PennState
College of Information
Sciences and Technology

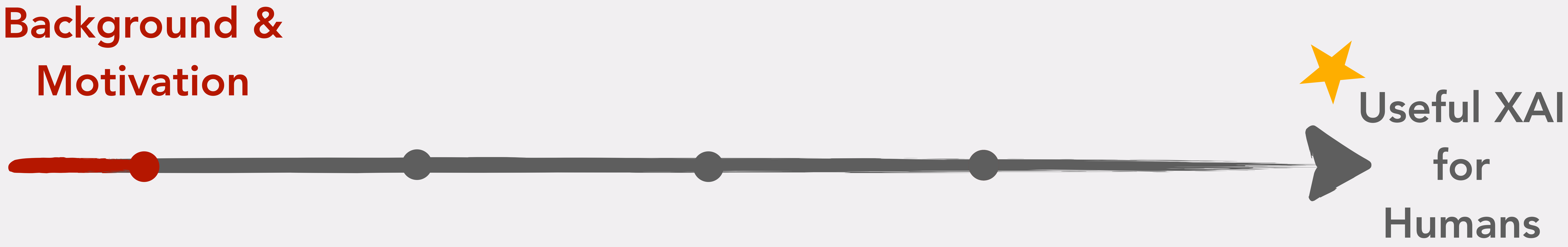
Towards Useful AI Interpretability for Humans via Interactive AI Explanations

Hua Shen

@huashen218 huashen218@psu.edu

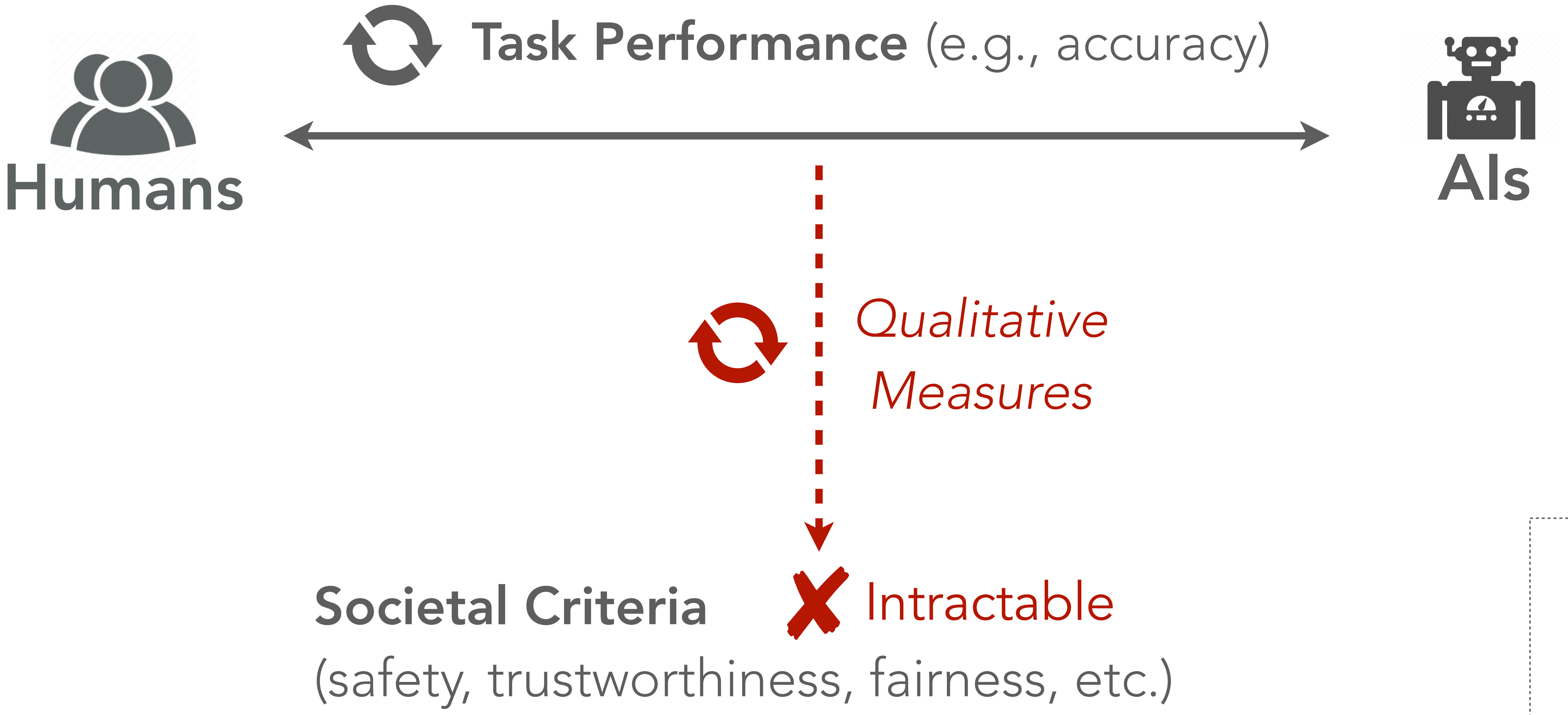
Penn State University

8/2023


Background &
Motivation
Useful XAI
for
Humans

Why do we **need AI interpretability**?

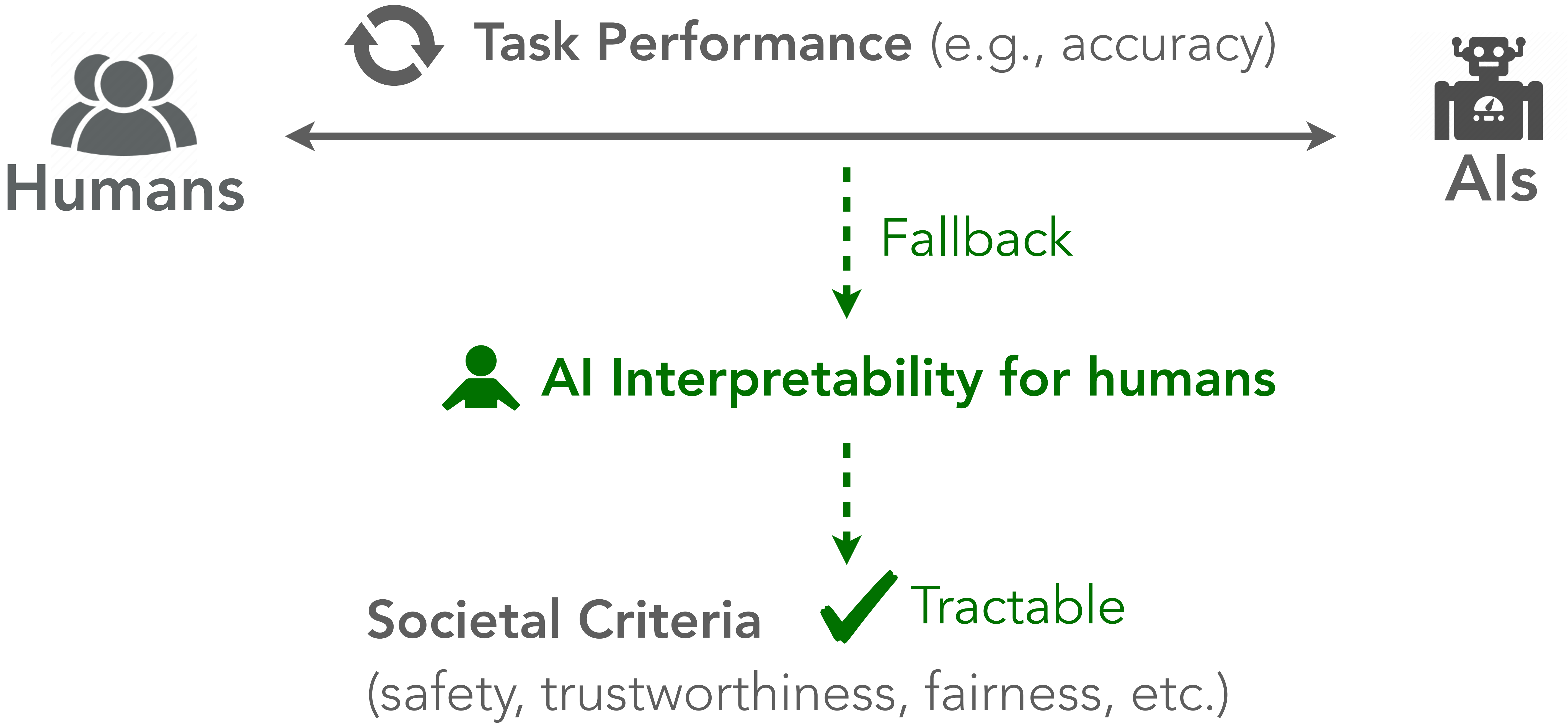
Human-AI collaborative systems are not only **optimized** for **task performance** (e.g., accuracy), but also are required to **satisfy** vital **societal criteria** (e.g., trustworthiness, safety, fairness, etc.).



Human-AI collaborative systems

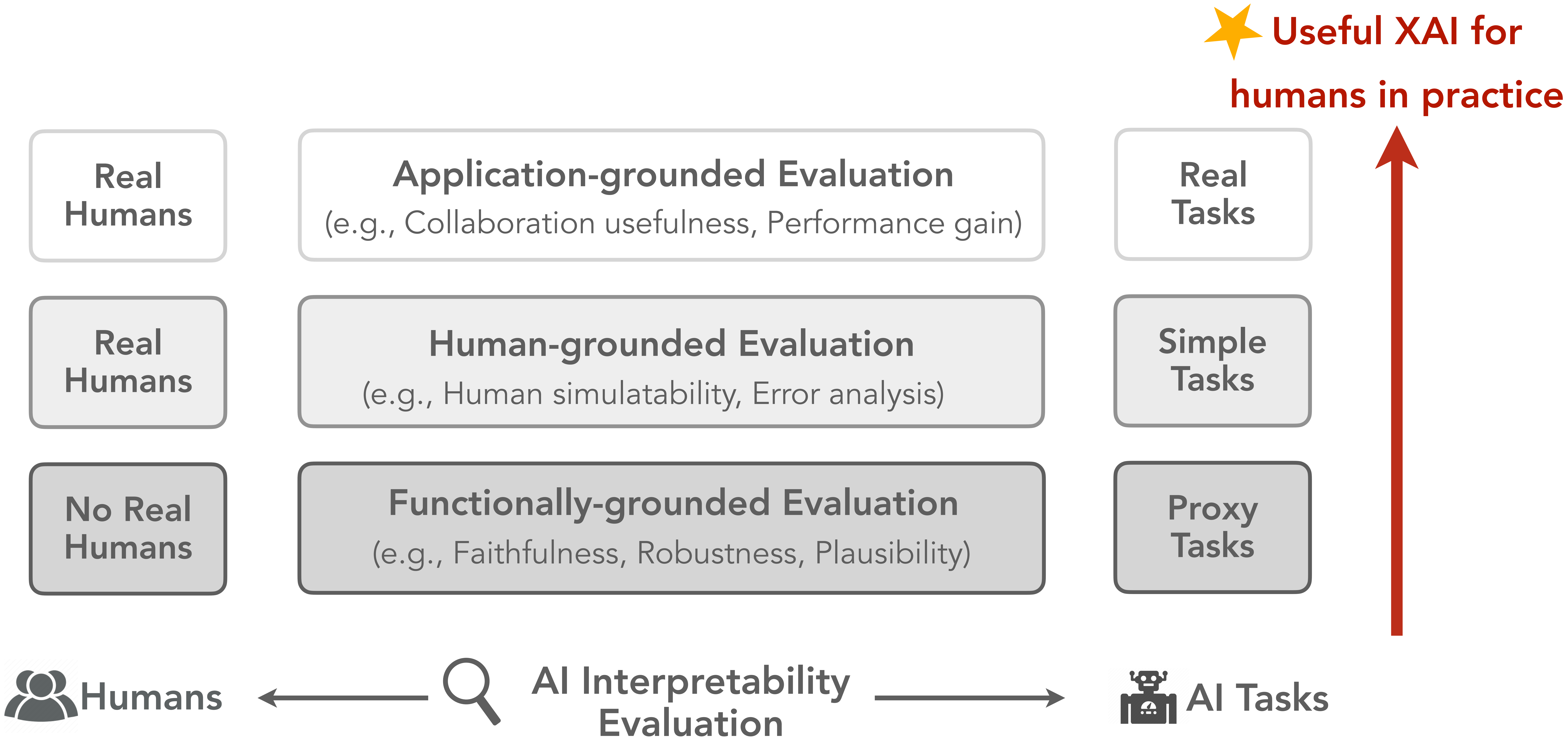
The usefulness of XAI for humans is crucial

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



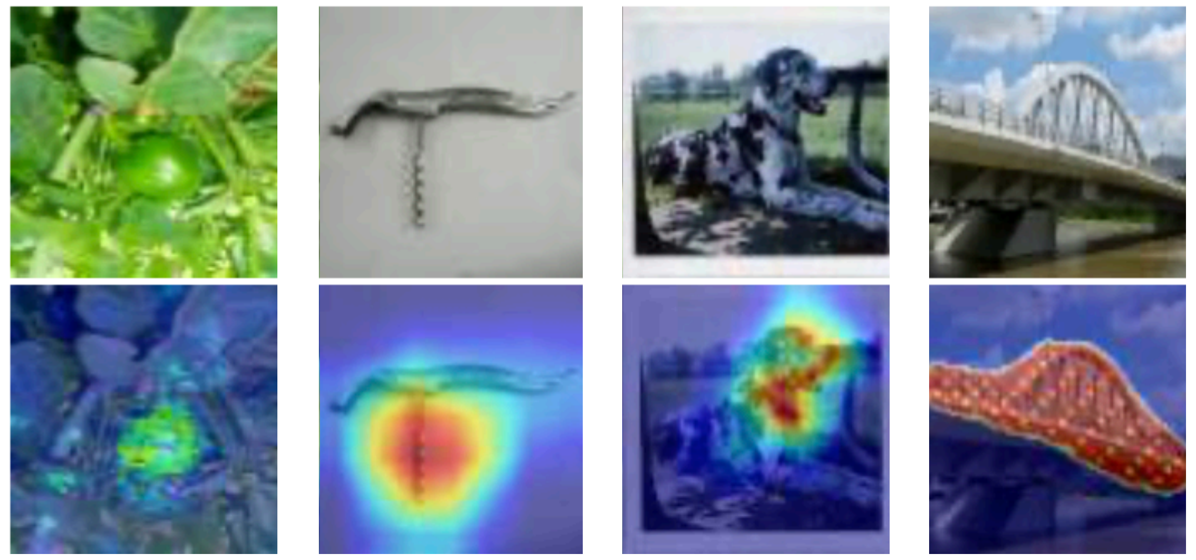
Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Evaluation of XAI usefulness



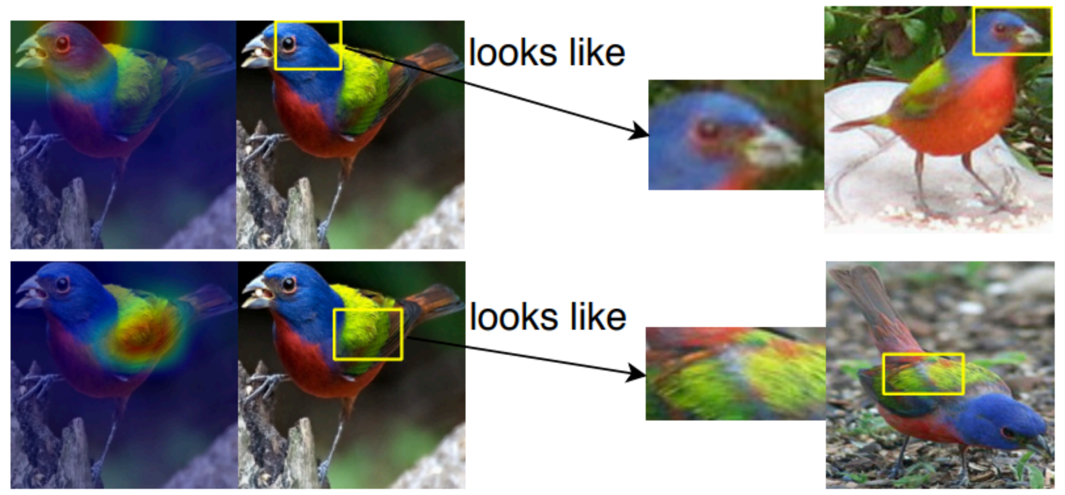
(Doshi-Velez, Finale, & Kim, Been. (2017))

Trajectory of XAI algorithms



Gradient CAM Mask RTS

Attribution Heatmaps



(e.g. ProtoPNet)

Prototype Explanations

(e.g. Influence Functions, Representor Point)

Training Examples

(e.g. SCOUT, Triggers)

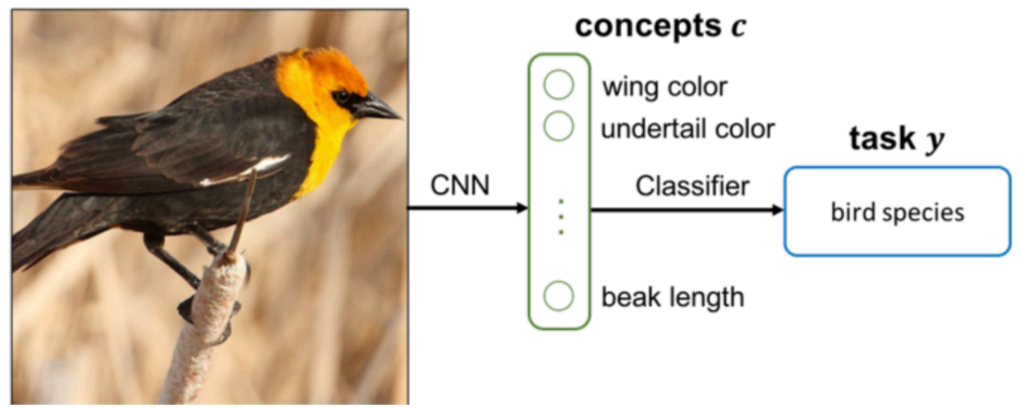
Counterfactuals

....

