THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

ENCORE

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Rising Star In Data Science

November
13th-14th, 2023

Towards Useful AI Interpretability for Humans via Interactive AI Explanations

Hua Shen

huashen@umich.edu @huashen218

University of Michigan

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

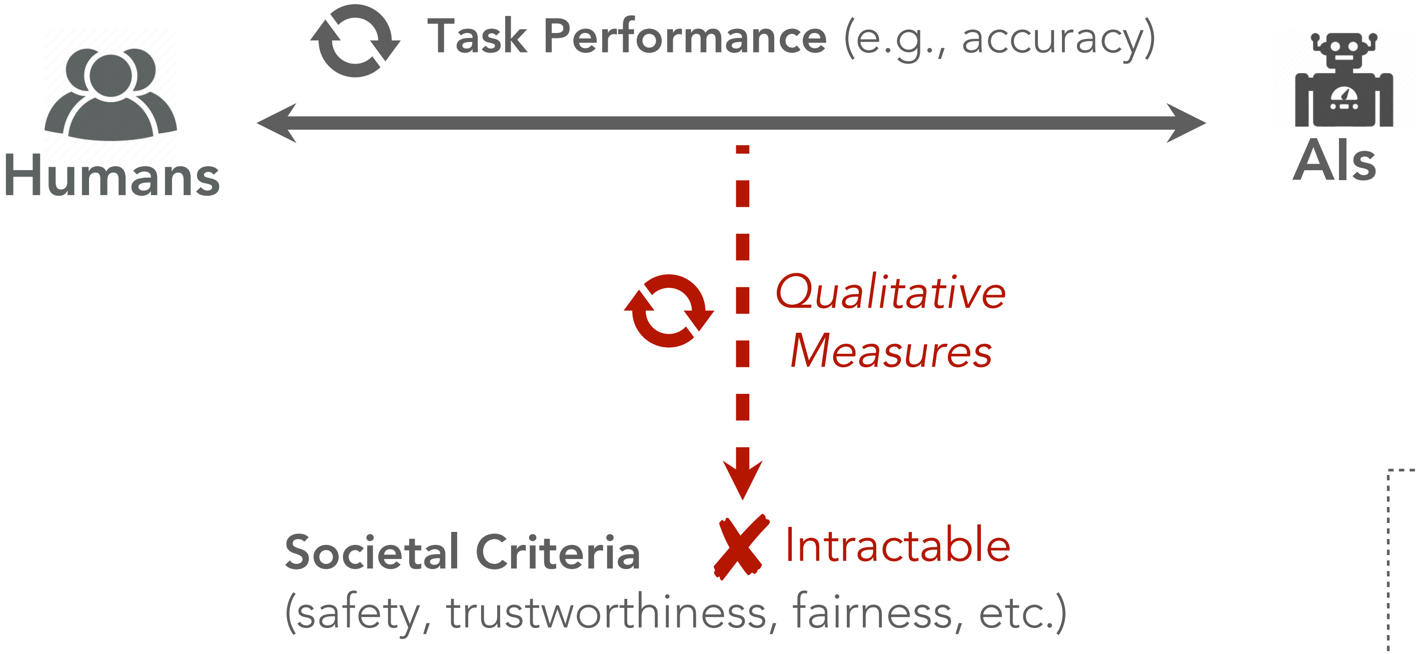
CSE COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF MICHIGAN

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.).

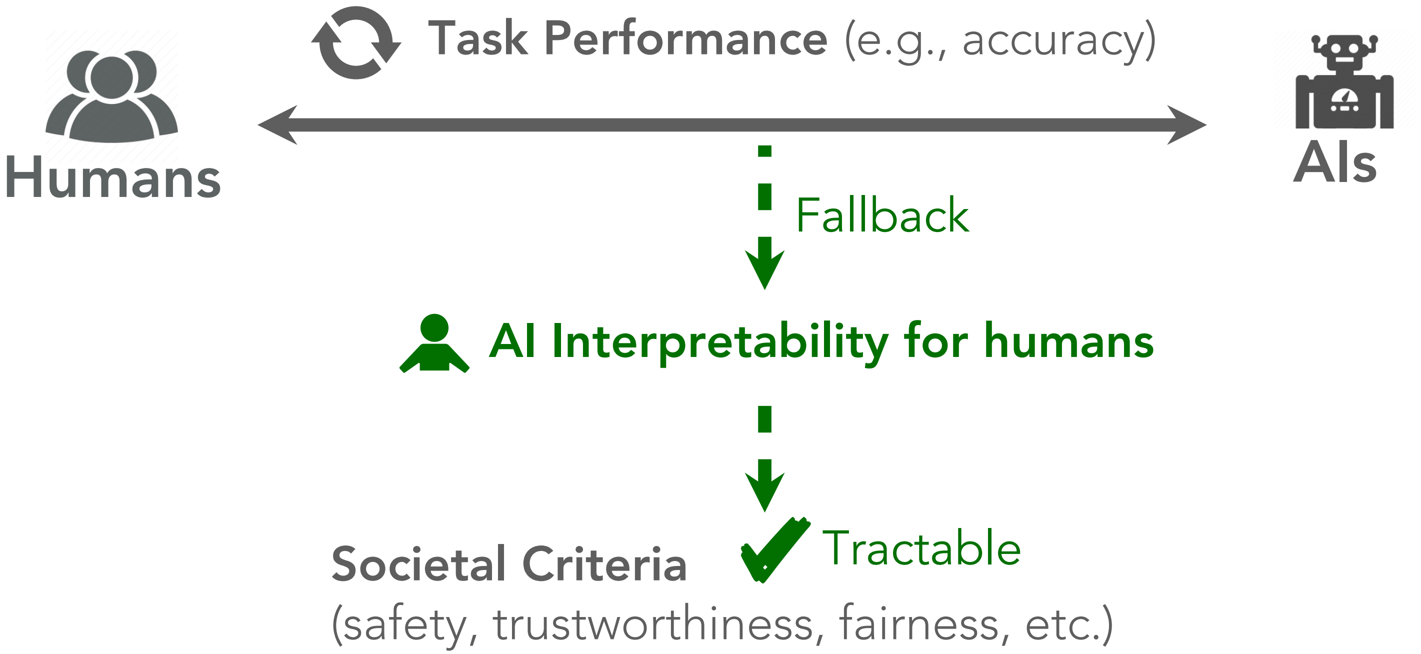


Task Performance (e.g., accuracy)
Humans
AIs
Qualitative Measures
✗ Intractable
Societal Criteria
(safety, trustworthiness, fairness, etc.)

