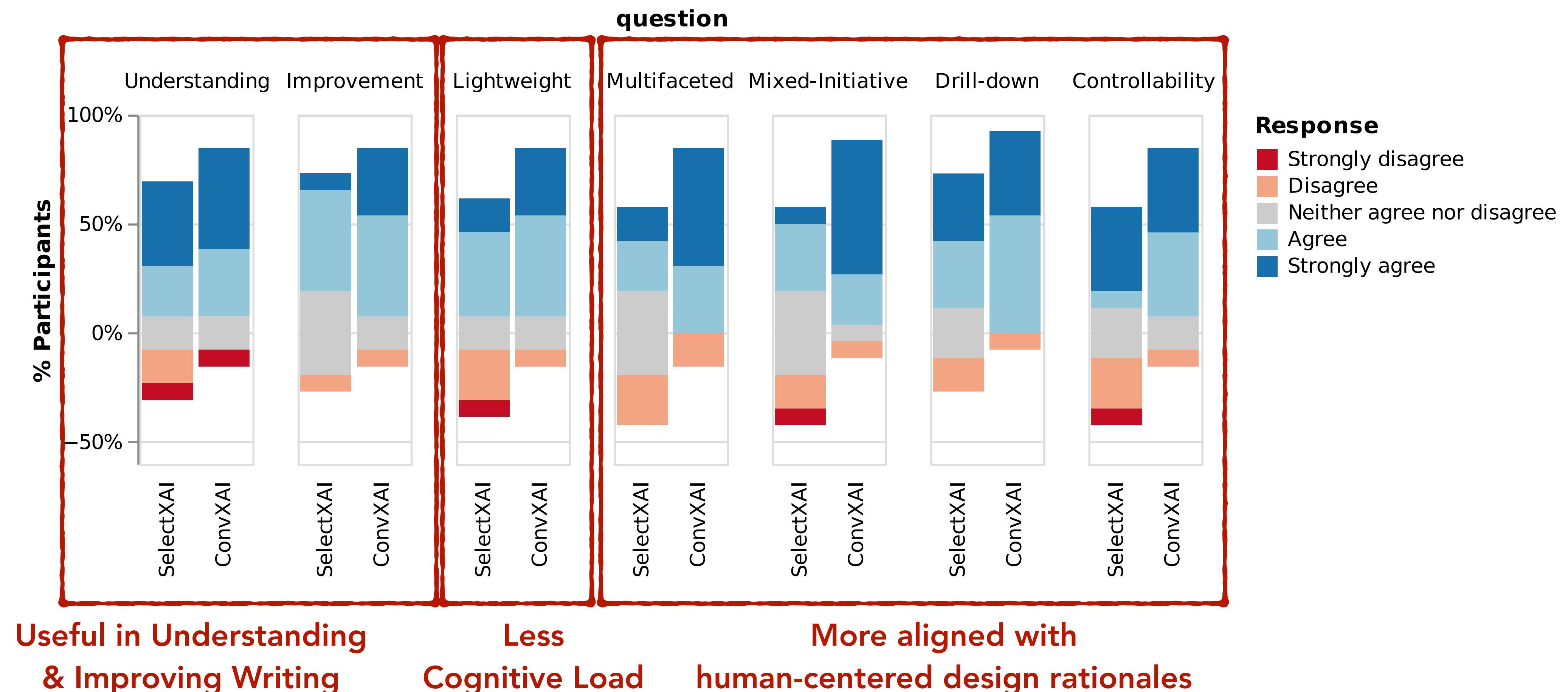
**AI Explanation (XAI) Panel (B):**

- Header: AI Explanation (XAI) Panel
- Section: Writing Feedback Summary
  - B Nice! I'm comparing your submission with 3235 CHI paper abstracts.
  - Your Overall Score of Structure and Style = 3 (out of 5).
- Section: Structure Suggestions:

Based on the sentence labels' percentage and order you used to write your background purpose here .
- Section: Sentence-wise Explanations
  - Data Statistics (What data did the system learn from?) (highlighted with a red arrow and hand icon)
  - 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?)