2012 → 2023

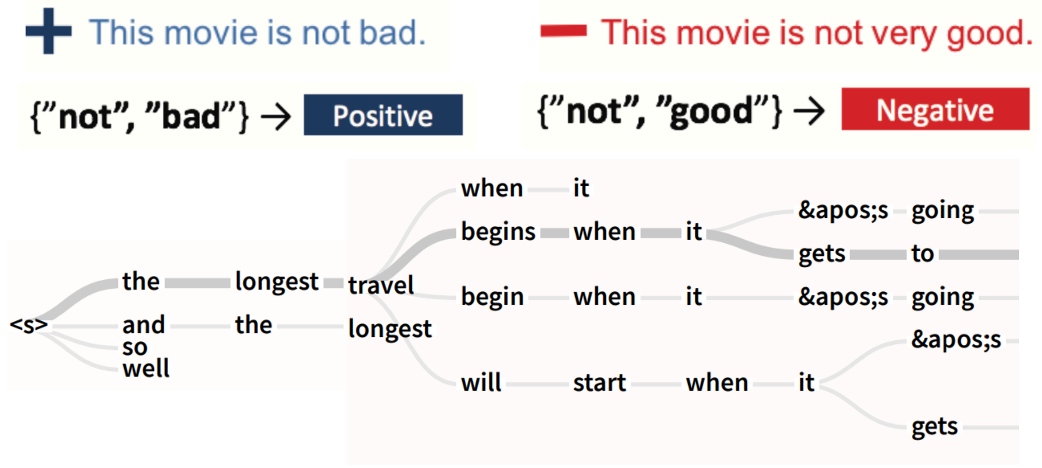
Concept-based Explanations

(e.g. TCAV, Concept Bottleneck)



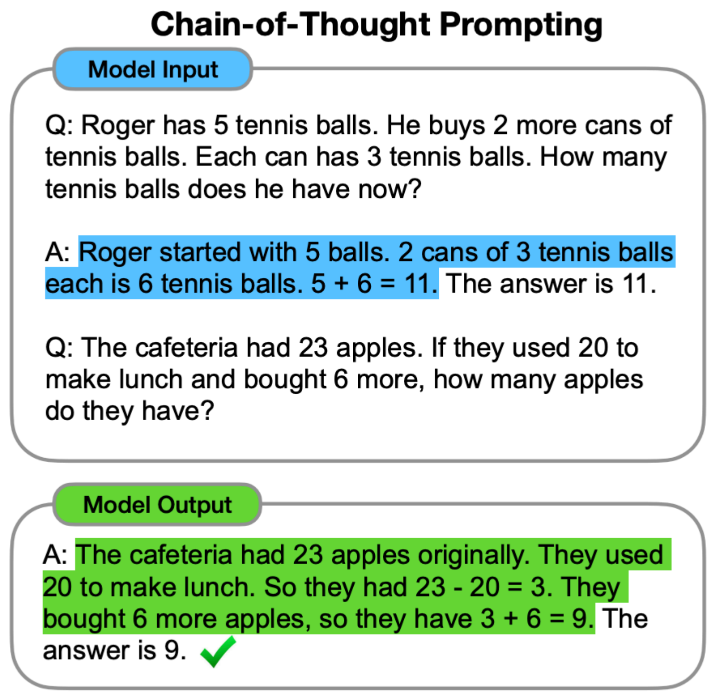
Decision Rules / Graphs

(e.g. Anchors)



Free-text Rationales

(e.g. Chain-of-thought)



Selvaraju et al., ICCV 2017; Fong et al., ICCV 2019; Kim et al., Koh & Liang ICML 2018; Koh*, Nguyen*, Tang* et al., ICML 2020; Chen* & Li* et al., NeurIPS 2019; Wang et al, CVPR 2020 , Ribeiro et al., KDD 2016; Lundberg & Lee, NeurIPS 2017; Ribeiro et al., AAAI 2018; Strobelt et al, IEEEVis 2018; Wallace et al, EMNLP, 2019; Wei et al, NeurIPS 2022.

Under-Explored: human evaluation of XAI usefulness

Selvaraju et al., ICCV 2017; Fong et al., ICCV 2019; Kim et al., Koh & Liang ICML 2018; Koh*, Nguyen*, Tang* et al., ICML 2020; Chen* & Li* et al., NeurIPS 2019; Wang et al, CVPR 2020 , Ribeiro et al., KDD 2016; Lundberg & Lee, NeurIPS 2017; Ribeiro et al., AAAI 2018; Strobelt et al, IEEEVis 2018; Wallace et al, EMNLP, 2019; Wei et al, NeurIPS 2022

Background & Motivation

Useful XAI for Humans

RQ1: Are XAI Useful for Humans?

Post-hoc Interpretability

Intrinsic Interpretability

Molnar, Christoph. Interpretable machine learning. 3.2 Taxonomy of Interpretability Methods. 2020.

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

Hua Shen

Kenneth Huang

Shen, Hua, and Ting-Hao Huang. "How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels." HCOMP. 2020.

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:

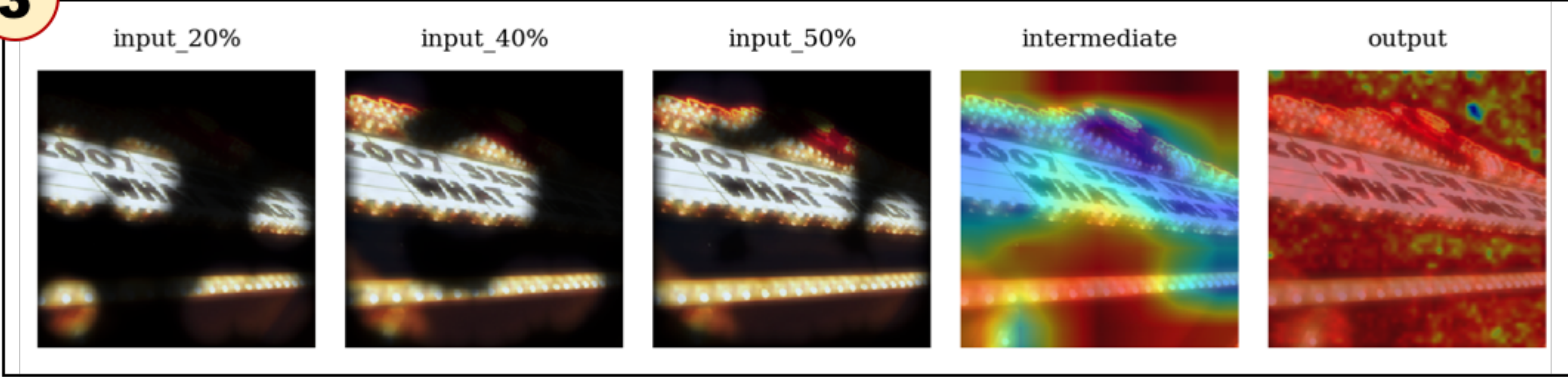
1

2 Correct Label: Movie Theater

Input Image

3

ExtremalPerturb GradCAM SmoothGrad



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 Misclasification (non-overlap users).

Conclusion

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

↓ (statistically significant)

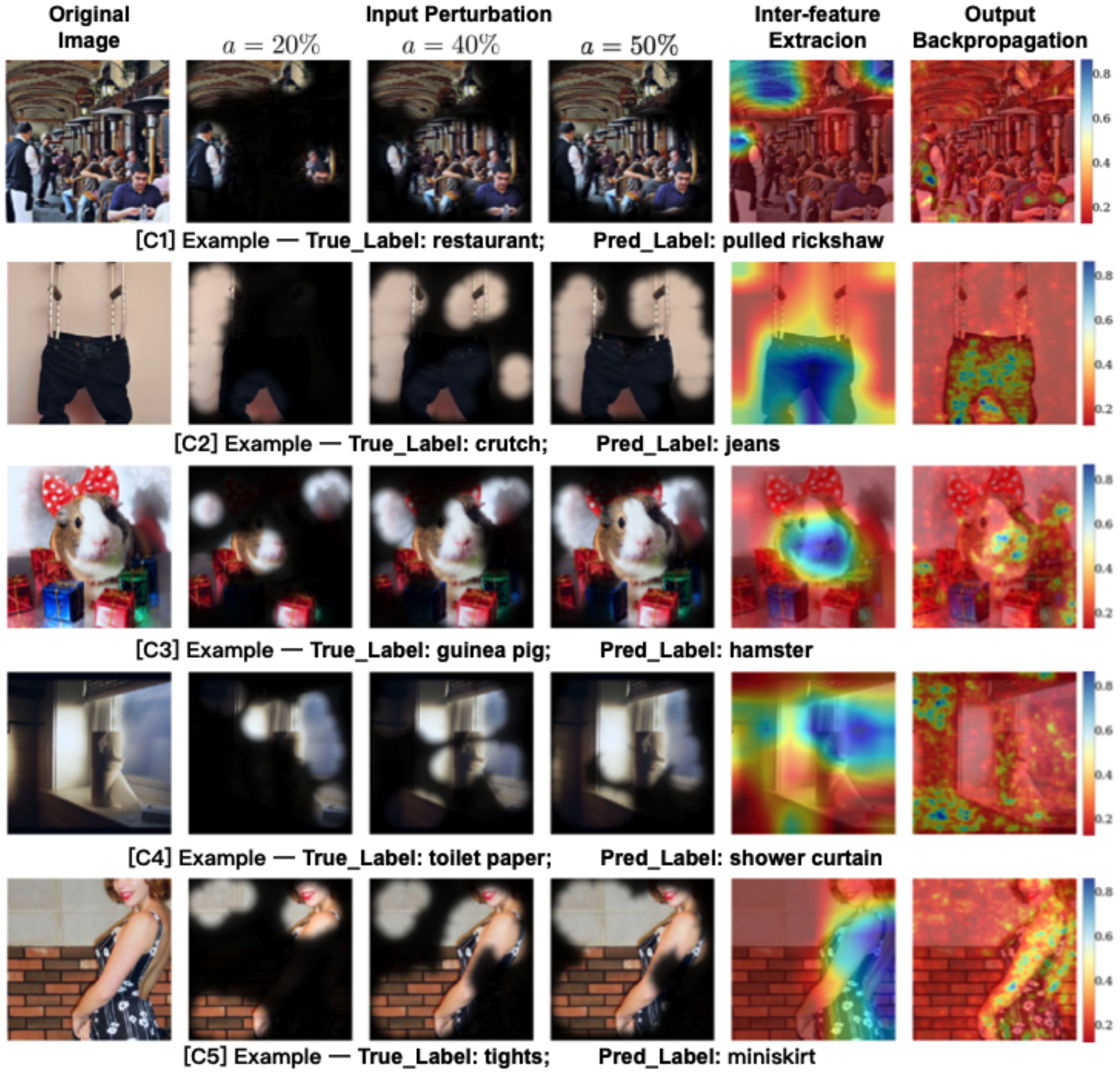
10% Accuracy Drop

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%.

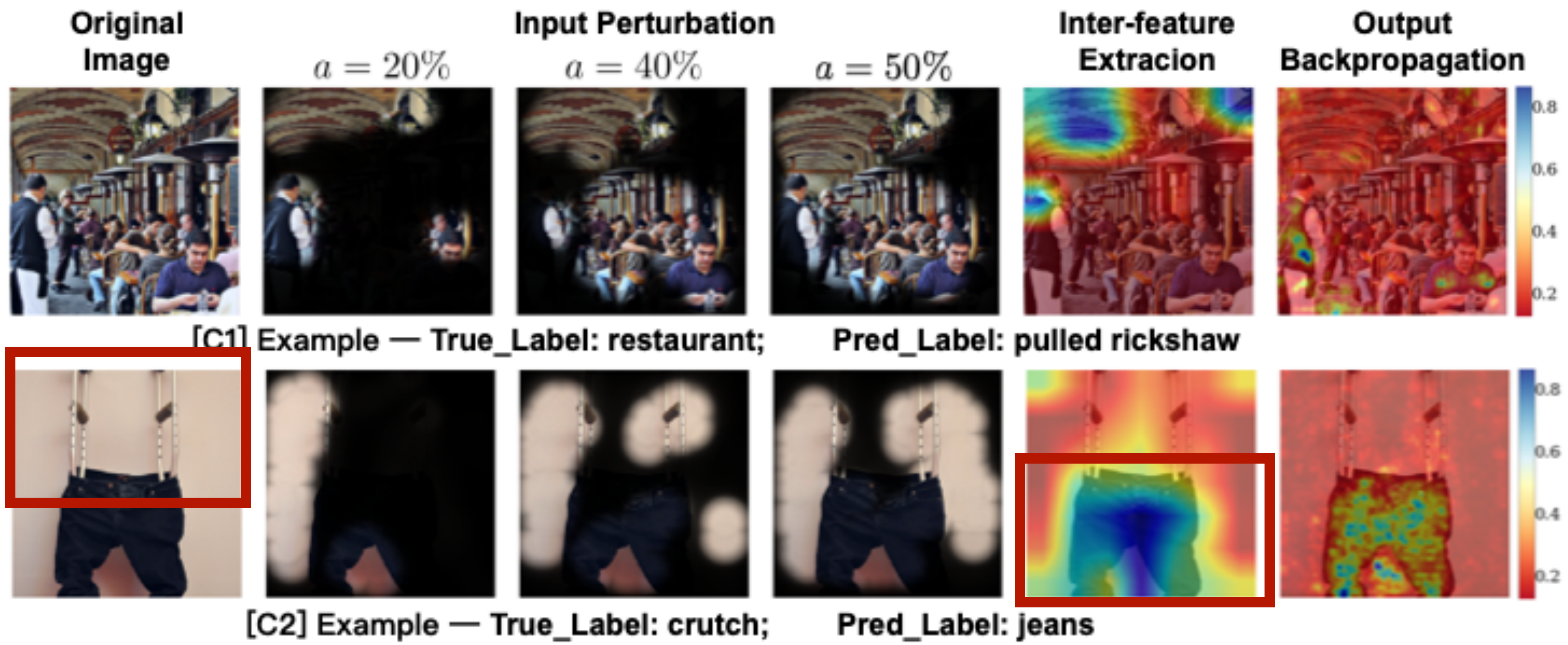
WHY?

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 Misclasification (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

Useful XAI for humans in practice

Real Humans	**Human-grounded Evaluation** (e.g., Human simulatability, Error analysis)	**Simple Tasks**
No Real Humans	**Functionally-grounded Evaluation** (e.g., Faithfulness, Robustness, Plausibility)	**Proxy Tasks**

HCOMP 2020

Humans ← AI Interpretability Evaluation → AI Tasks

Background & Motivation

Useful XAI for Humans

RQ1: Are XAI Useful for Humans?

Post-hoc Interpretability

Intrinsic Interpretability

Are Shortest Rationales the Best Explanations for Human Understanding?

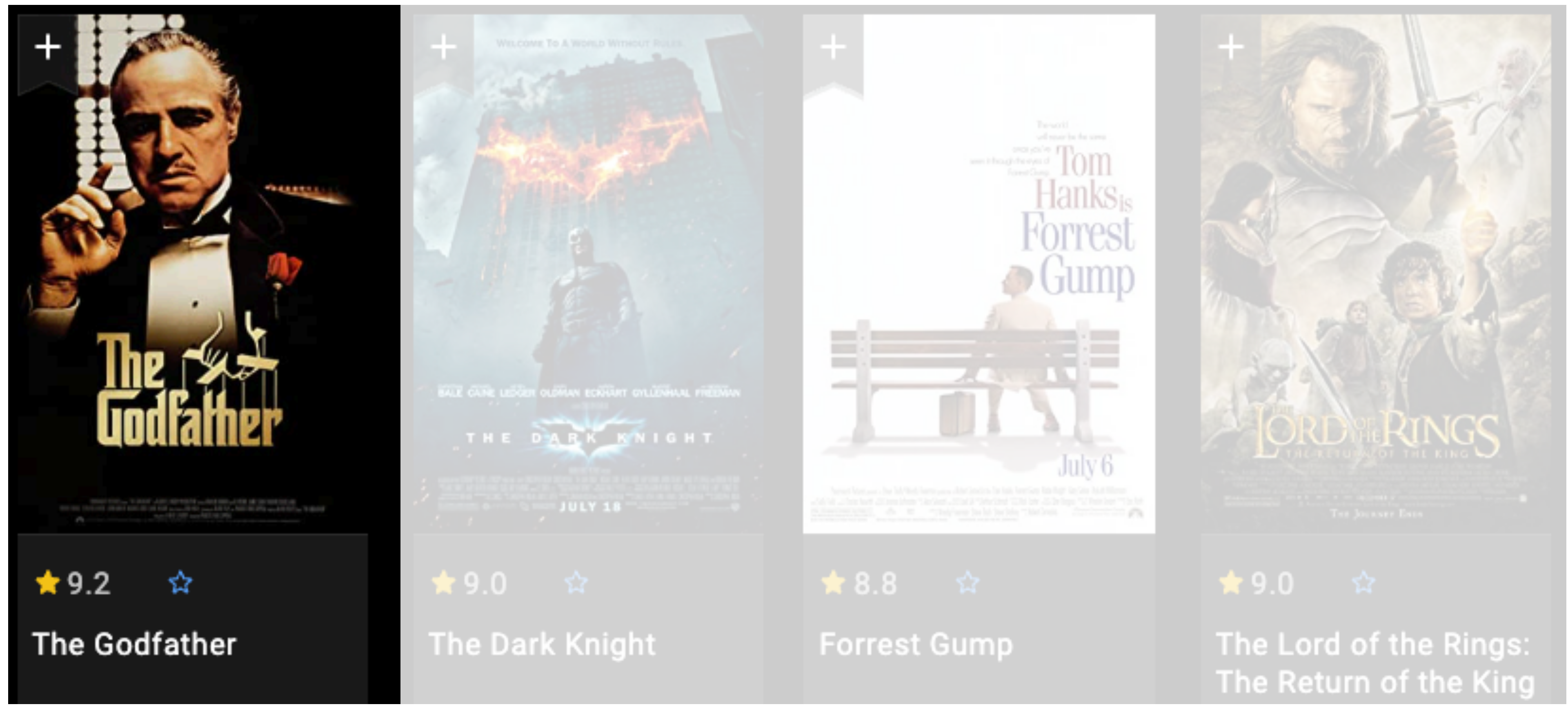


Hua Shen Sherry Wu Wenbo Guo Kenneth Huang

Shen, Hua, et al. "Are Shortest Rationales the Best Explanations for Human Understanding?." ACL. 2022.