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**"



Task Performance (e.g., accuracy)

Humans

AIs

Fallback

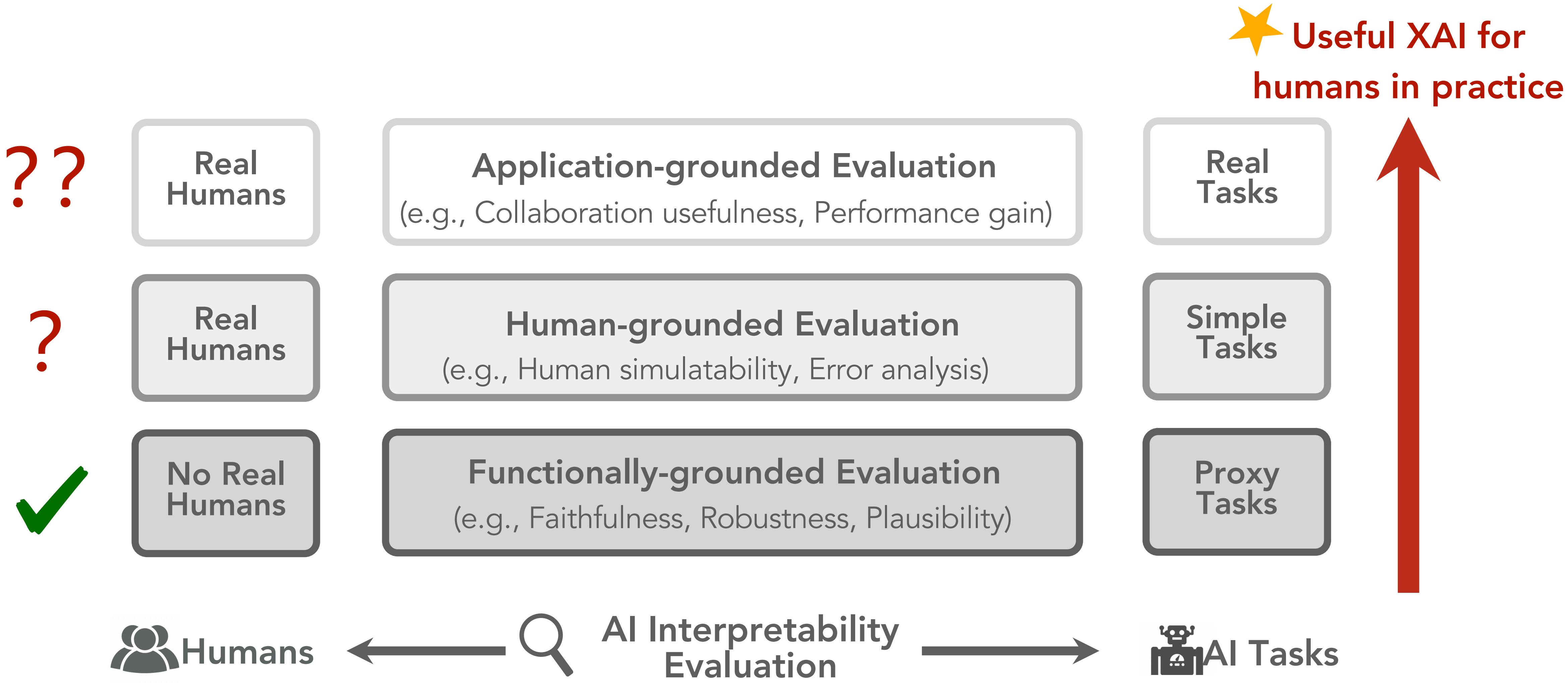
AI Interpretability for humans

Societal Criteria
(safety, trustworthiness, fairness, etc.)