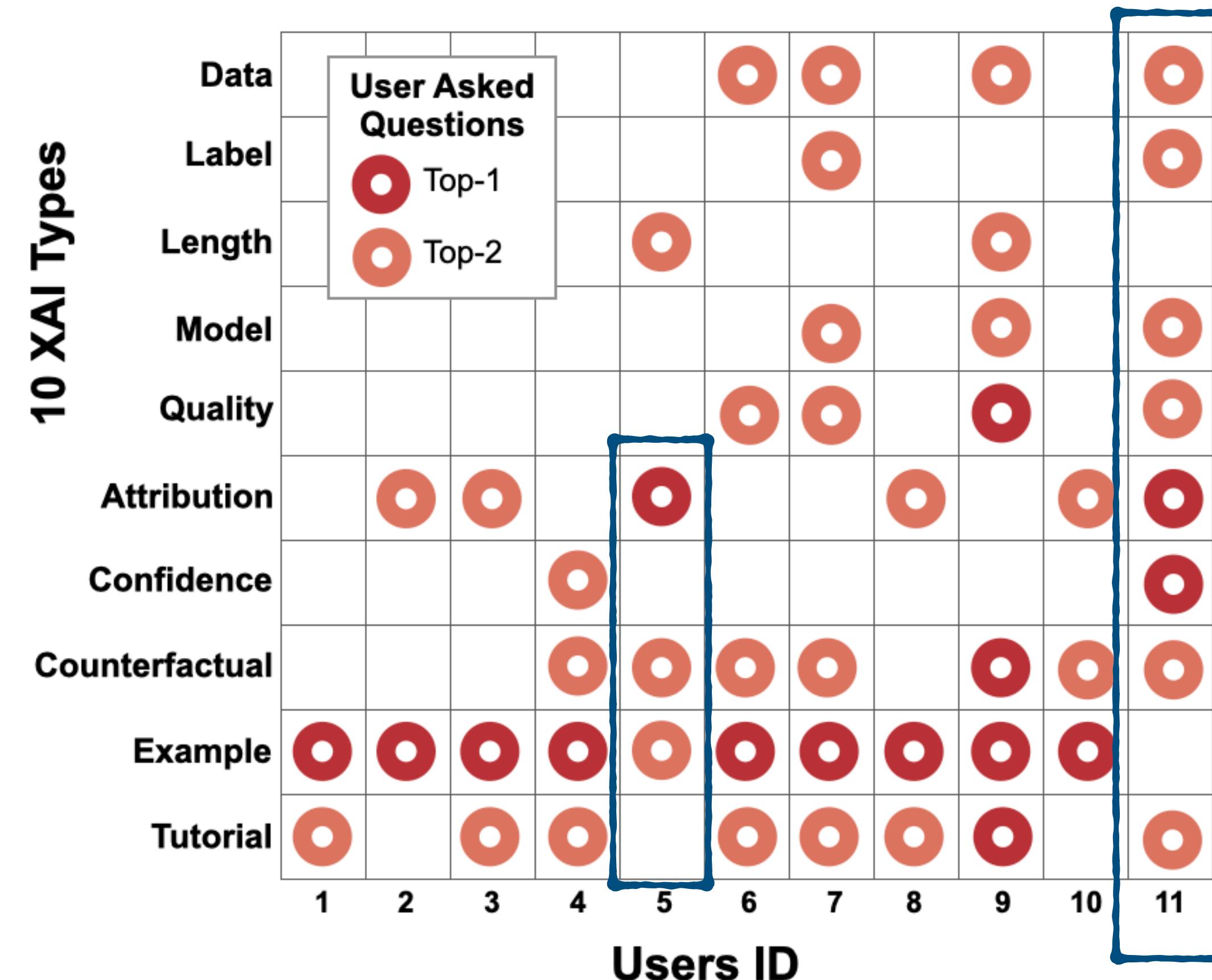
# Survey results of human study in Task1

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	<b>4.25 ( ±1.389)</b>	<b>4.375 ( ±1.408)</b>	<b>4 ( ±1.414)</b>
P	0.1248	0.1624	0.1489

Table1. Survey results of human-perceived usefulness rating.