Rationales in Text Classification Tasks

Sentiment Analysis For Movie Reviews



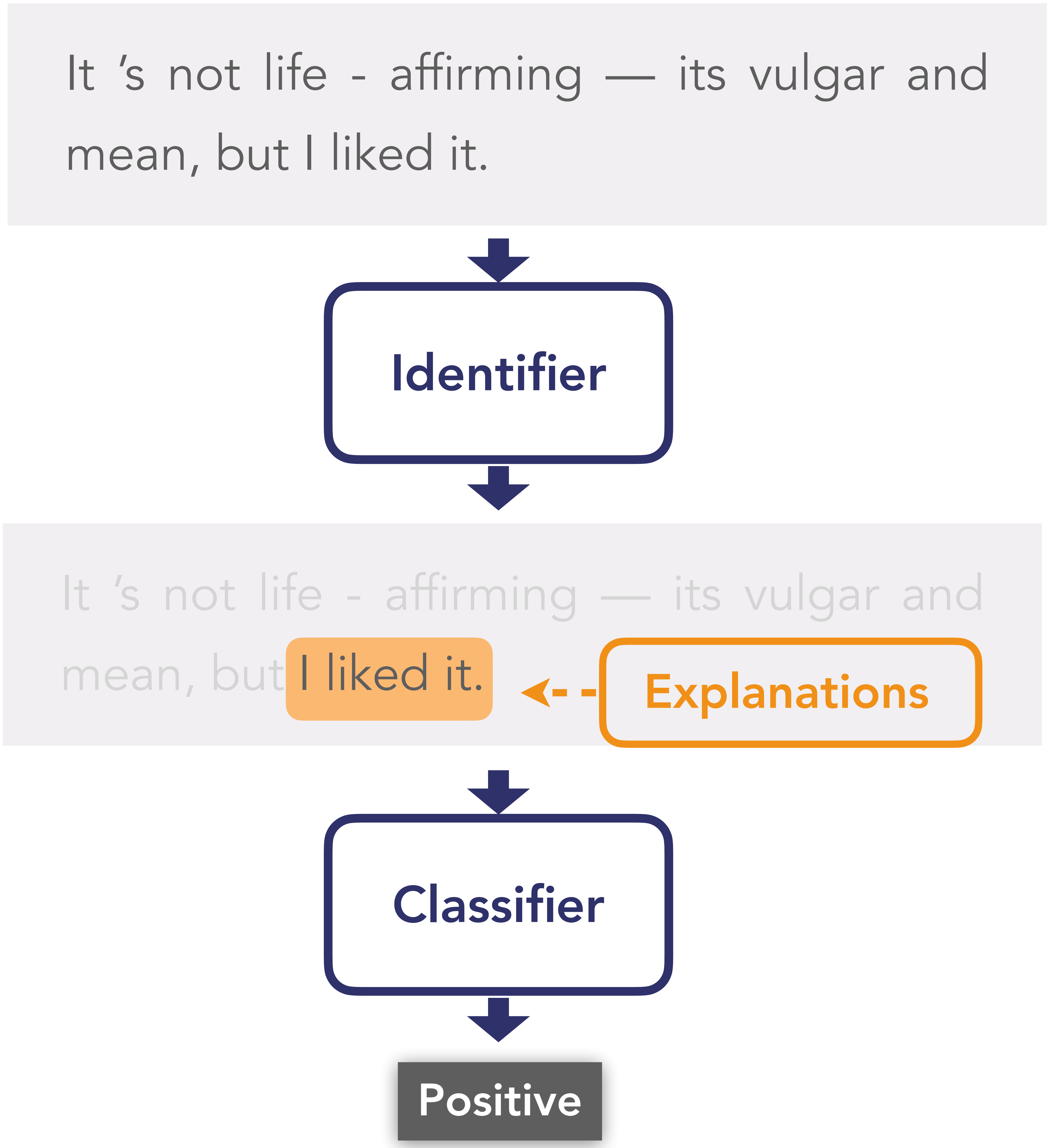
It 's not life - affirming — its vulgar and mean, but I liked it. ← Explanations

Positive Negative

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

Vafa, Keyon, et al. "Rationales for sequential predictions." EMNLP, 2021.
Bastings, Jasmijn, et al. "Interpretable neural predictions with differentiable binary variables." ACL, 2019.

Self-Explaining Models



Prior Study

Shorter Explanations are Better.

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

Lei, Tao, Regina Barzilay, and Tommi Jaakkola. "Rationalizing neural predictions." EMNLP, 2016.

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

Overview of Study Design

Goal: the impact of explanation lengths on human understanding:

Propose a **novel self-explaining model** to generate explanations with different lengths

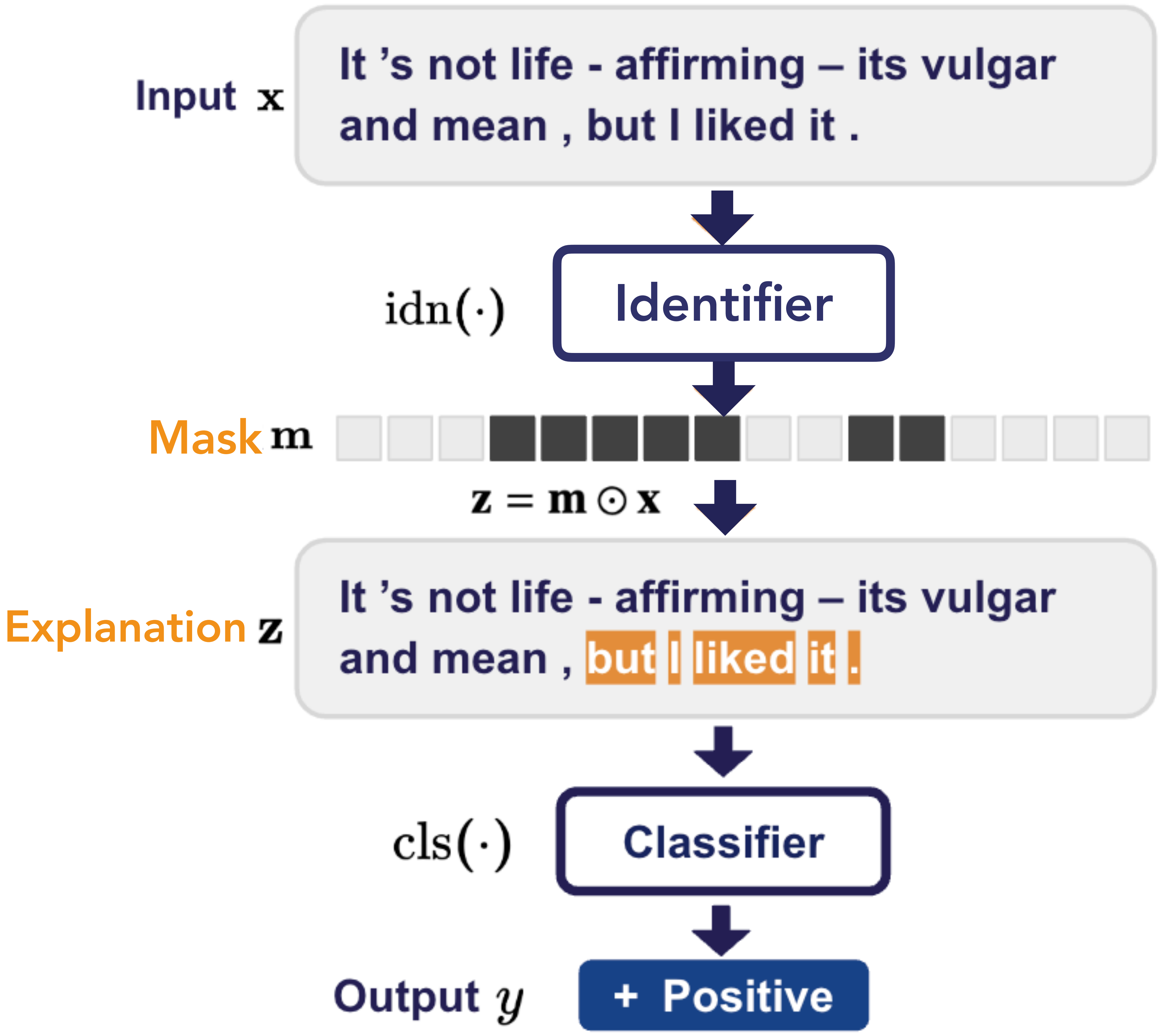
Humans guess the labels with explanations of different lengths

Contribution

A novel self-explaining model

human evaluation on intrinsic interpretability usefulness

LimitedInk model generates rationals with different length



Optimization Objective

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

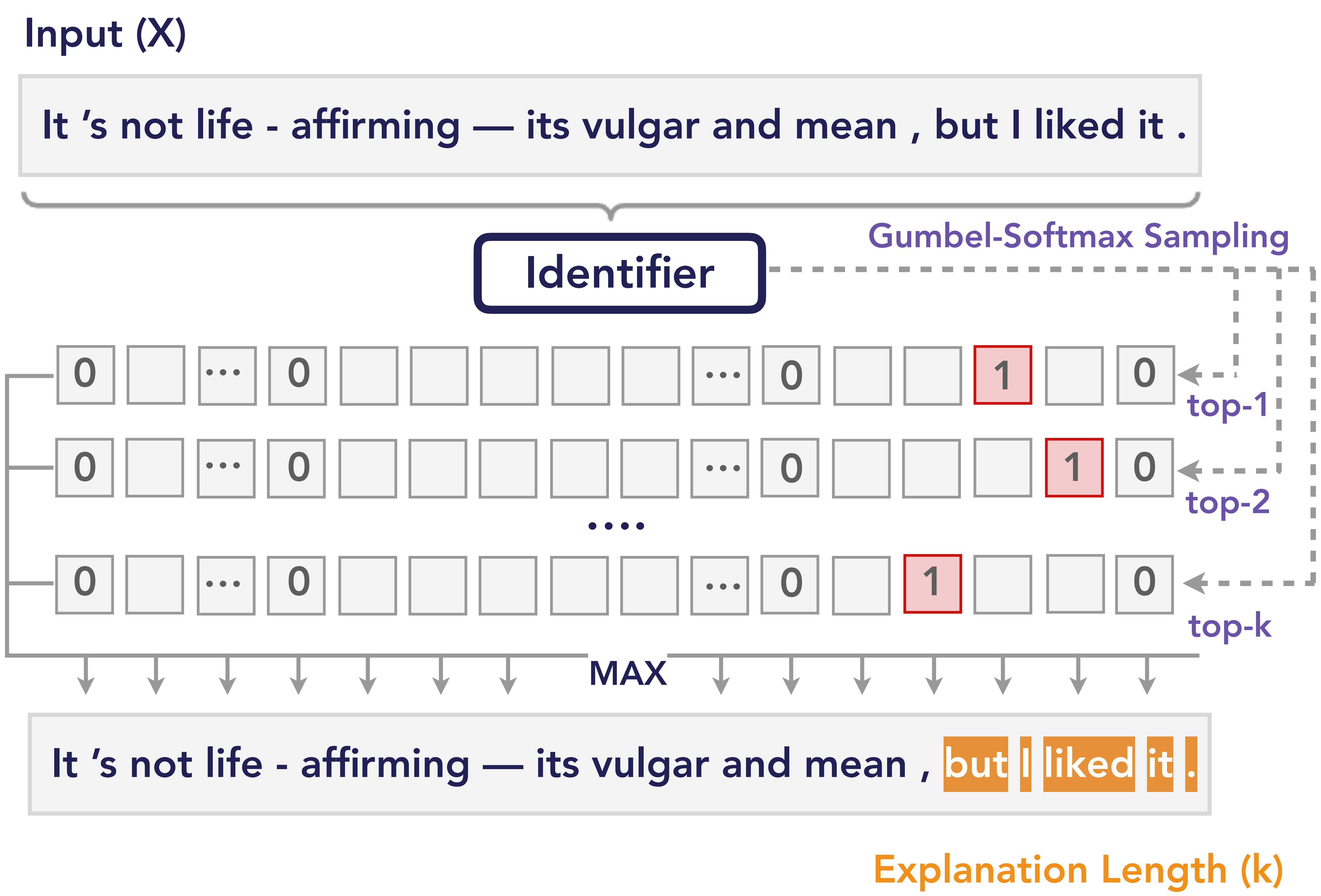
1. Gumbel-Softmax Sampling

2. Vector and Sort Regularization

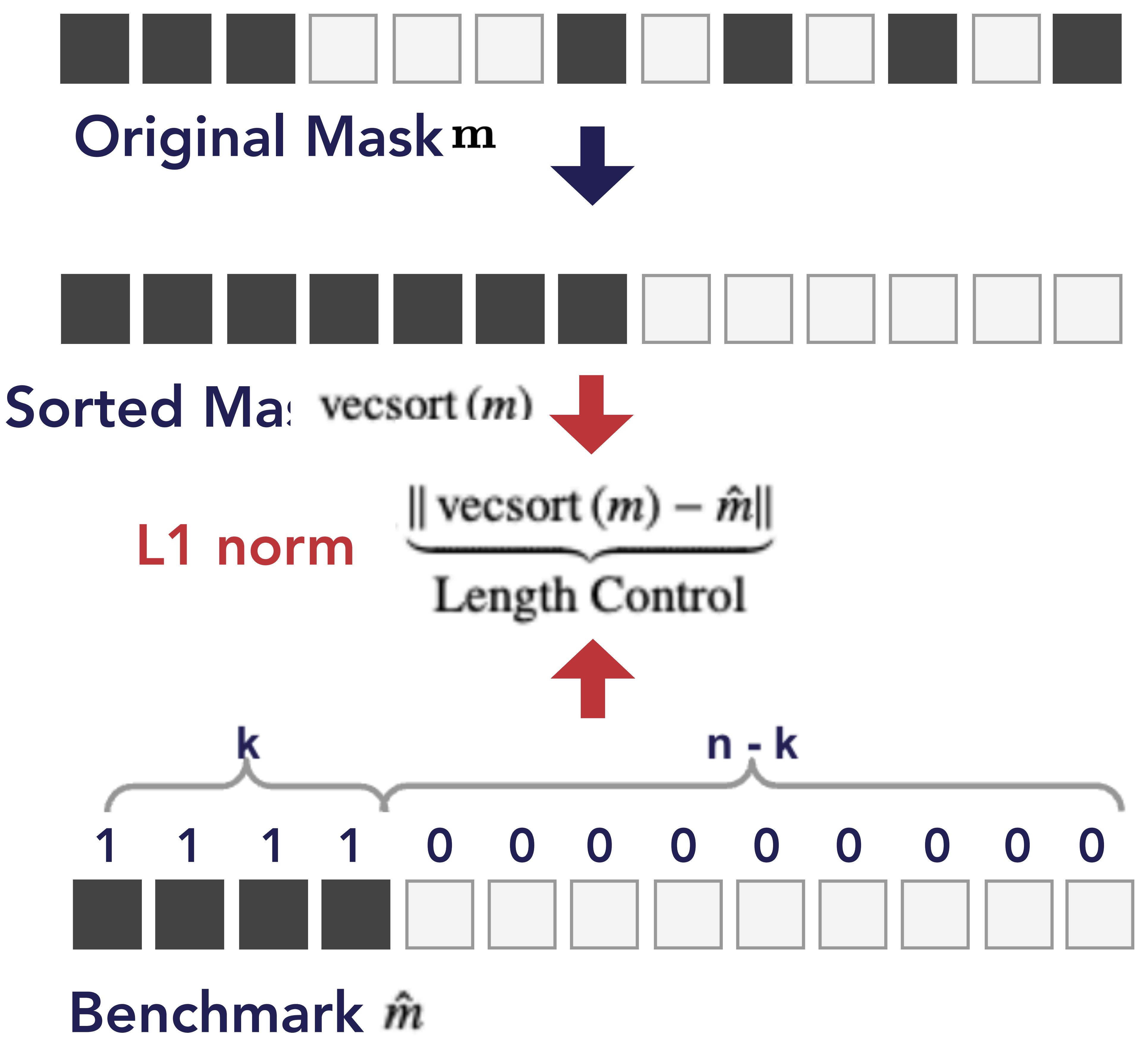
Control Different Explanation Length

How to control explanation length in LimitedInk

1. Gumbel-Softmax Sampling



2. Vector and Sort Regularization



Jang, E., Gu, S., & Poole, B. (2017, April). Categorical reparametrization with gumble-softmax. ICLR, 2017.

Can LimitedInk **perform well on classification**?