✓ Tractable

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

Evaluation of XAI usefulness



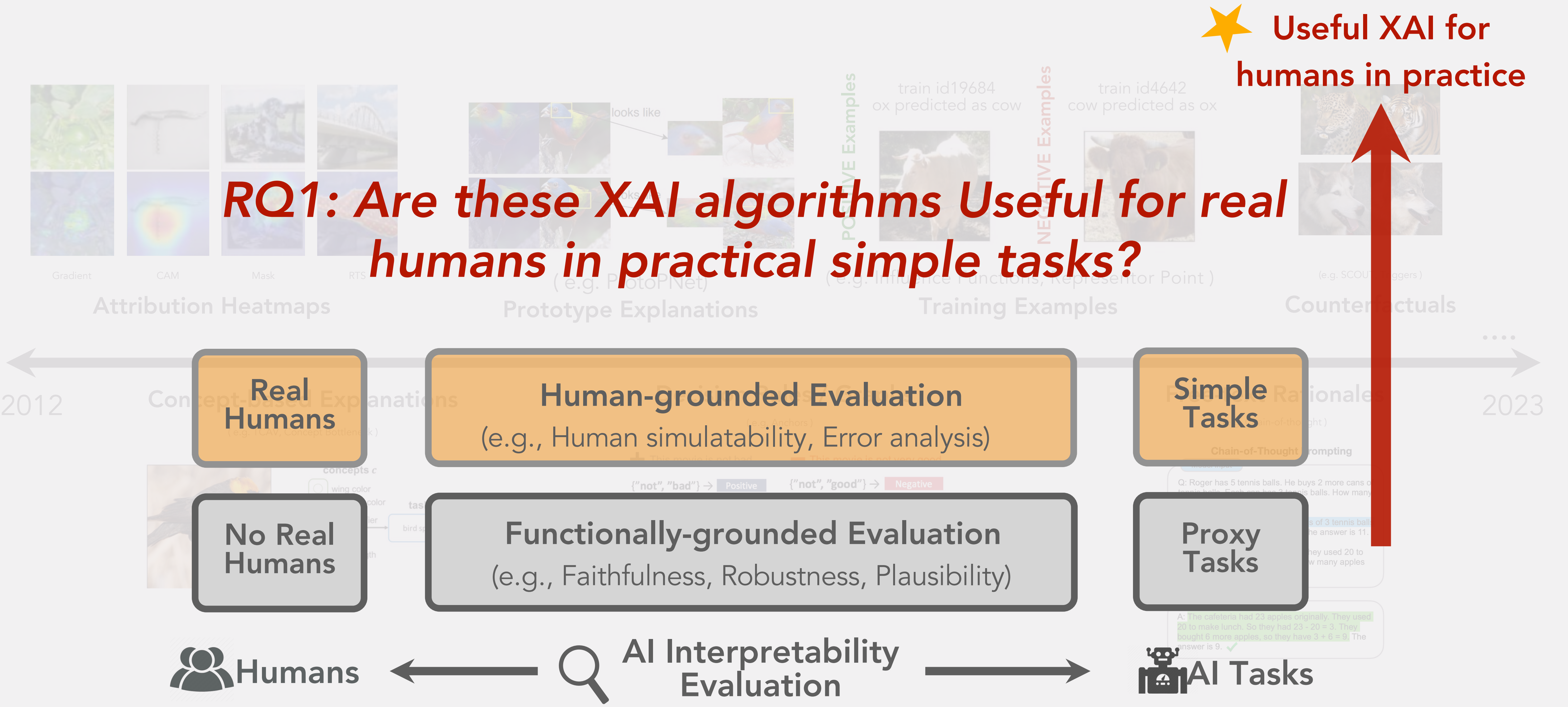
Useful XAI for humans in practice

??	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

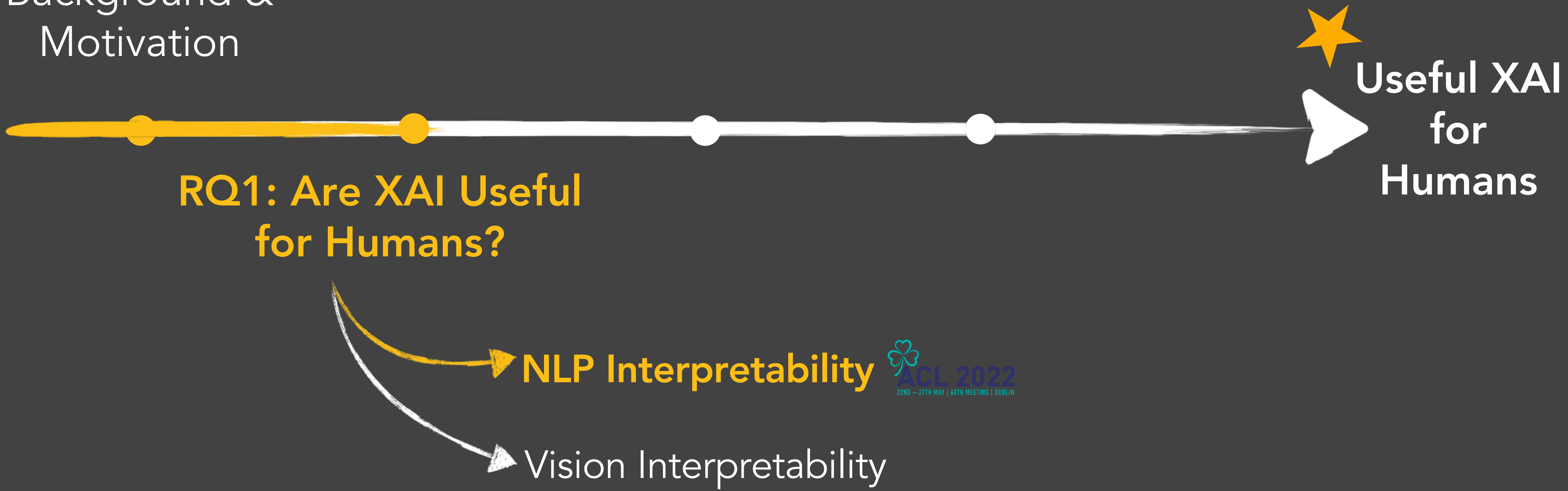
(Doshi-Velez, Finale, & Kim, Been. (2017))

Under-Explored: human evaluation of XAI usefulness



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

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

RQ1: Are XAI Useful for Humans?