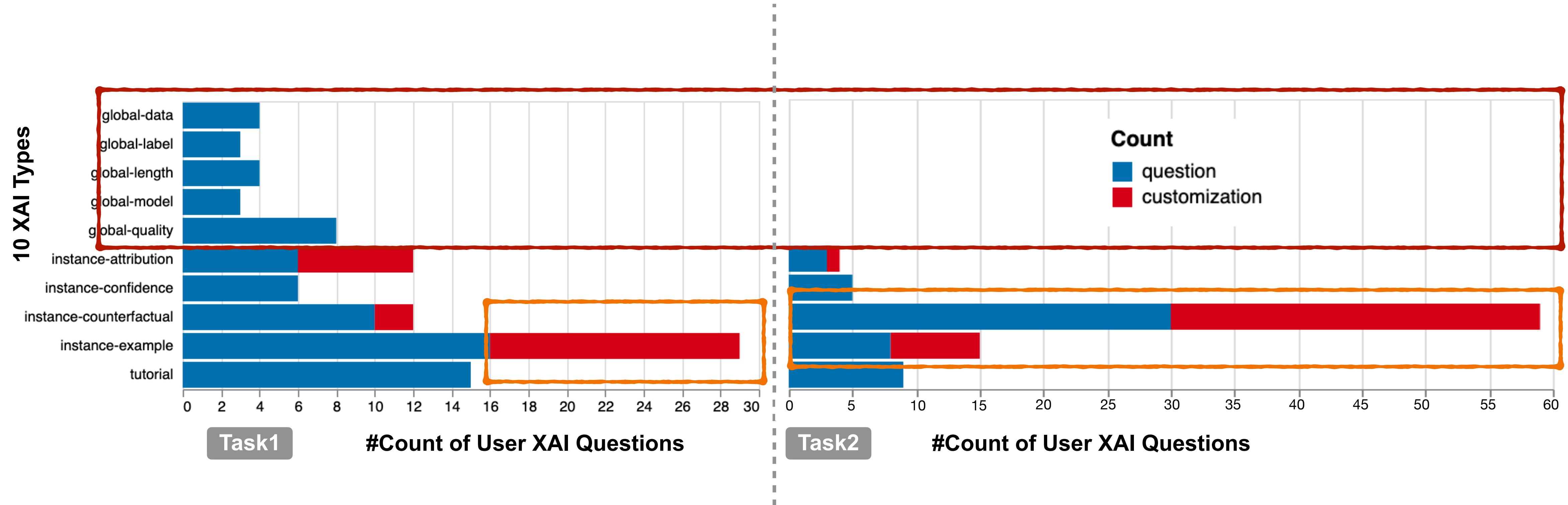
Condition ↑	Grammarly (1-100)		Model Quality (1-5)		Model Structure (1-5)		Human Quality (1-10)		Human Structure (1-10)	
	Original	Improved	Original	Improved	Original	Improved	Original	Improved	Original	Improved
SelectXAI	84.8 (±10.4)	85.1 (±5.52)	2.82 ( 0.75)	3.05 ( 0.64)	4.19 ( 0.37)	<b>4.75 ( 0.38)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

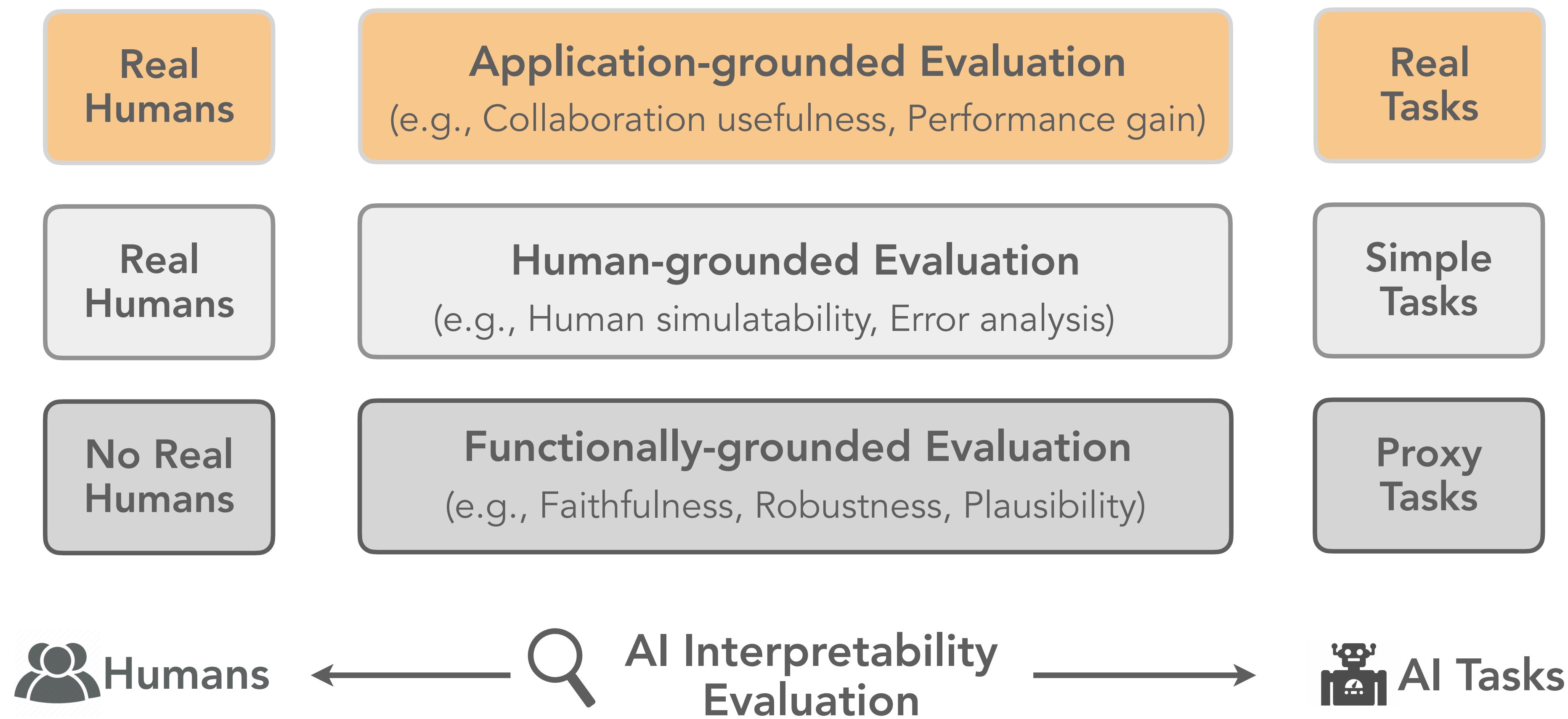
ConvXAI is a potentially useful human-centered XAI approach that empowers humans to interactively inquire heterogeneous AI Explanations via a simple conversation interface.