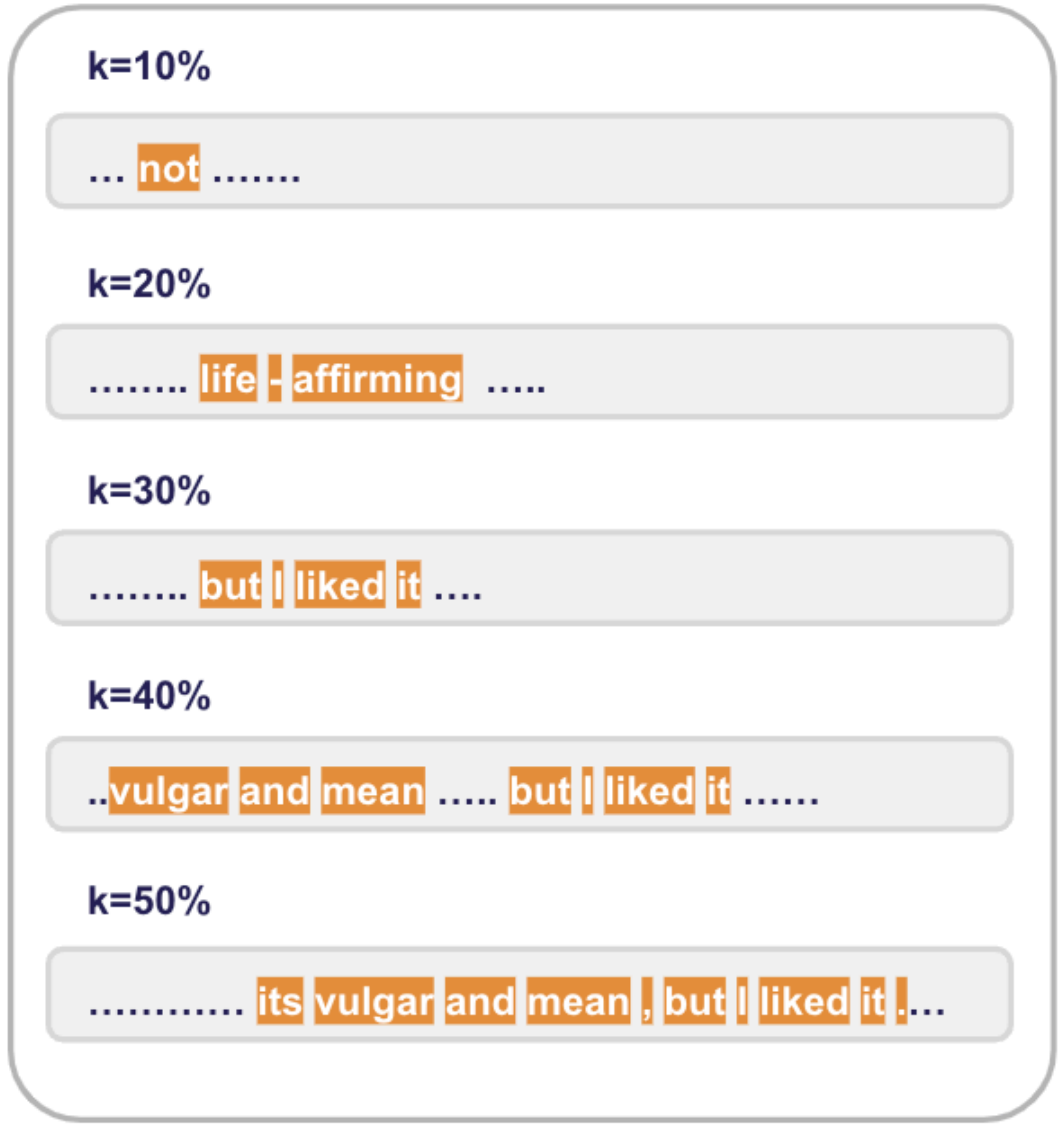
- **End-task classification: Task**, weighted average F1
- **Human Plausibility with annotated dataset: P**recision, **R**ecall, Token-level **F1**

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%			

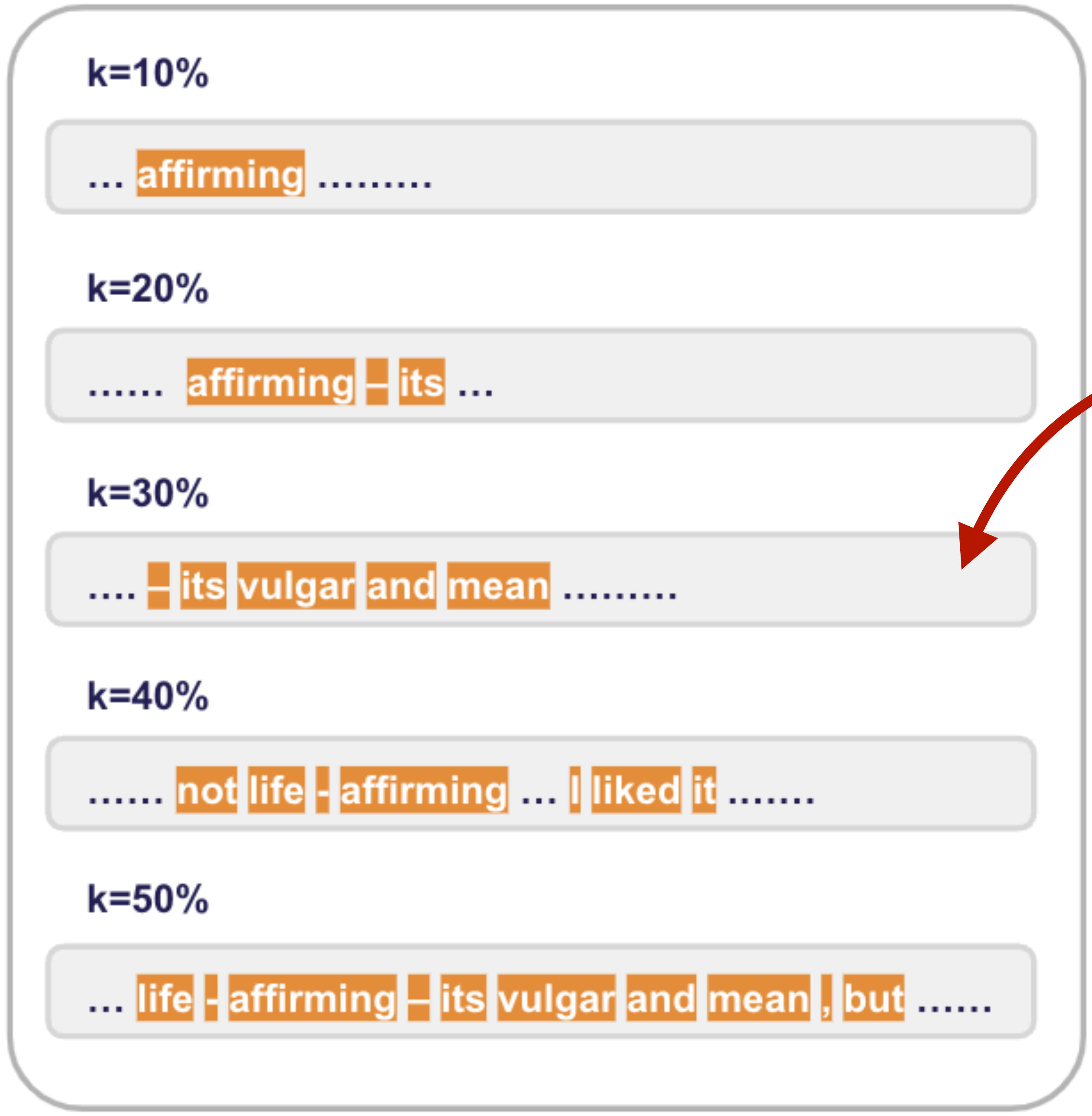
LimitedInk **performed compatible with three SOTA baselines** on the two common rationale metrics in five ERASER text classification benchmark datasets.

Step2 - Human Study Setups

LimitedInk Explanations



Random text spans (similar length)

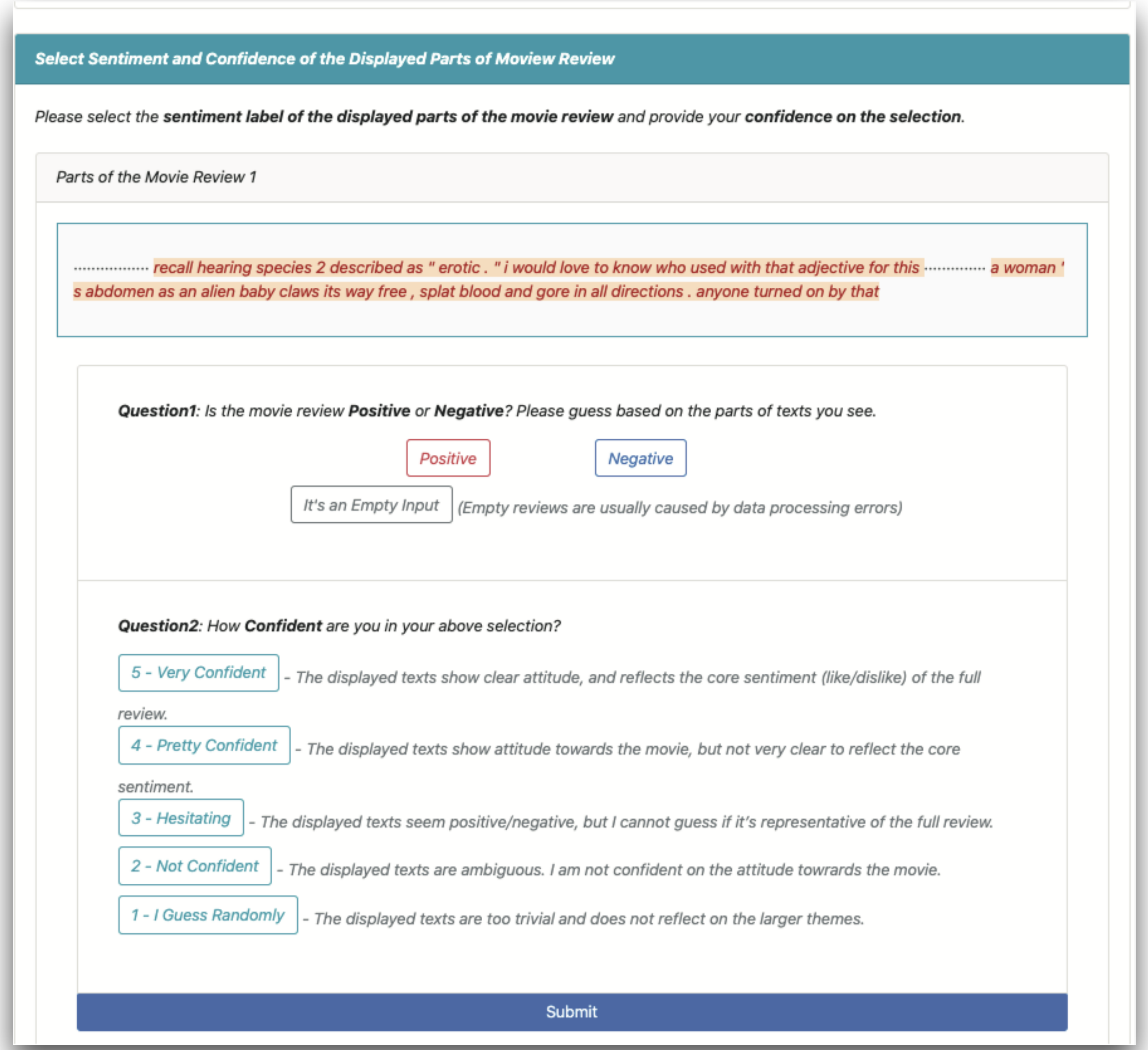


Only highlight explanations & hide other texts!

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 Moview 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.

Positive   Negative

It's an Empty Input (Empty reviews are usually caused by data processing errors)

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

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

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

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

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

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

Submit

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

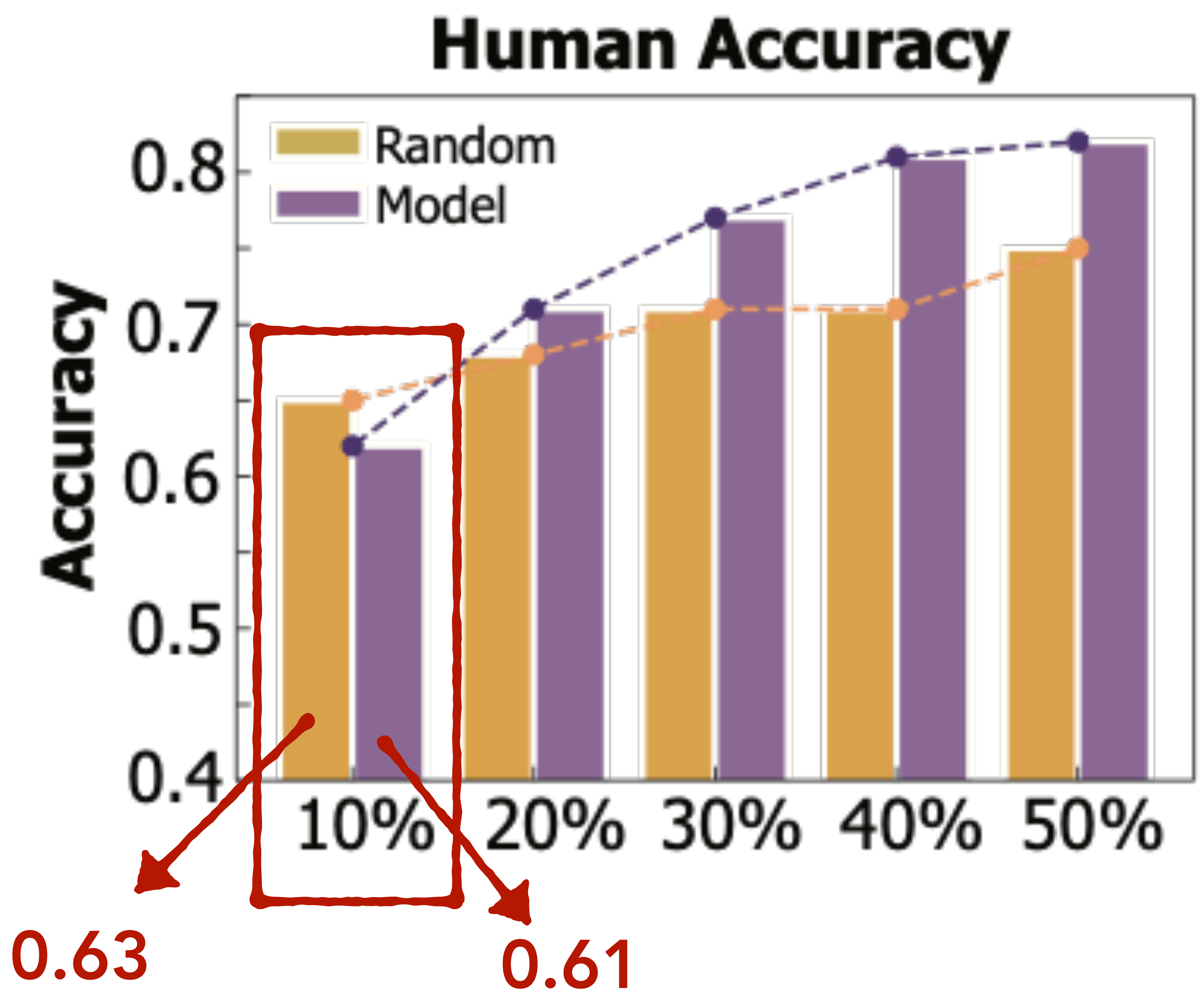
→ **prediction**

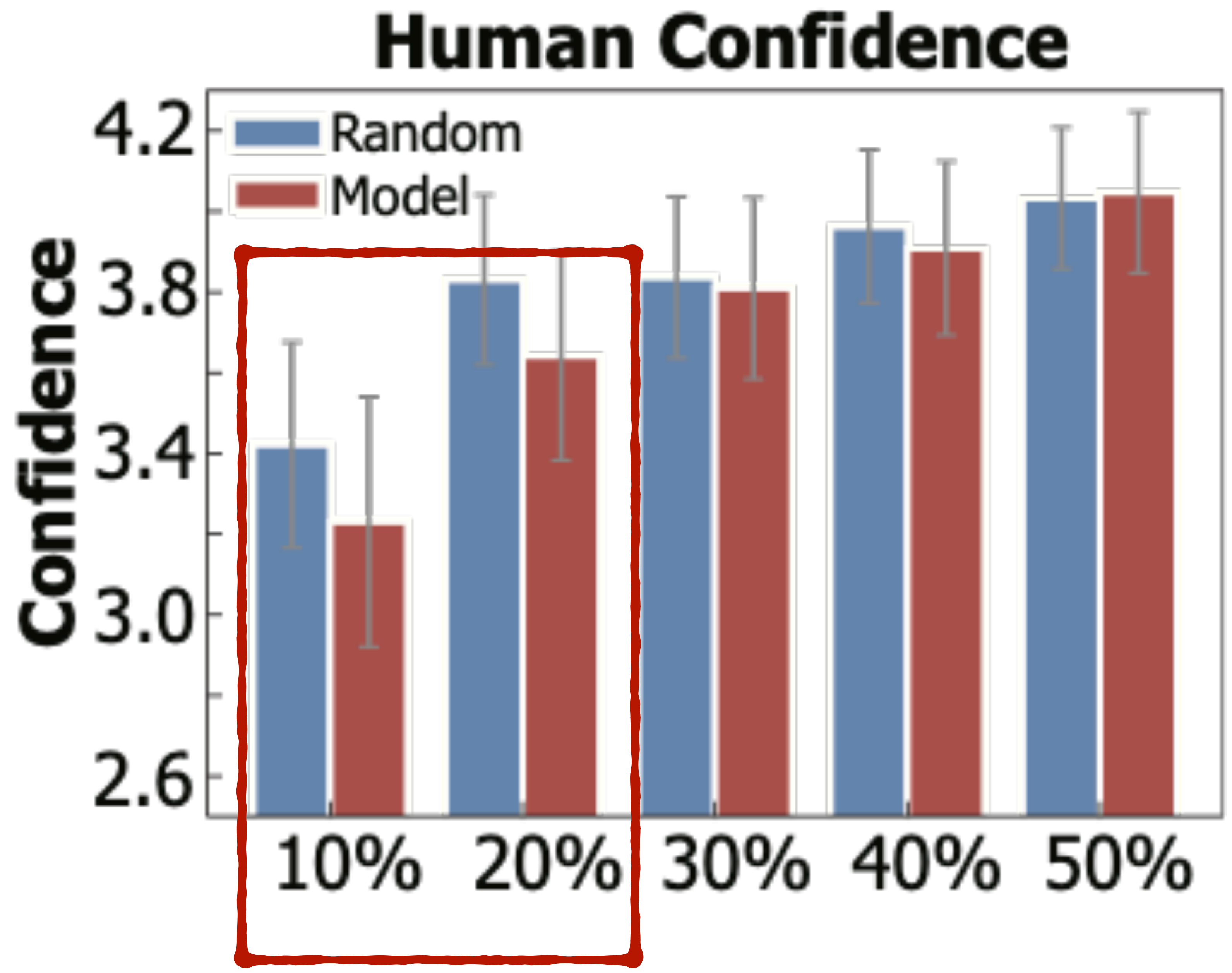
→ **confidence**

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

Useful XAI for humans in practice

Real Humans	**Human-grounded Evaluation** (e.g., Human simulatability, Error analysis)	**Simple Tasks**
No Real Humans	**Functionally-grounded Evaluation** (e.g., Faithfulness, Robustness, Plausibility)	**Proxy Tasks**

Humans ← AI Interpretability Evaluation → AI Tasks

Background & Motivation

Useful XAI for Humans

RQ1: Are XAI Useful for Humans?

RQ2: Why?

Post-hoc Interpretability

Intrinsic Interpretability

Explaining the Road Not Taken

Hua Shen

Ting-Hao 'Kenneth' Huang

Shen, Hua, and Ting-Hao Huang. "Explaining the Road Not Taken." HCXAI CHI Workshop. 2021.