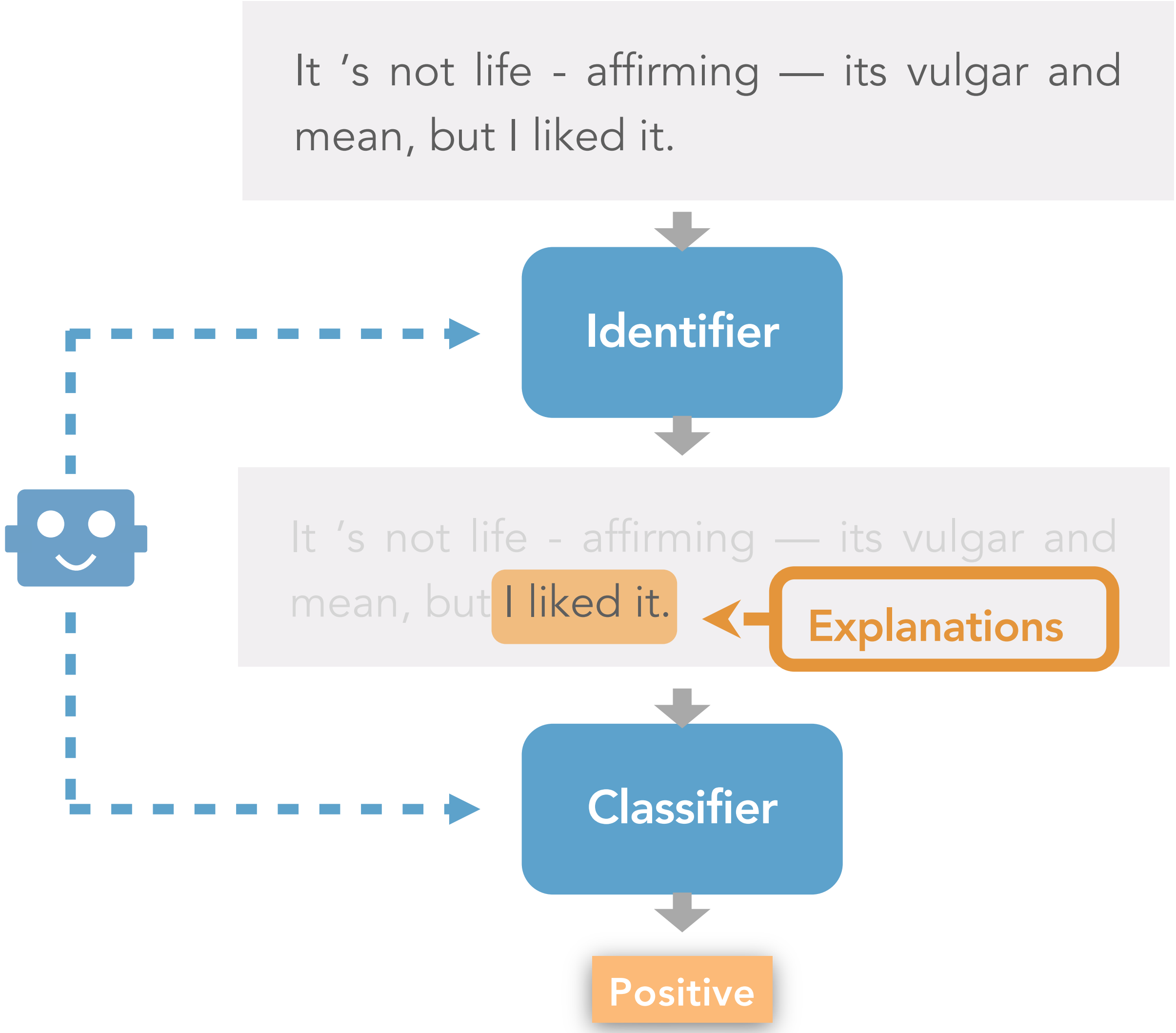
NLP Interpretability

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

Vision Interpretability

Useful XAI for Humans

Self-Explaining Language Models


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

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

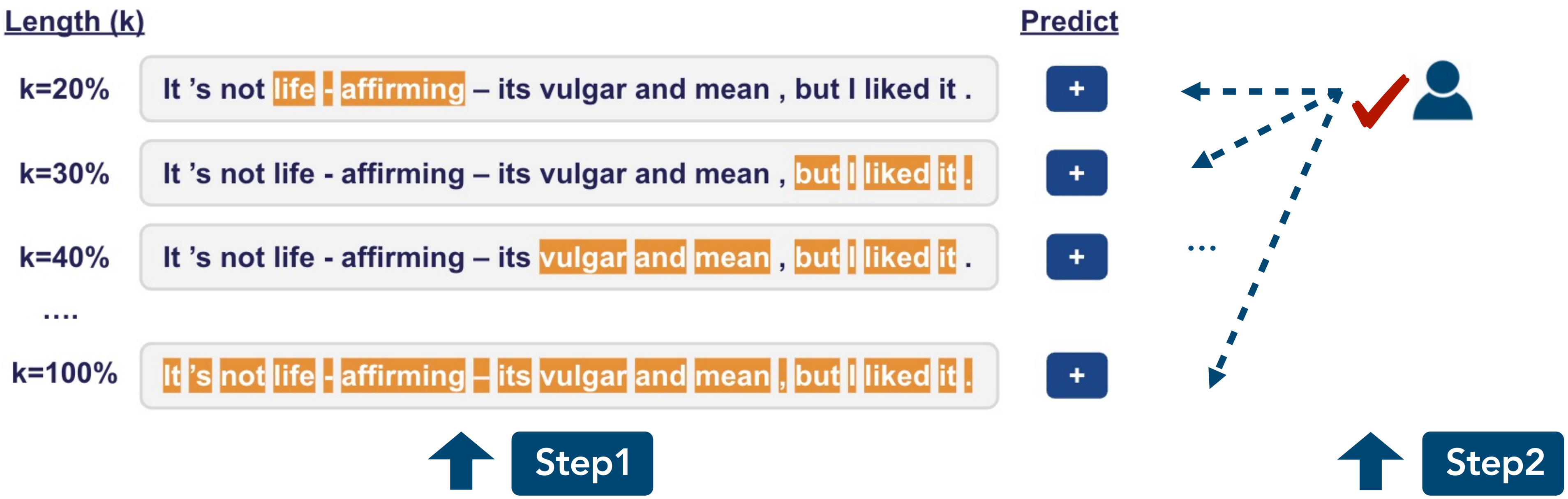
AI Researchers' Assumption

Shorter Explanations are Better for End Users.

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

Lei, Tao, Regina Barzilay, and Tommi Jaakkola. "Rationalizing neural predictions." EMNLP, 2016.
Vafa, Keyon, et al. "Rationales for sequential predictions." EMNLP, 2021.
Bastings, Jasmijn, et al. "Interpretable neural predictions with differentiable binary variables." ACL, 2019.

Are *Shortest* AI Explanations the *Most Useful* for *Human Understanding*?



k=20% It 's not life - affirming – its vulgar and mean , but I liked it .

k=30% It 's not life - affirming – its vulgar and mean , but I liked it .

k=40% It 's not life - affirming – its vulgar and mean , but I liked it .

k=100% It 's not life - affirming – its vulgar and mean , but I liked it .