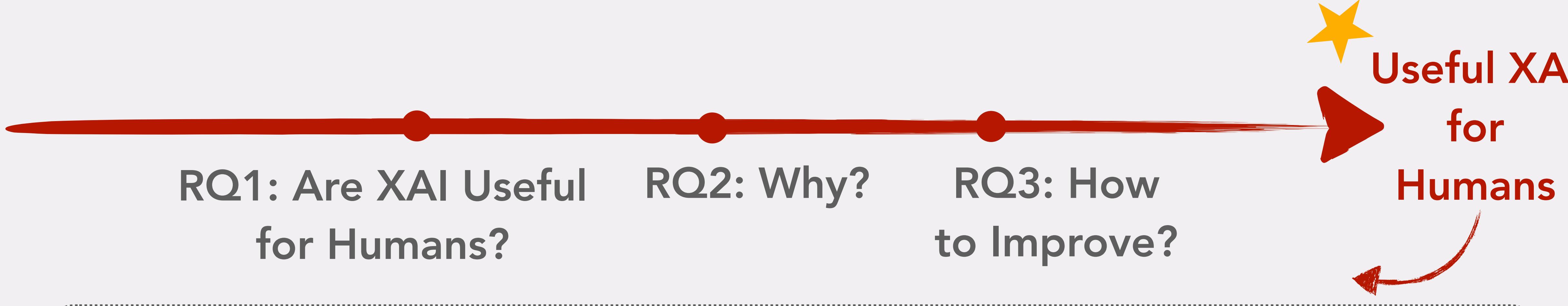
## **Key ingredients for useful XAI for humans in practice:**

- 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





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](https://github.com/huashen218/convxai/blob/main/notebooks/convxai_universal_xai_api.ipynb)

The screenshot shows three GitHub repository pages side-by-side:

- Top Left:** huashen218 / GuessWrongLabel (Public). This repository contains code for a human evaluation tool. It has 2 forks and 2 stars.
- Top Middle:** huashen218 / LimitedInk (Public). This repository contains code for a limited ink model and human study. It has 1 fork and 2 stars.
- Bottom Right:** huashen218 / convxai (Public). This repository contains code for the ConvXAI system. It has 0 forks and 0 stars. The repository description states: "ConvXAI: Towards Conversational Explainable AI Support for Scientific Writing". It includes sections for About, Contributors, Languages, and Suggested Workflows.

# 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.



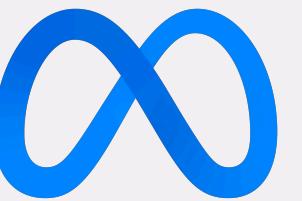
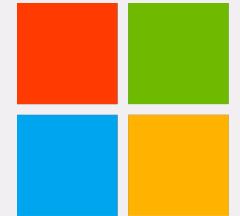
Carnegie  
Mellon  
University

## Committee Members



## Mentors & Collaborators

amazon



W