What are the real-world user needs for XAI?

XAI Question Bank

43 XAI 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?
- * How much data [like this] is the system trained on?

Output

- **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) ?

Performance

- **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…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - 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?
  - * How are the parameters set?

Why

- **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?

Why not

- **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?

What If

- **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]?

How to be that

- **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?

How to still be this

- **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?

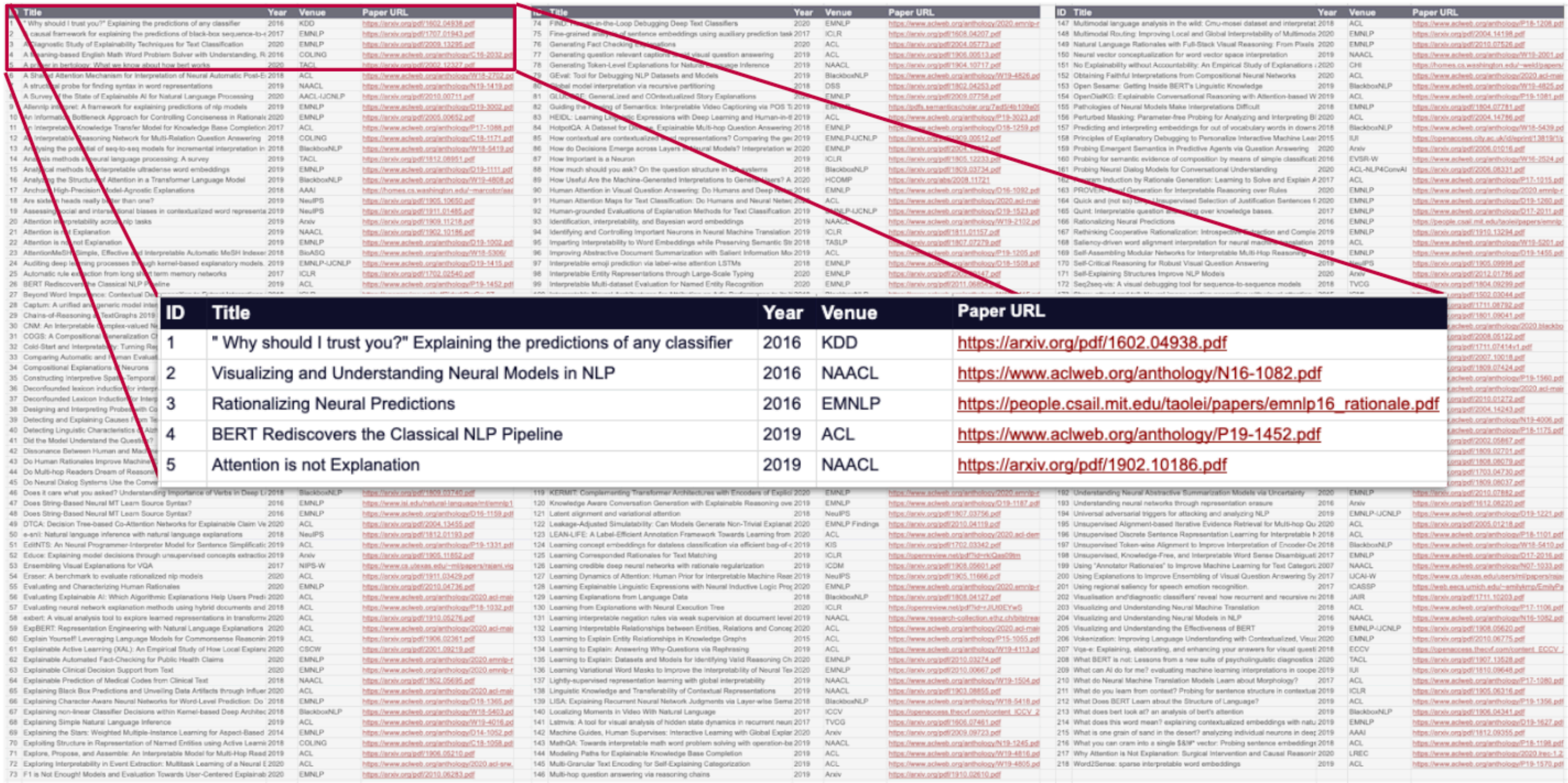
Others

- * 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)
- * What are the results of other people using the system? (social)

Liao, Q. Vera, Daniel Gruen, and Sarah Miller. "Questioning the AI: informing design practices for explainable AI user experiences." CHI. 2020.

*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 Rediscovers 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

The website listing 200+ XAI paper: https://human-centered-exnlp.github.io/

Matching XAI Papers with XAI Question Bank?

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)

?

218 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/antholog
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
4	BERT Rediscovers 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 Processin	2020	AACL-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 Uncertaint	2020	EMNLP	https://arxiv.org/pdf/2010.07882.

Manually Matching: 218 * 43 = 9,374 …

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/antholog
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
4	BERT Rediscovers 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 Processin	2020	AACL-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 Uncertaint	2020	EMNLP	https://arxiv.org/pdf/2010.07882.

1 Feature Attribution (FAT) 43.93%
2 Tuple / Graph (TUP) 10.15%
3 Concept / Sense (CPT) 9.72%
4 Rule / Grammar (RUL) 9.61%
5 Probing (PRB) 7.79%
6 Free Text (FRT) 7.09%
7 Example (EXP) 3.86%
8 Projection Space (PSP) 3.82%
9 Confidence Score (CFD) 1.18%
10 Word Cloud (WCL) 1.16%
11 Trigger (TRG) 0.93%
12 Images (IMG) 0.70%

XAI Question Bank

How (global) | Q19 - How does the system make predictions? ➡ 23.62%

10.15% (2 Tuple / Graph (TUP)) + 9.61% (4 Rule / Grammar (RUL)) + 3.86% (7 Example (EXP)) = 23.62%

Results: Heatmap for XAI Question Bank

0.0% 100.0%

Category	Question	Technique	Coverage
Input/Data (0.55%)	1-What kind of data does the system learn from?	EXP	3.86%
	2-What is the source of the data?		★
	3-How were the labels/ground-truth produced?		★
	4-What is the sample size?		★
	5-What data is the system NOT using?		●
	6-What are the limitations/biases of the data?		●
	7-How much data [like this] is the system trained on?		★
Output (0.77%)	8-What kind of output does the system give?	EXP	3.86%
	9-What does the system output mean?		★
	10-How can I best utilize the output of the system?		●
	11-What is the scope of the system's capability?		●
	12-How's the output used for other systems modules?		●
Performance (2.03%)	13-How accurate/precise/reliable are the predictions?	CFD	1.18%
	14-How often does the system make mistakes?		★
	15-In what situations is the system to be incorrect?	CFD/EXP/TRG	5.97%
	16-What are the limitations of the system?		●
	17-What kind of mistake is the system likely to make?	EXP	5.05%
	18-Is the system's performance good enough for…?		●
How (Global) (30.31%)	19-How does the system make predictions?	TUP/RUL/EXP	23.63%
	20-What features does the system consider?	FAT	43.99%
	21-What is the system's overall logic?	RUL/FAT	53.60%
	22-What kind of algorithm is used?		★

Category	Question	Technique	Coverage
Why / Why not (45.14%)	23-Why/how is this instance given this prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	24-What instance feature leads to the system's prediction?	FAT	43.99%
	25-Why are [instance A and B] given the same prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	26-Why/how is this instance NOT predicted?	TRG	0.93%
	27-Why is the instance predicted P instead of Q?	TRG	0.93%
	28-Why are [instance A and B] given different predictions?	TRG/RUL/TUP/FAT/FRT/EXP	75.62%
What if / How to be (15.54%)	29-What would the system predict if this instance changes to ..?	CFD/EXP/TRG	5.97%
	30-What would system predict if this instance feature changes to..?	CFD/FAT/TRG	46.10%
	31-What would the system predict for [a different instance]?	CFD/TRG	2.11%
	32-How should this instance change to get a different prediction?	TRG	0.93%
	33-How should instance feature change to get different prediction?	TRG	0.93%
	34-What kind of instance gets a different prediction?	TRG/EXP	4.79%
	35-What's the scope of change permitted to get the same prediction?	TRG	0.93%
	36-What's the highest feature can have to get the same prediction?	TRG/FAT	44.91%
	37-What is necessary feature present to guarantee this prediction?	TRG/FAT	44.91%
	38-What kind of instance gets this prediction?	EXP	3.86%
Others (11.49%)	39-How/what/why will the system change/improve/drift over time?		●
	40-How to improve the system?		●
	41-Why using or not using this feature/rule/data?	FAT/RUL/EXP	57.46%
	42-What does [ML terminology] mean?		★
	43-What are the results of other people using the system?		●

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

Results: Heatmap for XAI Question Bank

0.0% — 100.0%

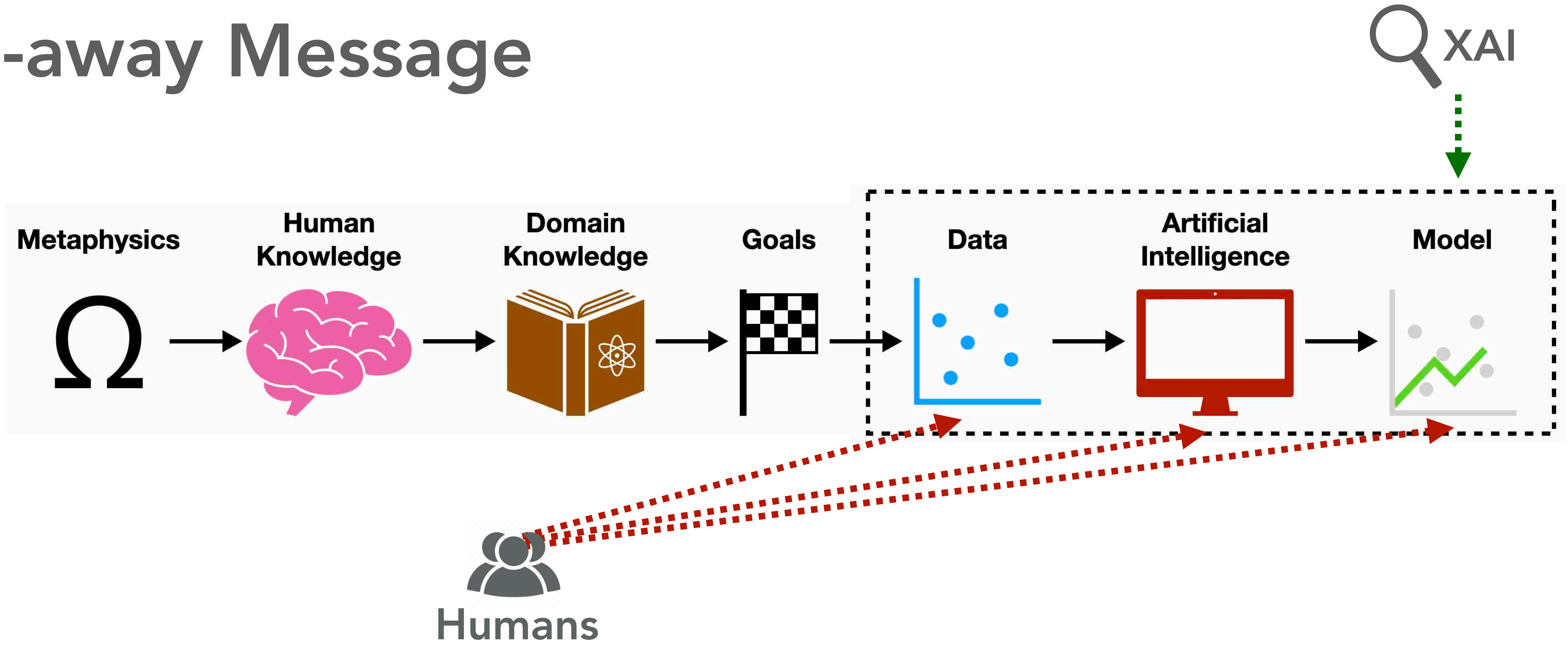
Input/Data (0.55%)	1-What kind of data does the system learn from?	EXP	3.86%
	2-What is the source of the data?		★
	3-How were the labels/ground-truth produced?		★
	4-What is the sample size?		★
	5-What data is the system NOT using?		●
	6-What are the limitations/biases of the data?		●
	7-How much data [like this] is the system trained on?		★
Output (0.77%)	8-What kind of output does the system give?	EXP	3.86%
	9-What does the system output mean?		★
	10-How can I best utilize the output of the system?		●
	11-What is the scope of the system's capability?		●
	12-How's the output used for other systems modules?		●
Performance (2.03%)	13-How accurate/precise/reliable are the predictions?	CFD	1.18%
	14-How often does the system make mistakes?		★
	15-In what situations is the system to be incorrect?	CFD/EXP/TRG	5.97%
	16-What are the limitations of the system?		●
	17-What kind of mistake is the system likely to make?	EXP	5.05%
	18-Is the system's performance good enough for…?		●
How (Global) (30.31%)	19-How does the system make predictions?	TUP/RUL/EXP	23.63%
	20-What features does the system consider?	FAT	43.99%
	21-What is the system's overall logic?	RUL/FAT	53.60%
	22-What kind of algorithm is used?		★

Why / Why not (45.14%)	23-Why/how is this instance given this prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	24-What instance feature leads to the system's prediction?	FAT	43.99%
	25-Why are [instance A and B] given the same prediction?	RUL/TUP/FAT/FRT/EXP	74.70%
	26-Why/how is this instance NOT predicted?	TRG	0.93%
	27-Why is the instance predicted P instead of Q?	TRG	0.93%
	28-Why are [instance A and B] given different predictions?	TRG/RUL/TUP/FAT/FRT/EXP	75.62%
What if / How to be (15.54%)	29-What would the system predict if this instance changes to ..?	CFD/EXP/TRG	5.97%
	30-What would system predict if this instance feature changes to..?	CFD/FAT/TRG	46.10%
	31-What would the system predict for [a different instance]?	CFD/TRG	2.11%
	32-How should this instance change to get a different prediction?	TRG	0.93%
	33-How should instance feature change to get different prediction?	TRG	0.93%
	34-What kind of instance gets a different prediction?	TRG/EXP	4.79%
	35-What's the scope of change permitted to get the same prediction?	TRG	0.93%
	36-What's the highest feature can have to get the same prediction?	TRG/FAT	44.91%
	37-What is necessary feature present to guarantee this prediction?	TRG/FAT	44.91%
	38-What kind of instance gets this prediction?	EXP	3.86%
Others (11.49%)	39-How/what/why will the system change/improve/drift over time?		●
	40-How to improve the system?		●
	41-Why using or not using this feature/rule/data?	FAT/RUL/EXP	57.46%
	42-What does [ML terminology] mean?		★
	43-What are the results of other people using the system?		●

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

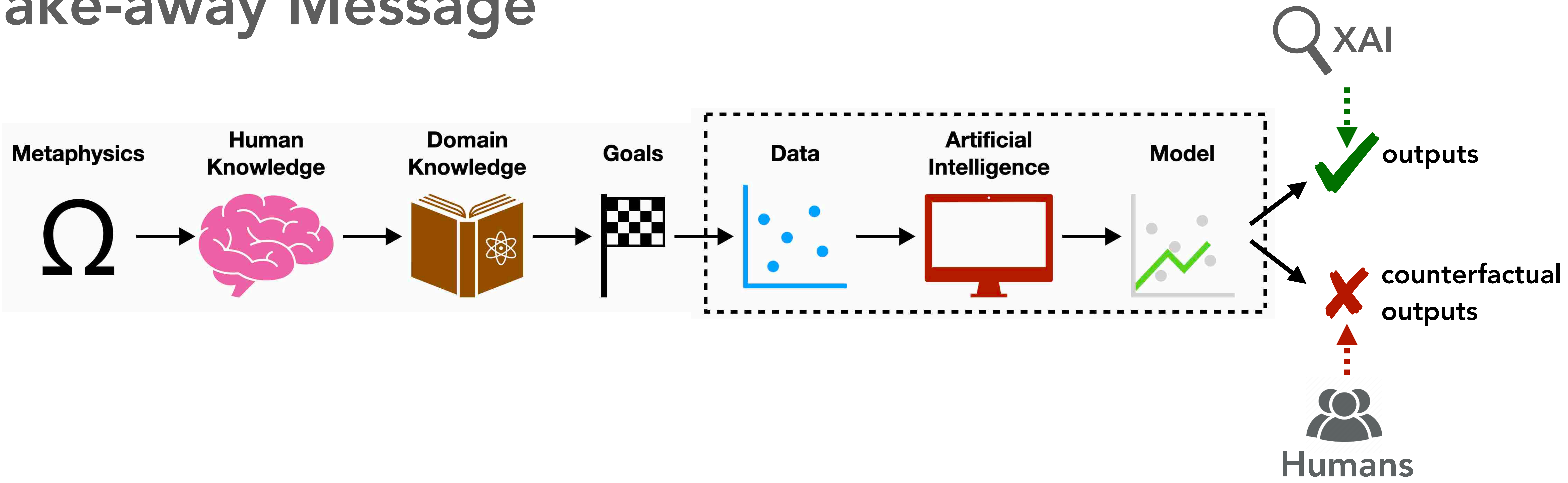
➔ XAI largely **ignored**: **diverse information** across AI development process (data, model, deployment, etc.)

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

RQ1: Are XAI Useful for Humans?

Post-hoc Interpretability

Intrinsic Interpretability

RQ2: Why?

RQ3: How to Improve?