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

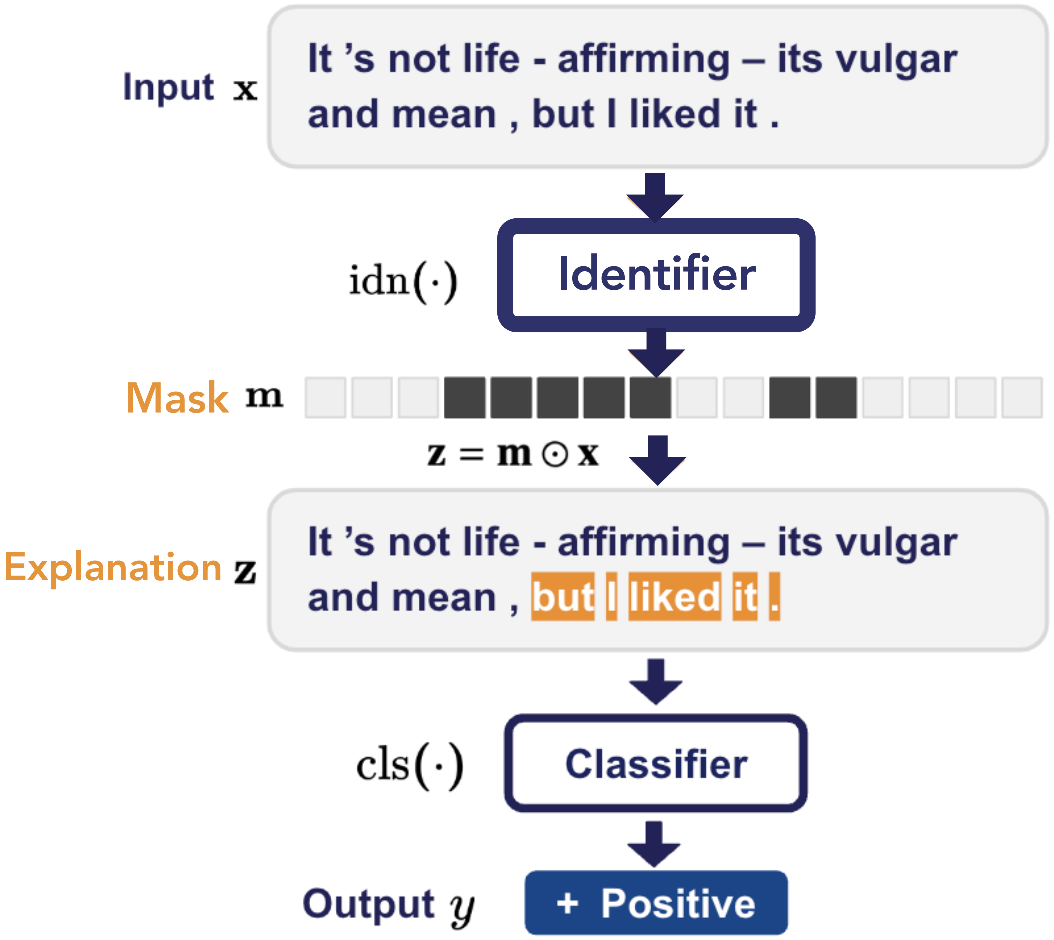
Humans guess the labels with explanations of different lengths