ConvXAI:
Interactive XAI via Conversations

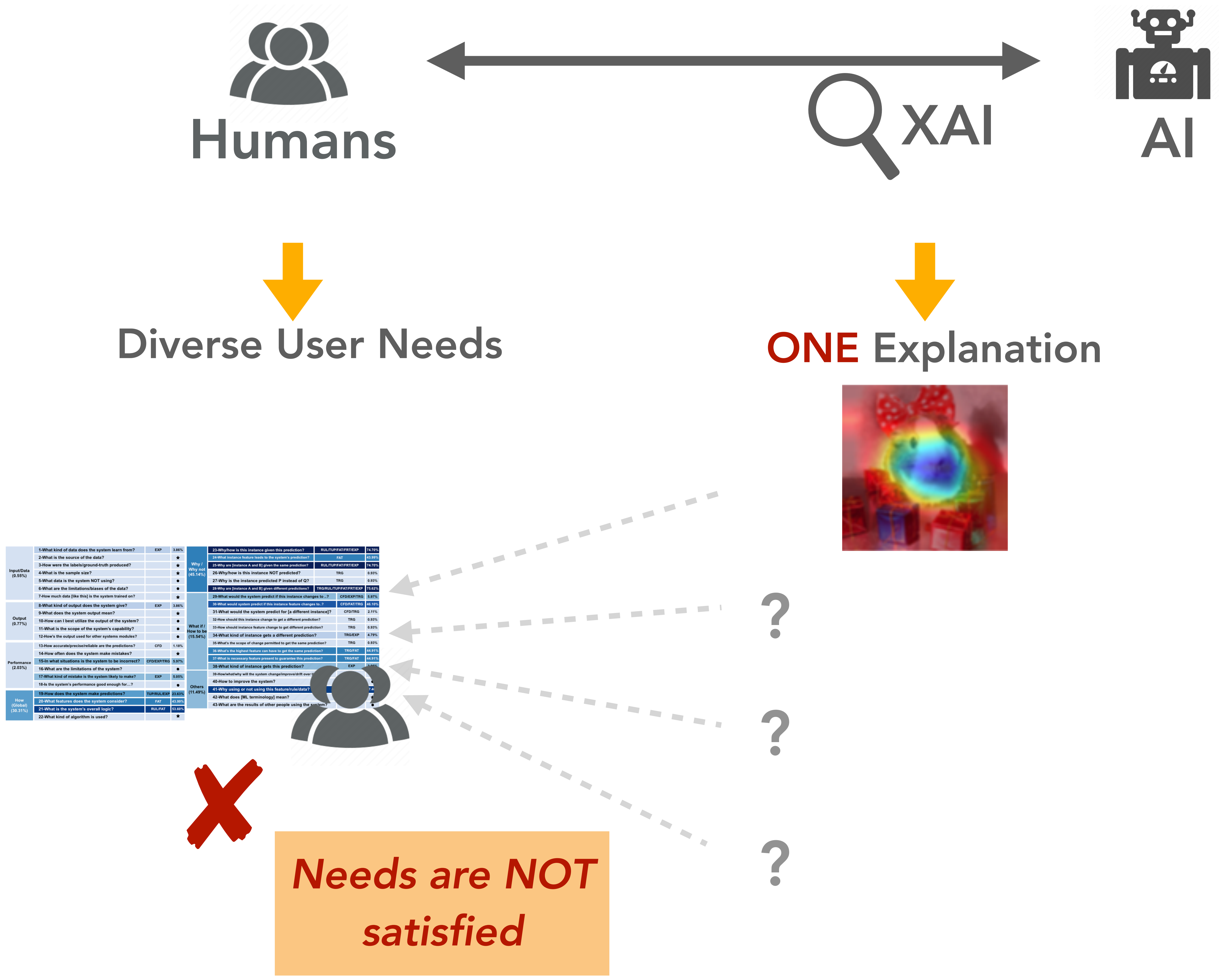
Useful XAI for Humans

CSCW 2023
Demo

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

Hua Shen Chieh-Yang Huang Sherry Wu Kenneth Huang

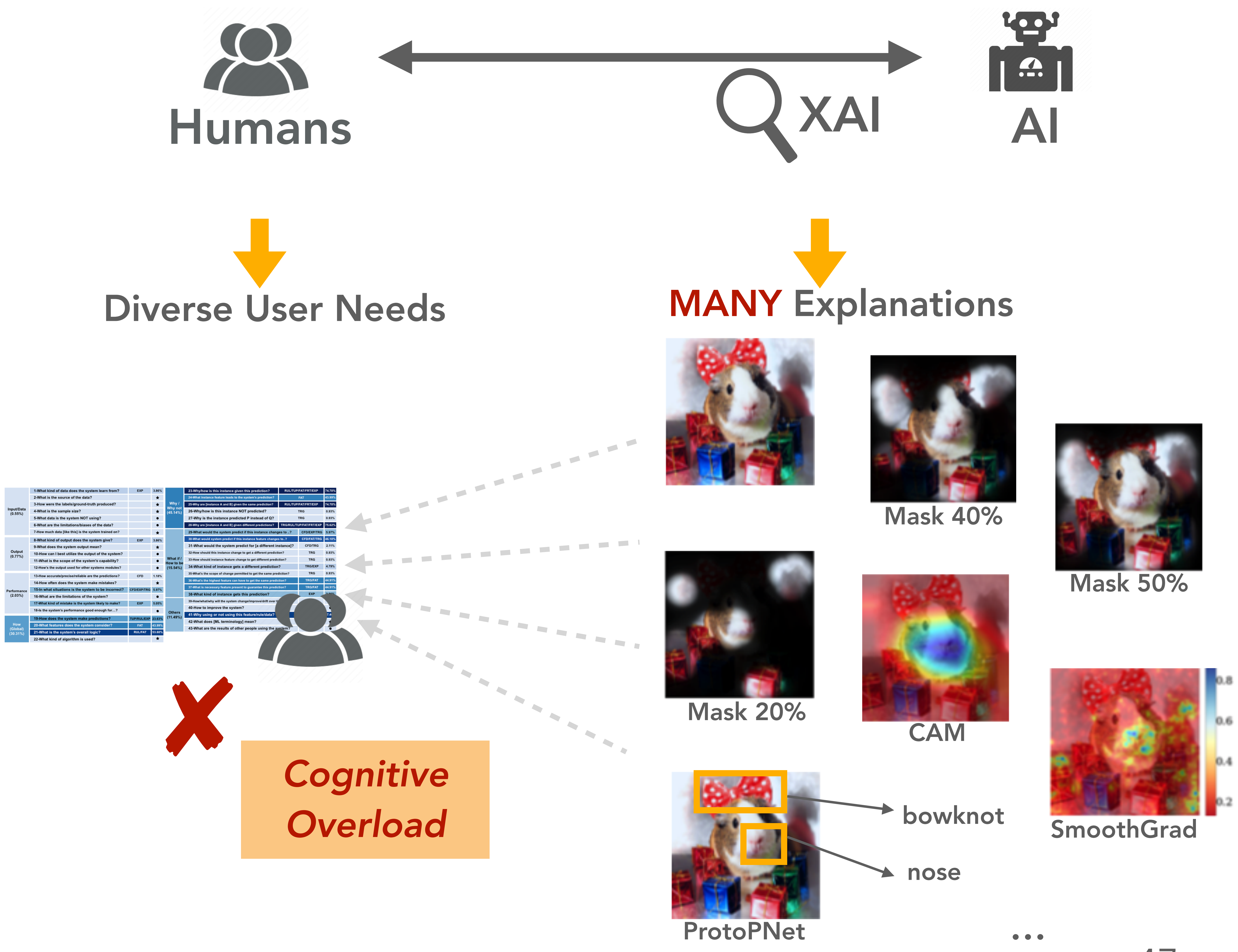
Challenges of Existing XAI



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

Shen, Hua, and Ting-Hao Huang. "How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels." HCOMP. 2020.

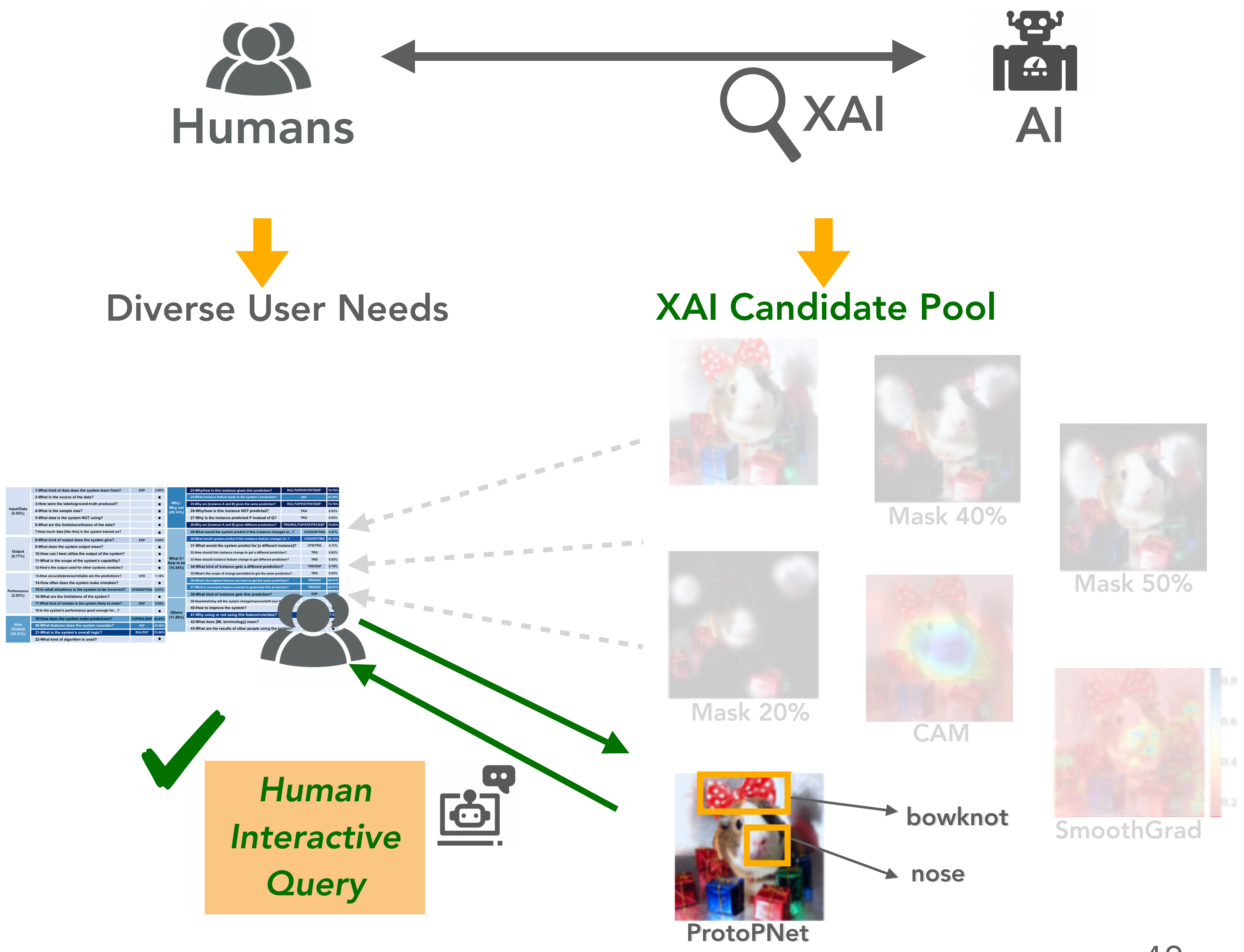
Challenges of Existing XAI



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

Poursabzi-Sangdeh, Forough, et al. "Manipulating and measuring model interpretability." CHI. 2021.

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?

Useful XAI for humans in practice

Real Humans	**Application-grounded Evaluation** (e.g., Collaboration usefulness, Performance gain)	Real Tasks
Real Humans	**Human-grounded Evaluation** (e.g., Human simulatability, Error analysis)	Simple Tasks
No Real Humans	**Functionally-grounded Evaluation** (e.g., Faithfulness, Robustness, Plausibility)	Proxy Tasks

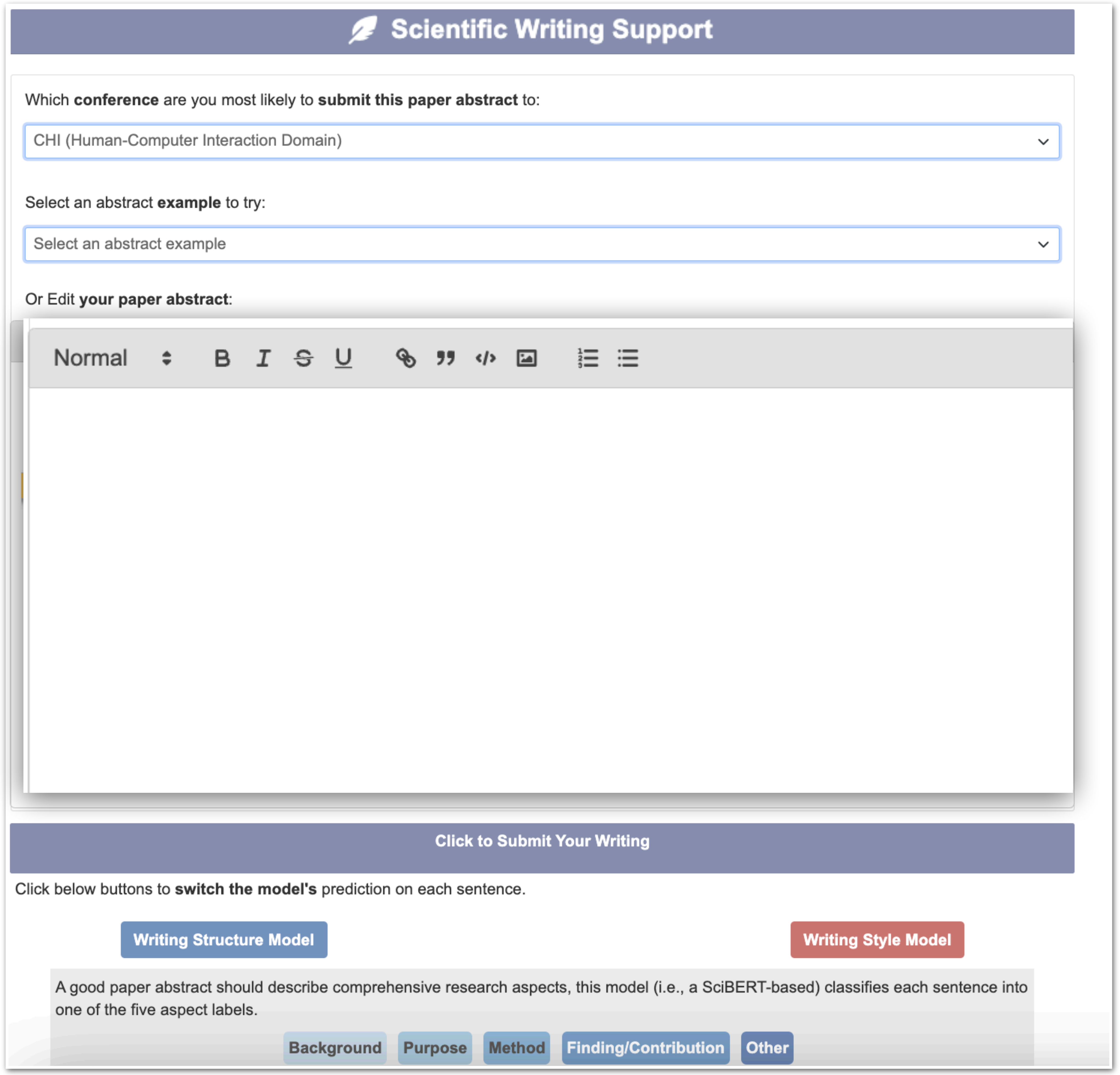
?

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

HCOMP 2020

Humans ← AI Interpretability Evaluation → AI Tasks

Real Task: AI-assisted scientific writing by humans



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

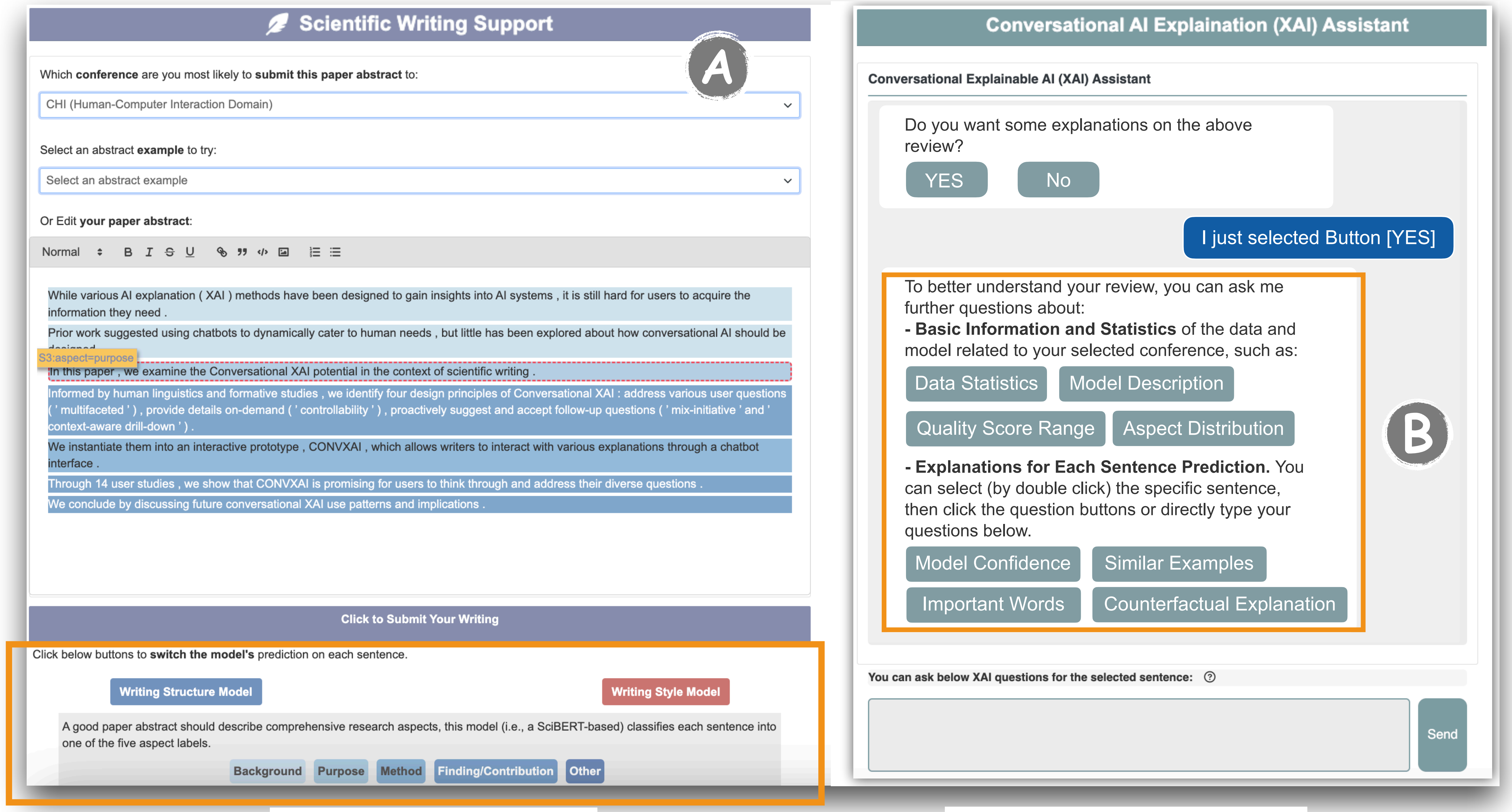
Participants

- 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”