Contribution

A novel self-explaining model

human interactively guess and select the LM output

LimitedInk: A novel self-explaining LM



Input $\mathbf{x}$

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

$\mathrm{idn}(\cdot)$ Identifier

Mask $\mathbf{m}$

$\mathbf{z} = \mathbf{m} \odot \mathbf{x}$

Explanation $\mathbf{z}$

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

$\mathrm{cls}(\cdot)$ Classifier

Output y

+ Positive

Optimization Objective

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

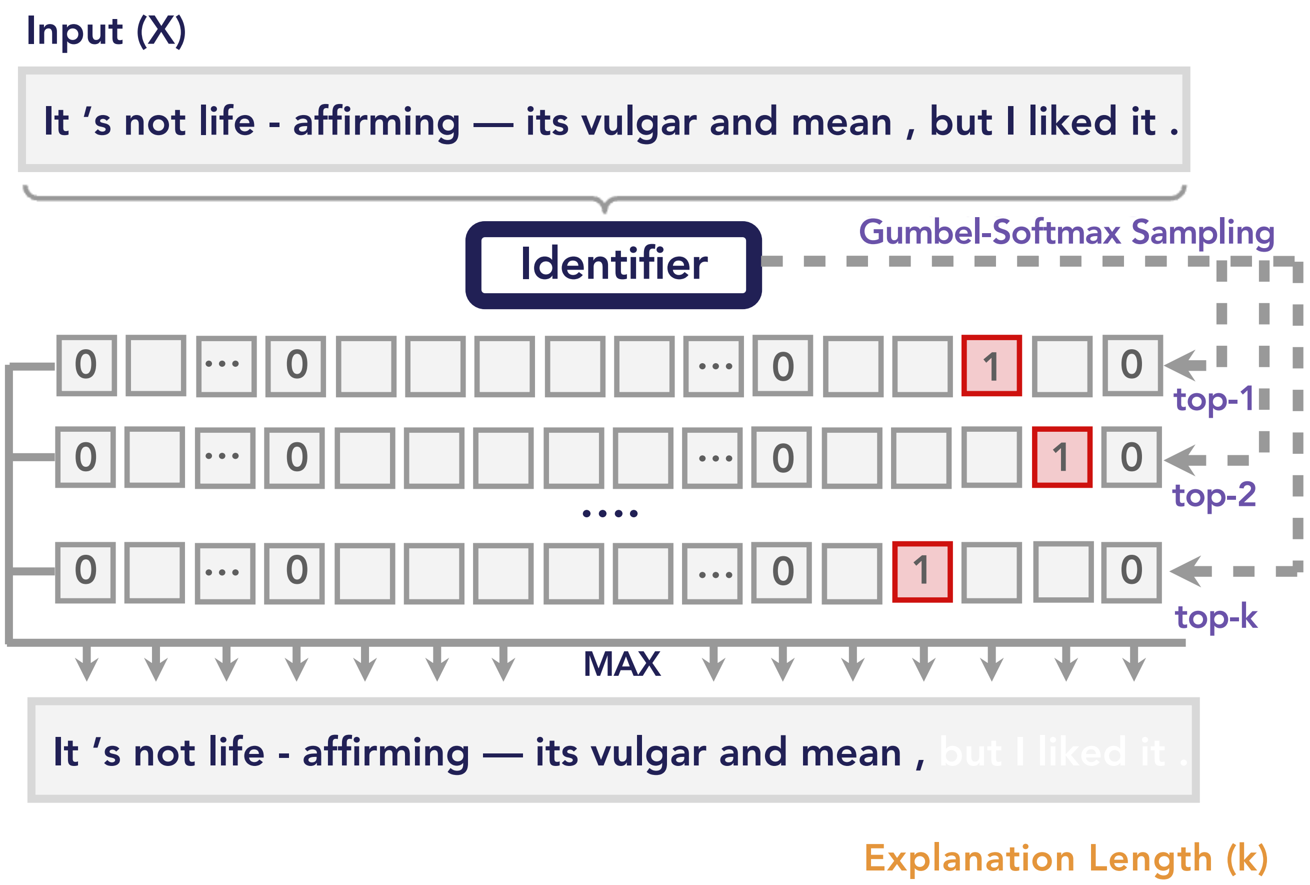
1. Gumbel-Softmax Sampling

2. Vector and Sort Regularization

Control Different Explanation Length

How to control explanation length in LimitedInk

1. Gumbel-Softmax Sampling

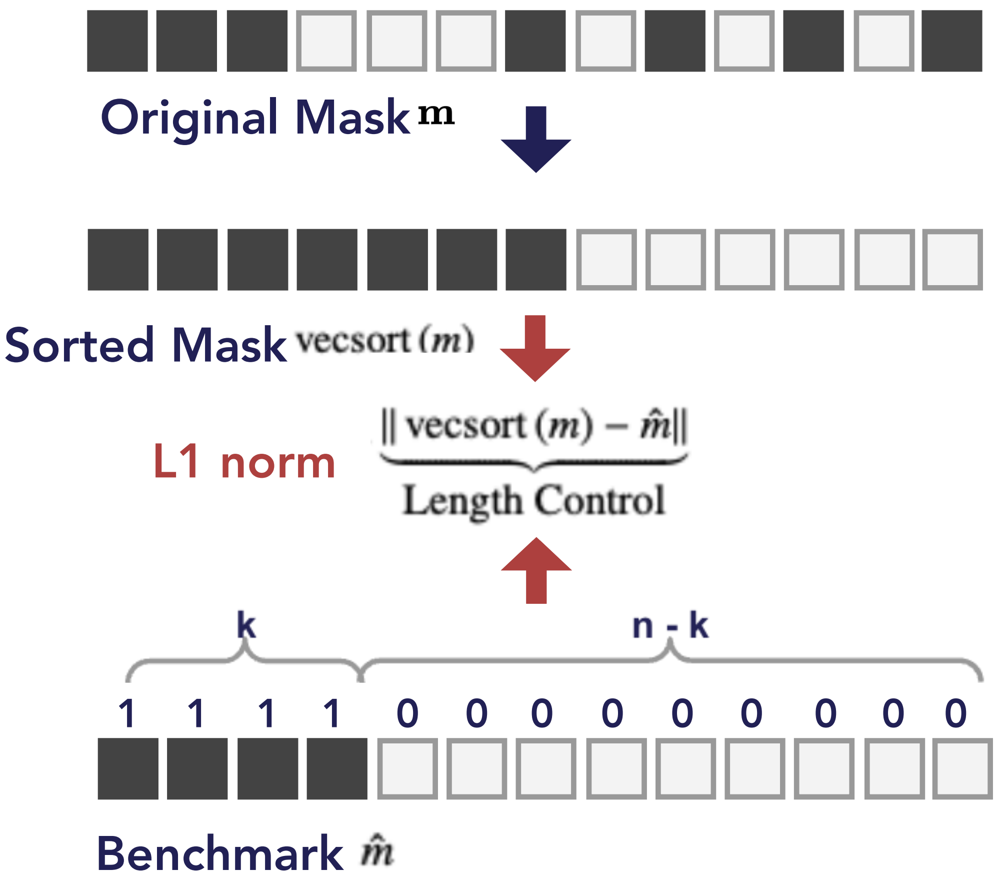

Input (X)

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

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

Explanation Length (k)

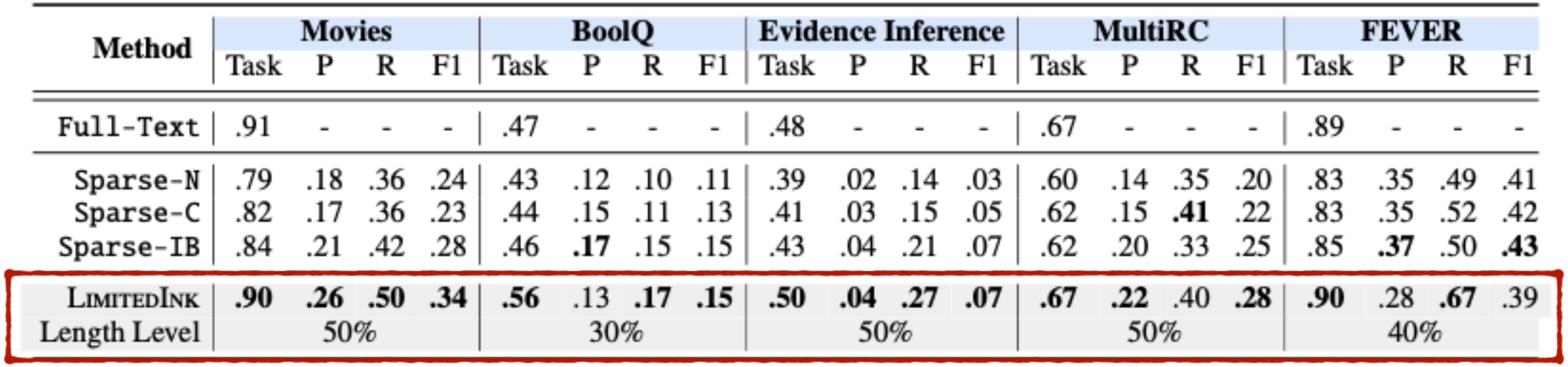
2. Vector and Sort Regularization


Original Mask $\mathbf{m}$
Sorted Mask $\text{vecsort}(m)$
L1 norm
$\underbrace{\|\text{vecsort}(m) - \hat{m}\|}_{\text{Length Control}}$
k
n - k
1 1 1 1 0 0 0 0 0 0 0 0 0
Benchmark $\hat{m}$

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**

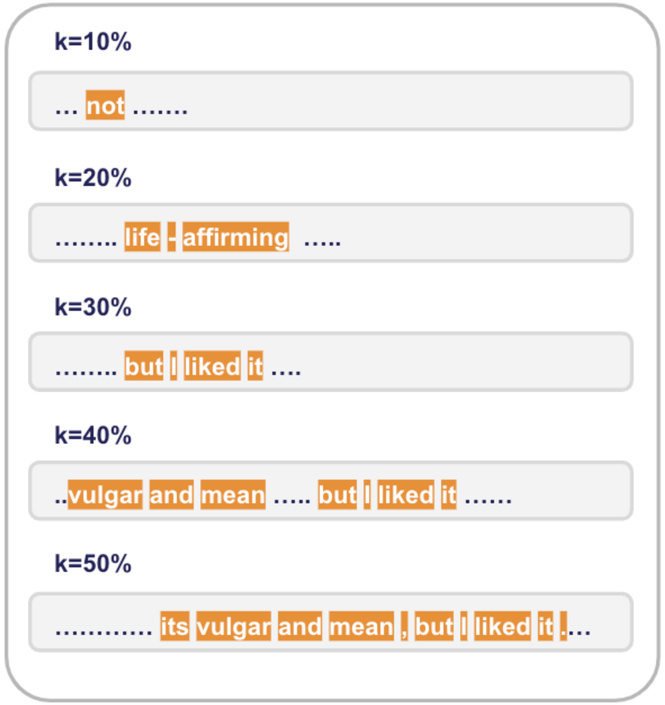


Method	Movies				BoolQ				Evidence Inference				MultiRC				FEVER			
	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1
Full-Text	.91	-	-	-	.47	-	-	-	.48	-	-	-	.67	-	-	-	.89	-	-	-
Sparse-N	.79	.18	.36	.24	.43	.12	.10	.11	.39	.02	.14	.03	.60	.14	.35	.20	.83	.35	.49	.41
Sparse-C	.82	.17	.36	.23	.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



k=10%

... not ....

k=20%

.... life - affirming .....

k=30%

.... but I liked it ....

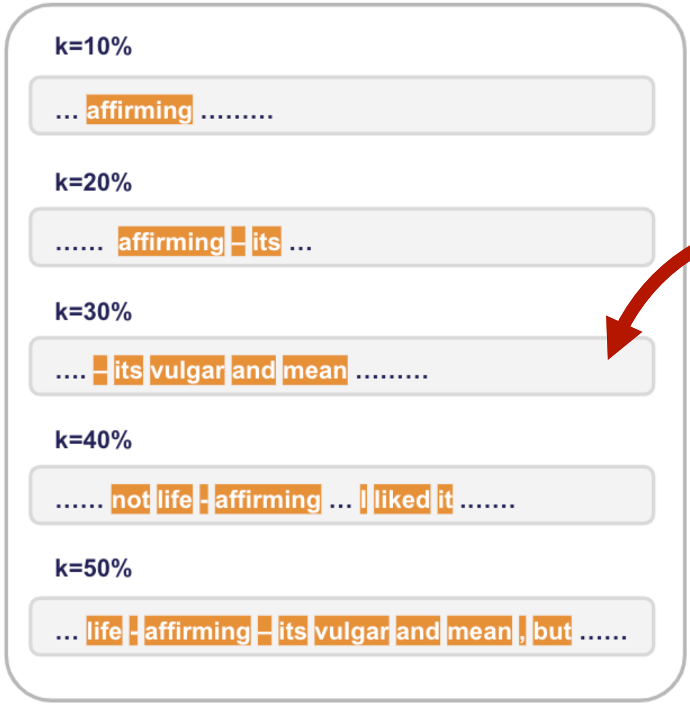
k=40%

..vulgar and mean ..... but I liked it ......

k=50%

.... its vulgar and mean , but I liked it ....

Random text spans (similar length)



k=10%

... affirming ....

k=20%

...... affirming - its ...

k=30%

.... - its vulgar and mean ....

k=40%

...... not life - affirming ... I liked it ....

k=50%

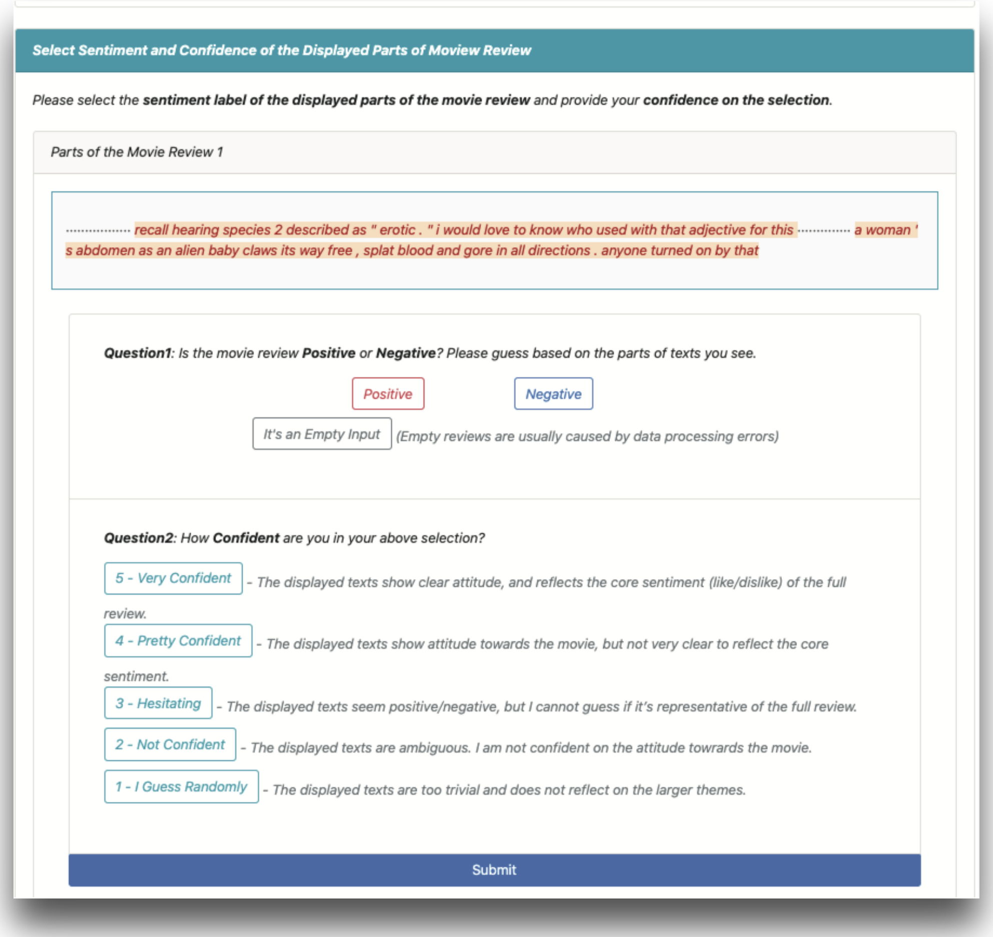
... life - affirming - its vulgar and mean , but ......

Only highlight explanations & hide other texts!