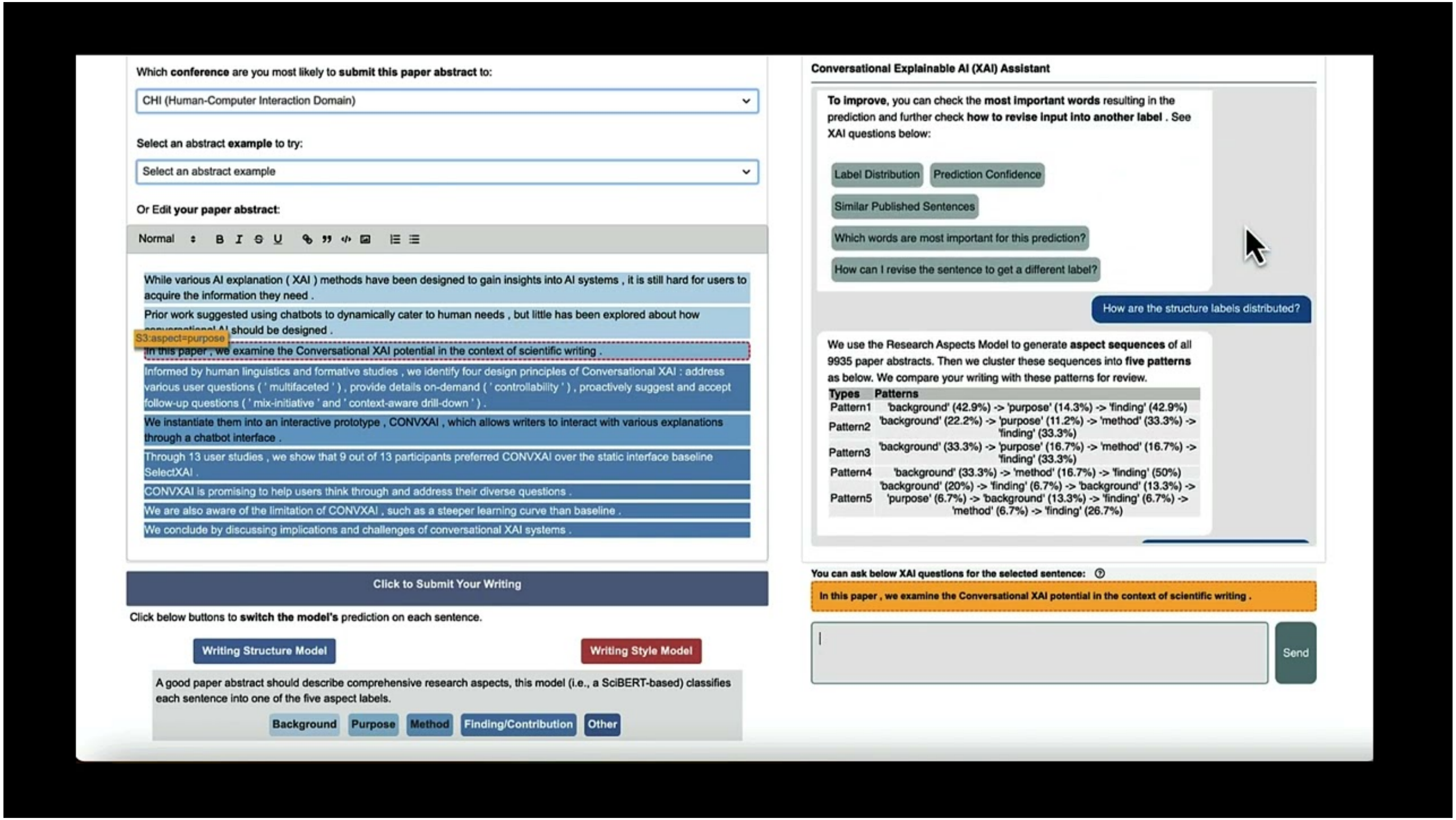
P3 Context-aware Drill-down

Maintain the conversation history to generate responses with user needs

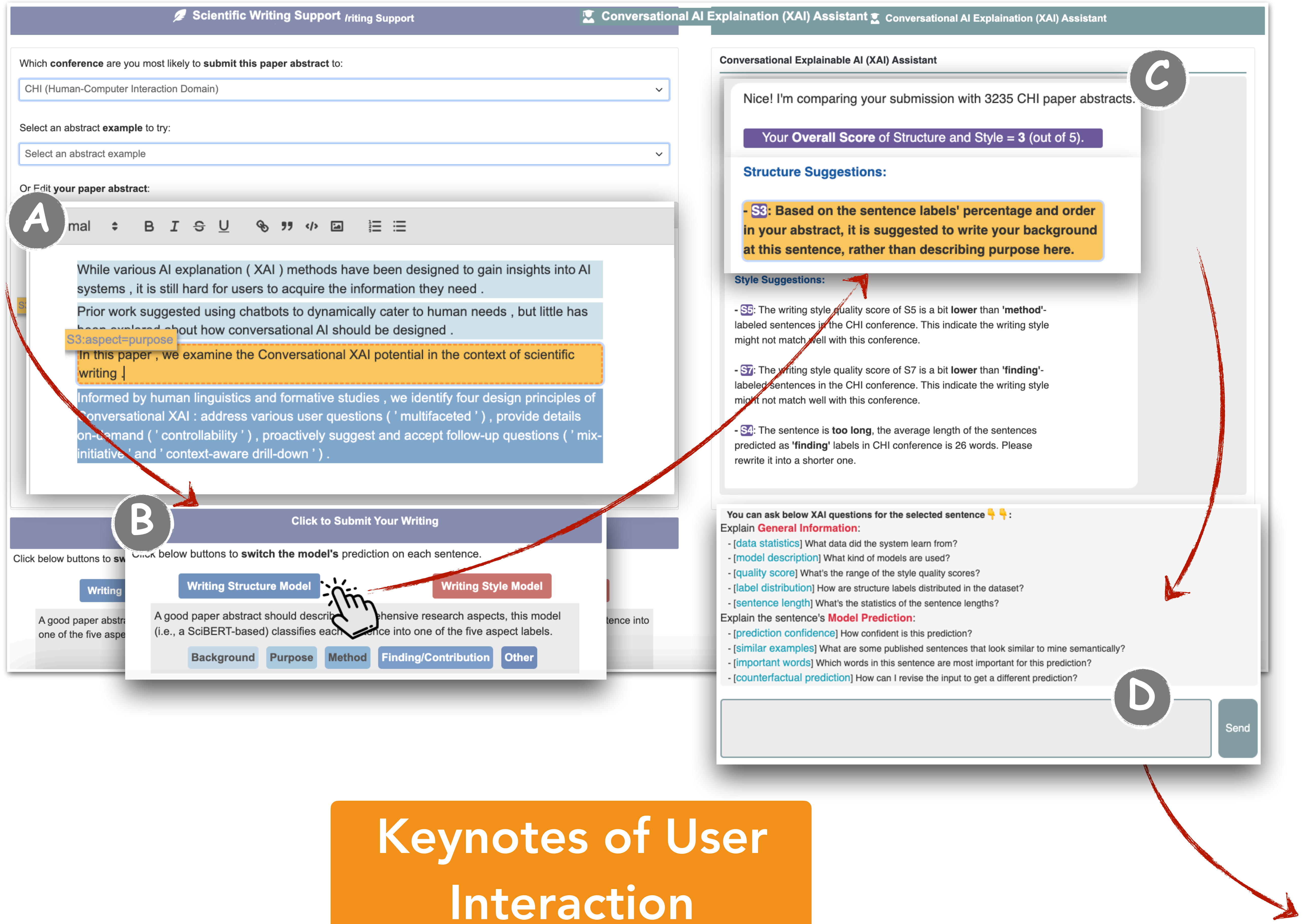
P4 Controllability

Enable humans to customize XAI with personalized needs

ConvXAI Demo:

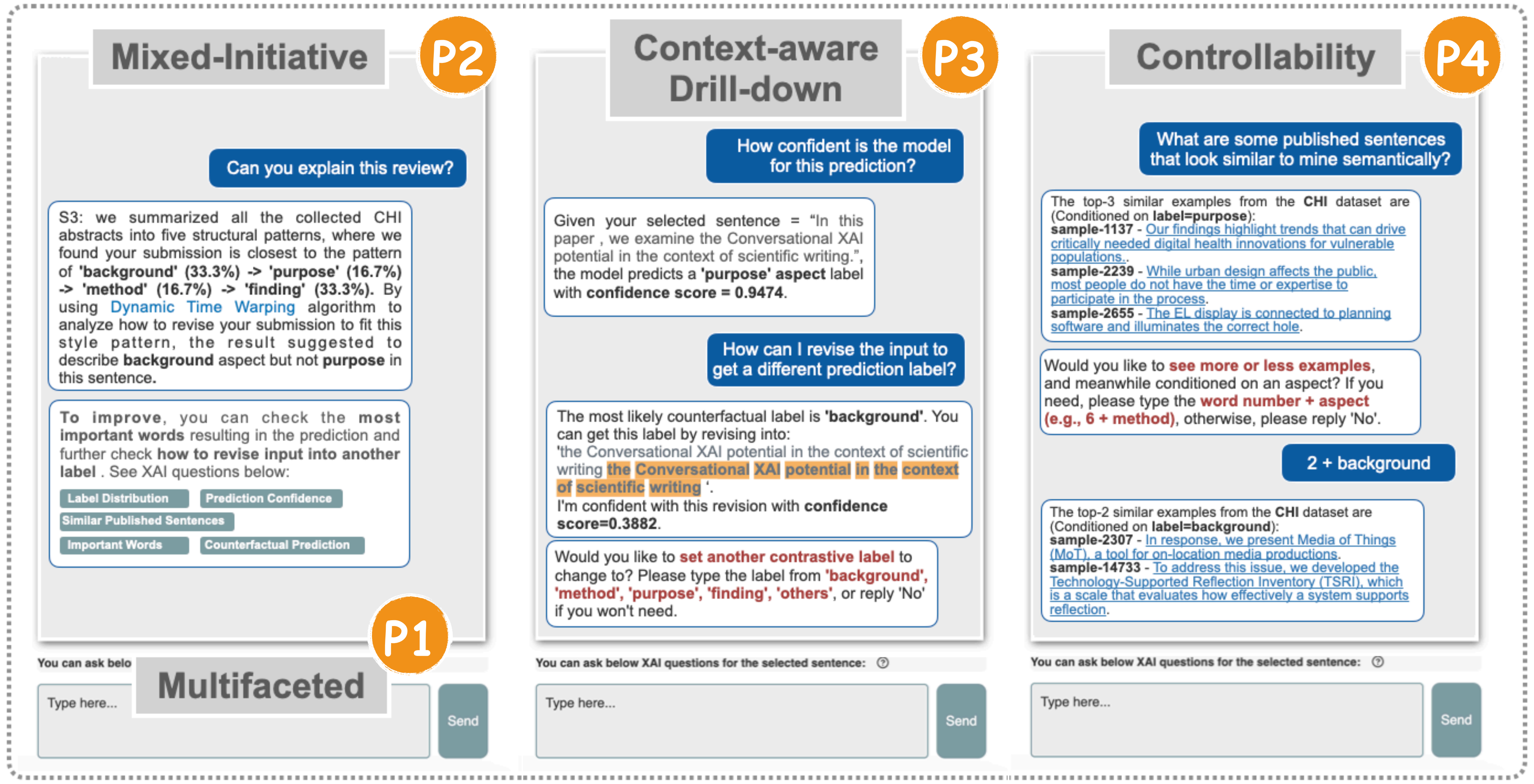


ConvXAI



Keynotes of User Interaction

Implementing Four Design Principles



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

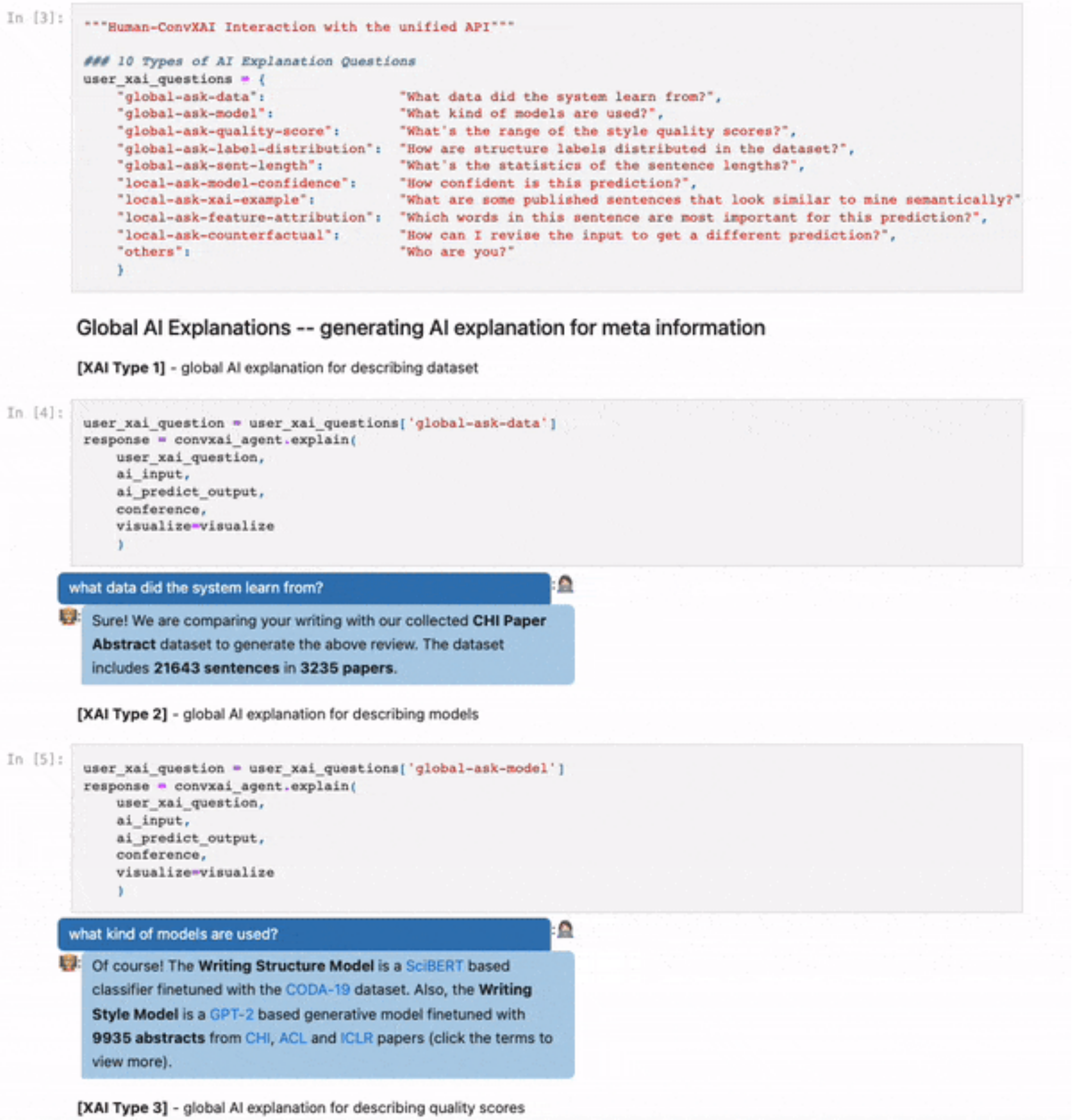
Technical Challenges & Contributions

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**.

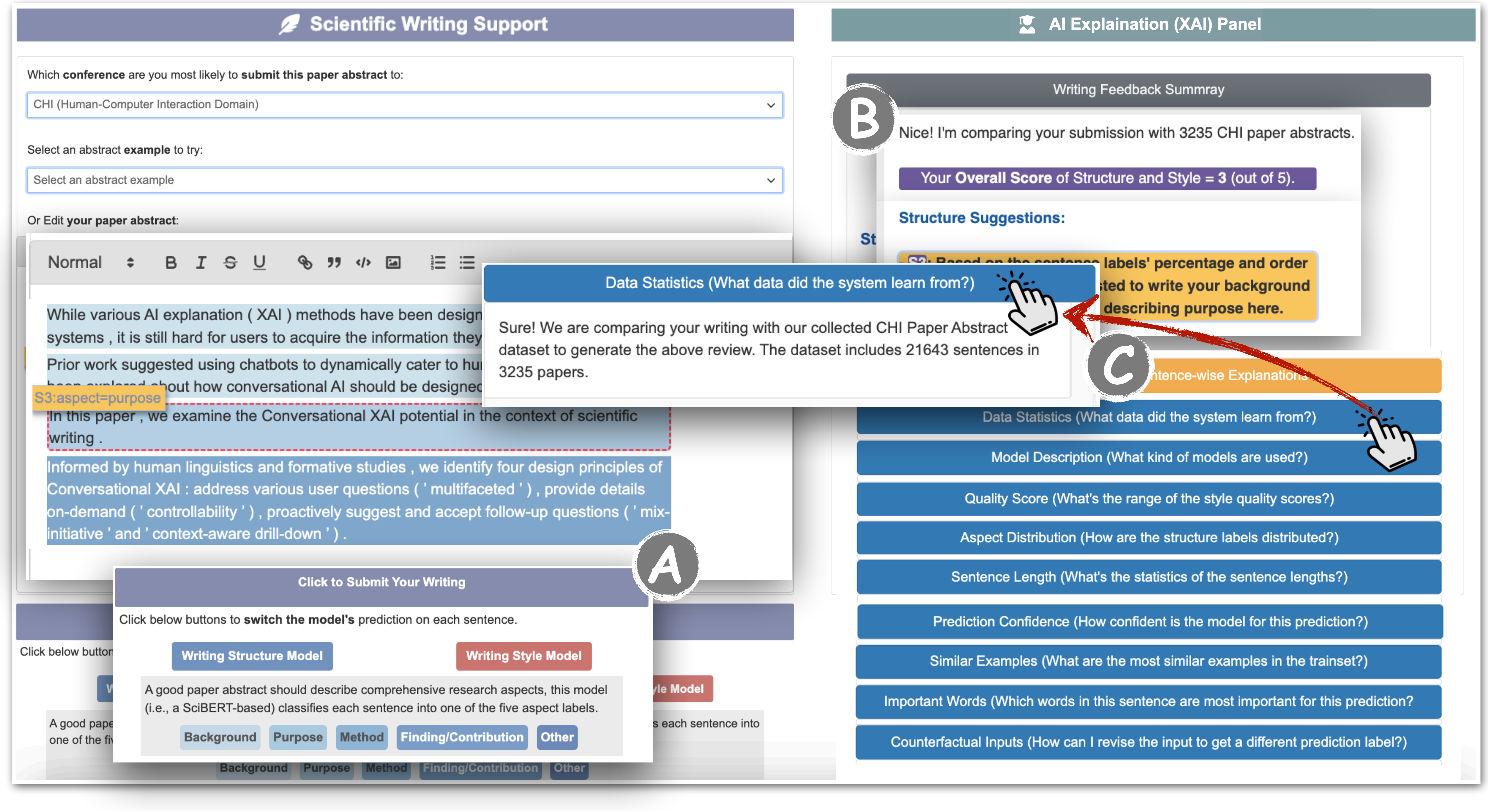


Evaluate ConvXAI with real human studies

	Task1	Task2
Who is studies	**13** graduate researchers	**8** researchers
When	09/2022 (90min)	12/2022 (90min) (**rejoin**)
How it's studied	1. **Two** think-aloud **scientific writing tasks**: • Within-Subjects Study: ConvXAI vs. Baseline • Improve a paper's abstract; • Paper domains: NLP, or HCI, or AI 2. Post **Survey -** Questionnaires 3. Semi-**structured Interviews**	
What is the goal	• Effective system design • Self-perceived usefulness	• Writing output quality • Temporal change of user need

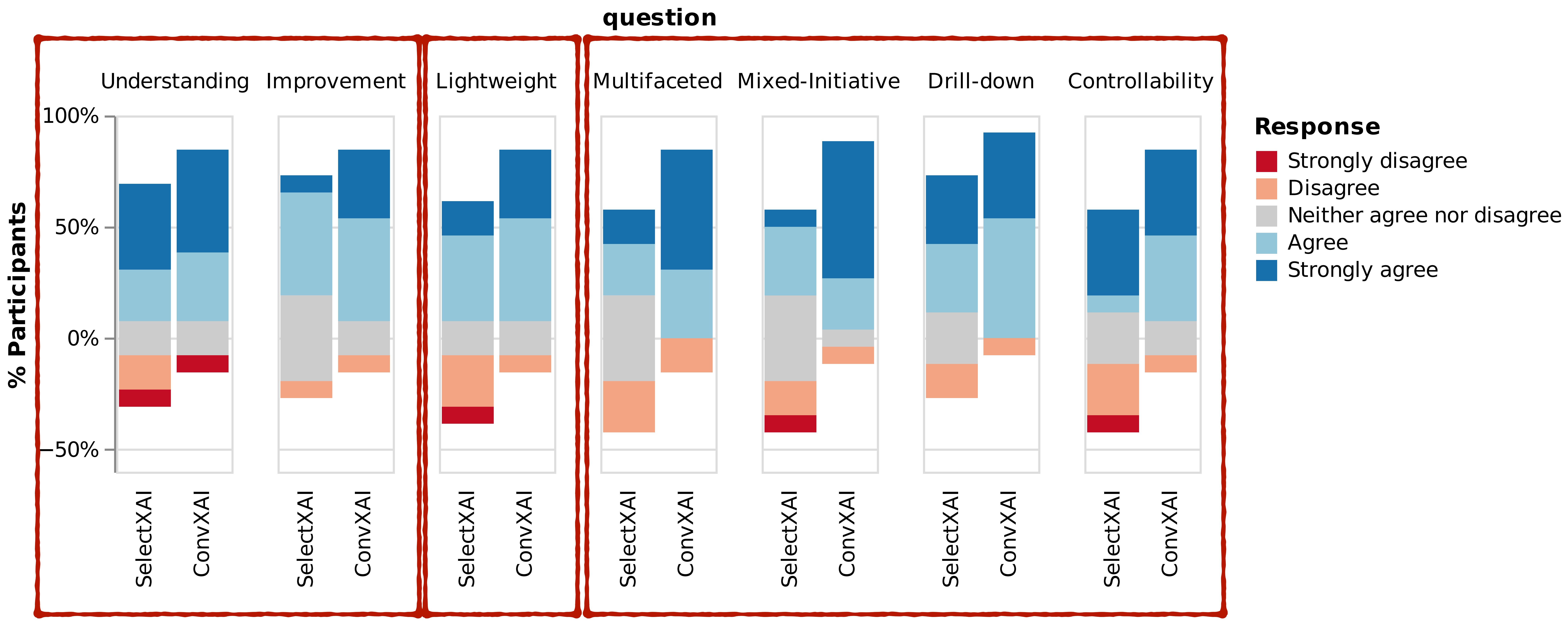
Baseline System (SelectXAI)

Within-Subjects Study Design



Survey results of human study in Task1

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



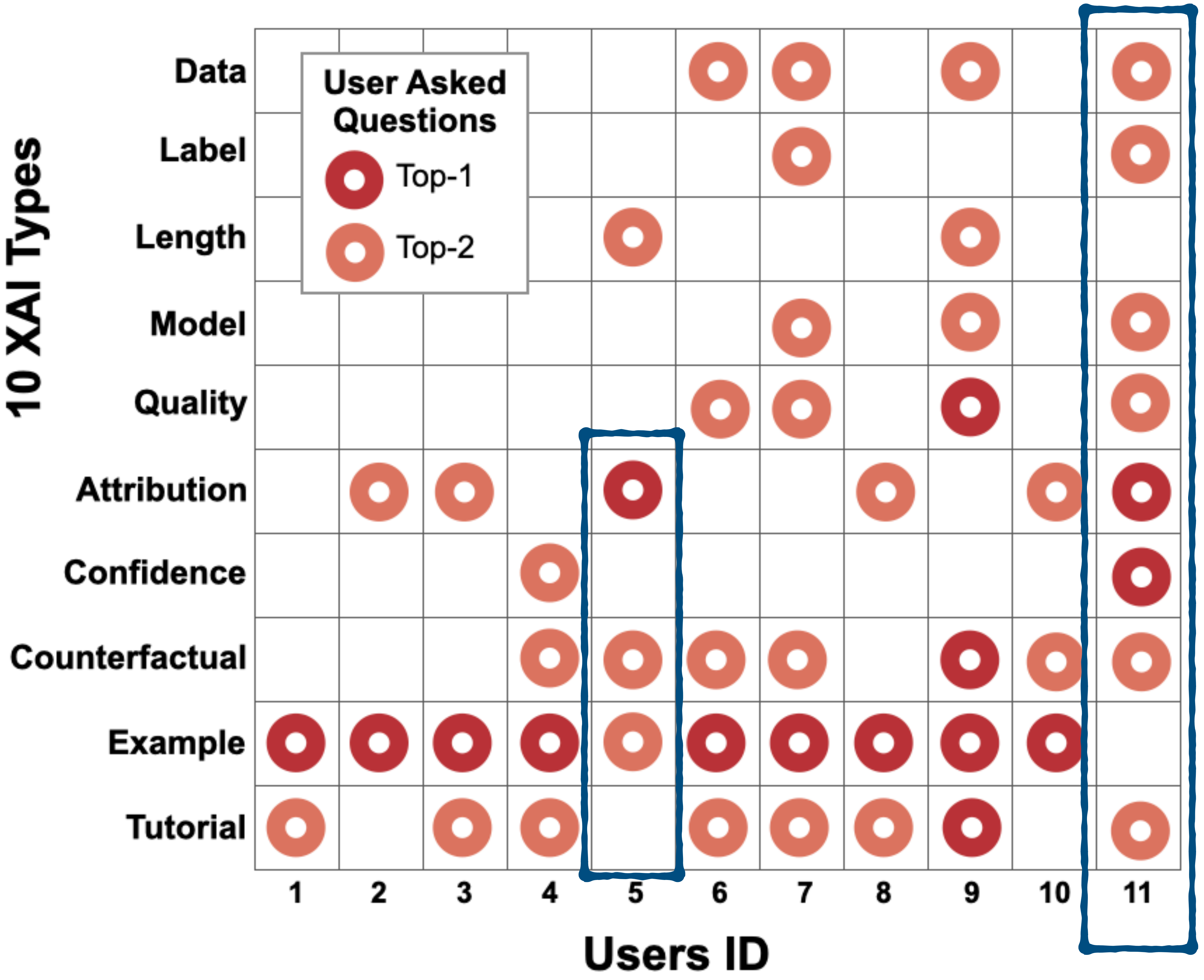
Useful in Understanding & Improving Writing

Less Cognitive Load

More aligned with human-centered design rationales

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.

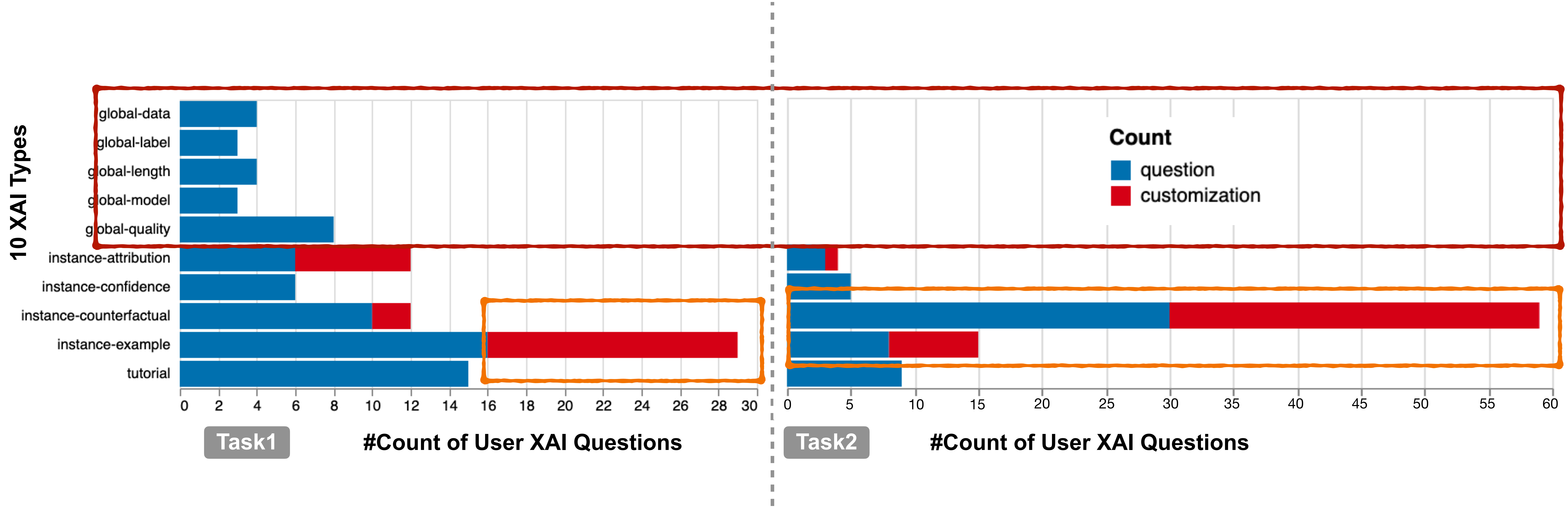
Condition ↑	Grammarly (1-100)		Model Quality (1-5)		Model Structure (1-5)		Human Quality (1-10)		Human Structure (1-10)	
	Original	Improved	Original	Improved	Original	Improved	Original	Improved	Original	Improved
SelectXAI	84.8 (±10.4)	85.1 (±5.52)	2.82 (0.75)	3.05 (0.64)	4.19 (0.37)	**4.75 (0.38)**	6.5 (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

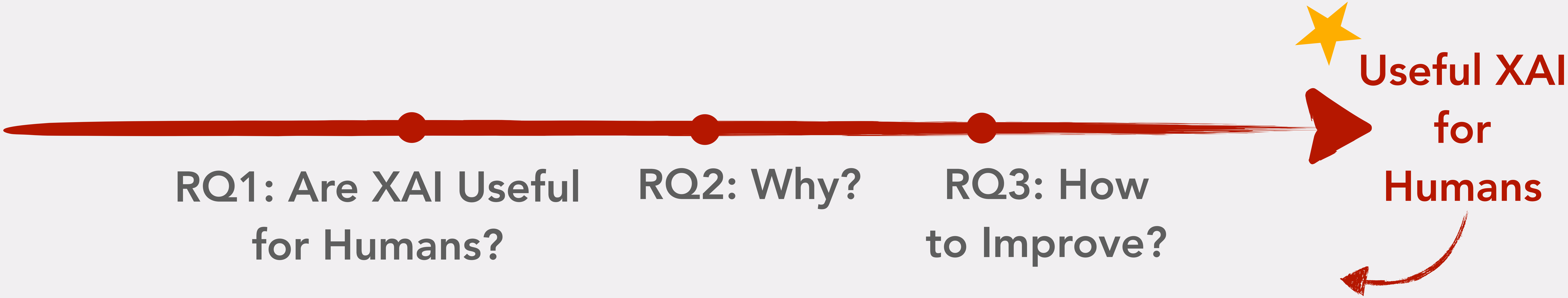
Real Humans	**Application-grounded Evaluation** (e.g., Collaboration usefulness, Performance gain)	Real Tasks
Real Humans	**Human-grounded Evaluation** (e.g., Human simulatability, Error analysis)	Simple Tasks
No Real Humans	**Functionally-grounded Evaluation** (e.g., Faithfulness, Robustness, Plausibility)	Proxy Tasks

Humans ← AI Interpretability Evaluation → AI Tasks

CSCW 2023 Demo

ACL 2022

HCOMP 2020



Limitation

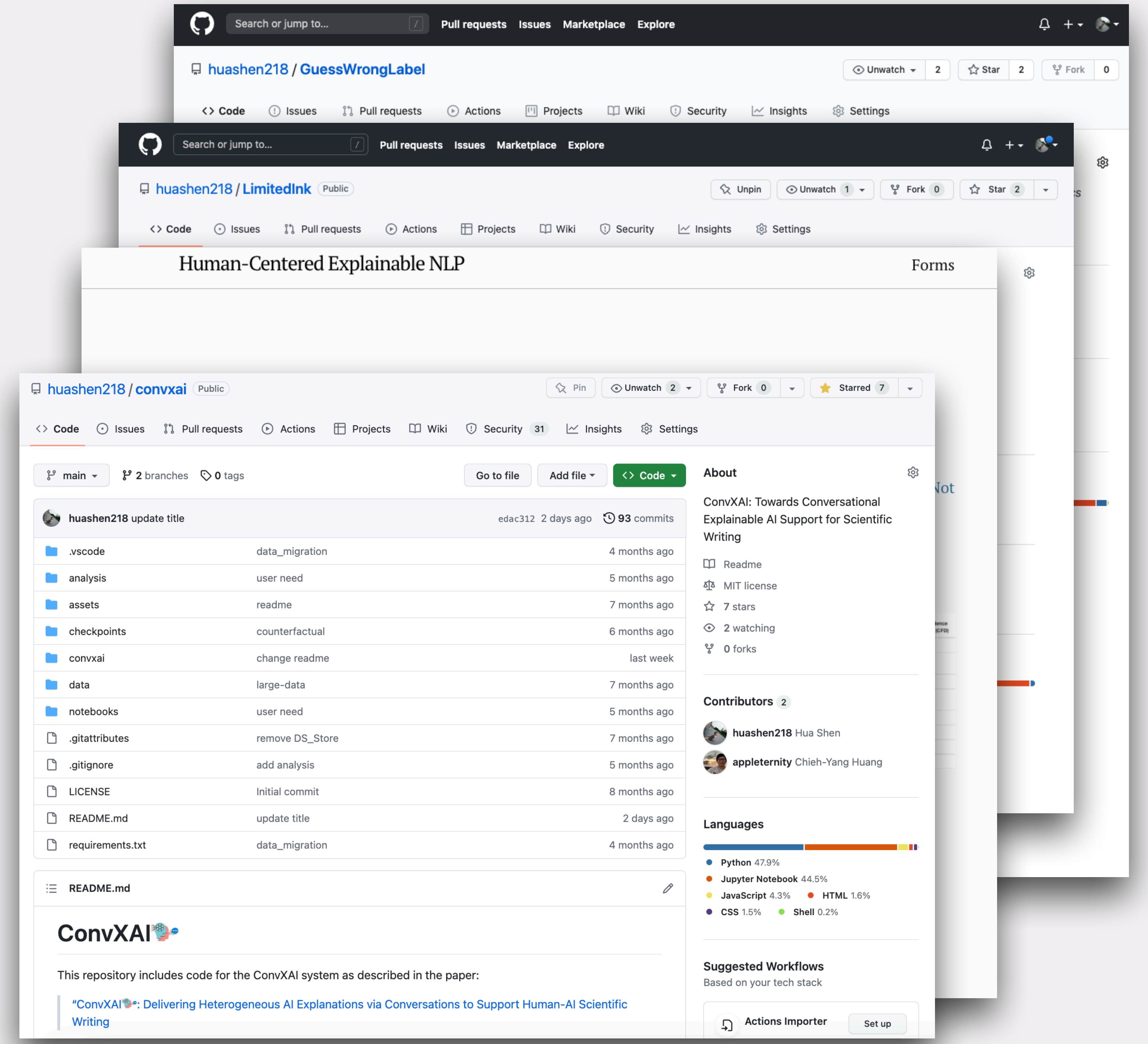
- In real world human-AI tasks, “how to quantify human’s subjective goal of XAI usefulness, and align it with objective AI predictions” is still challenging.

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

amazon

thank you!