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

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

User Interface for Human Interaction



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 towrards the movie.*

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

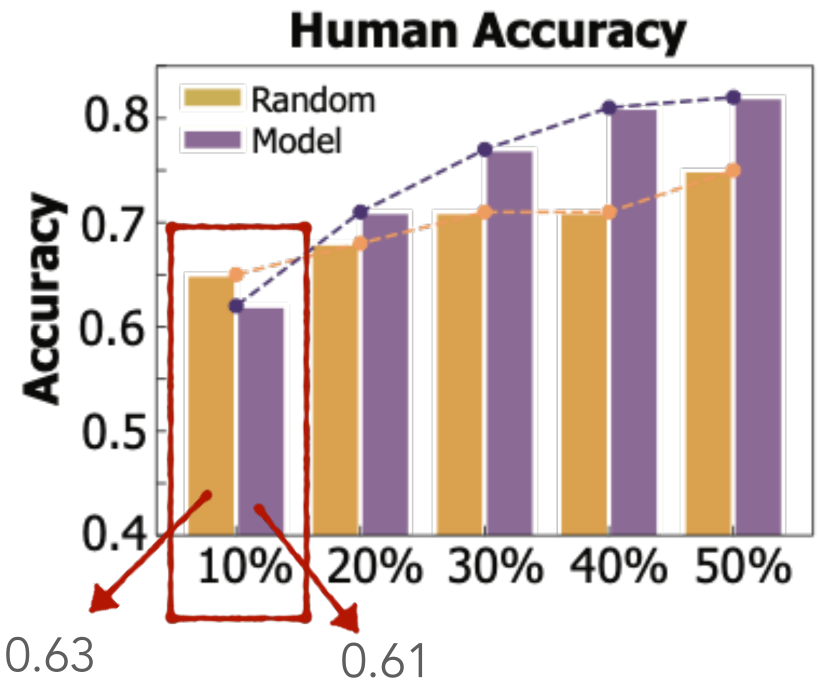
Submit

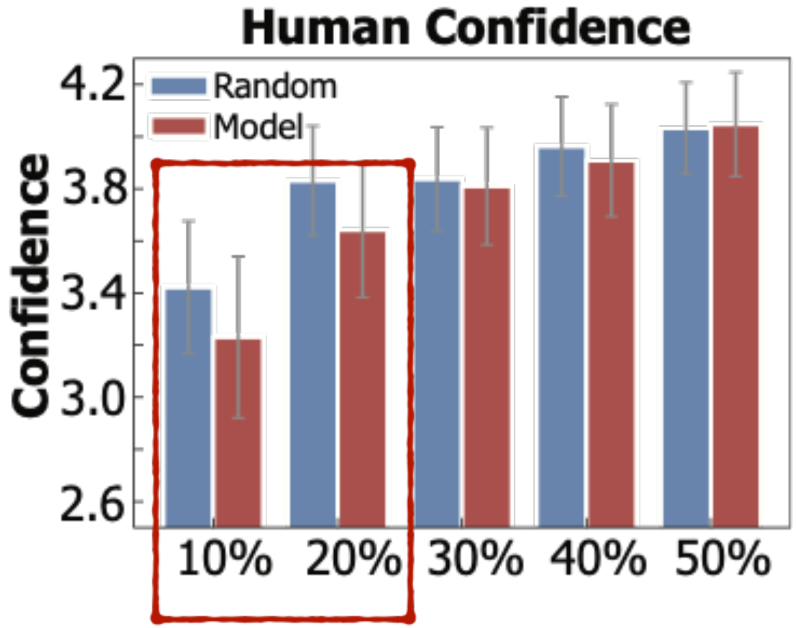
Sentiment Analysis:
we randomly sampled **100** reviews (correct prediction) from the **Movie review** test set

1. simulate model predictions
2. provide the confidence

MTurk Workers

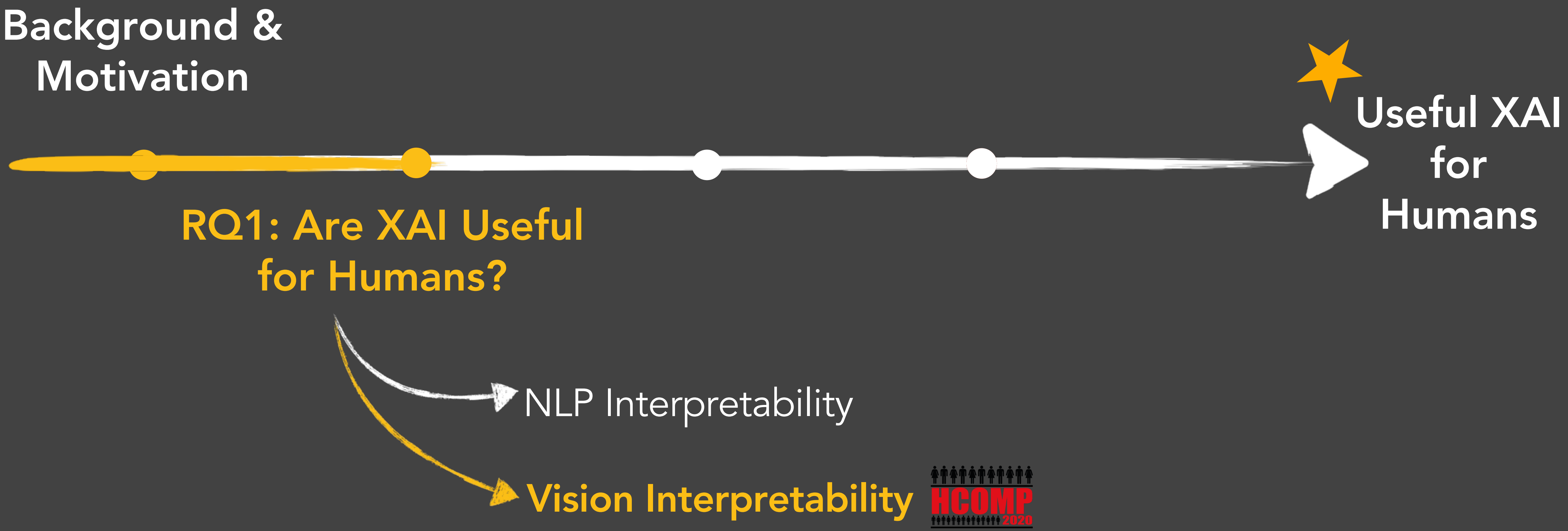
Key Findings





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

The **shortest AI explanations** are **NOT always Useful** for humans to understand the AI’s decision-making.



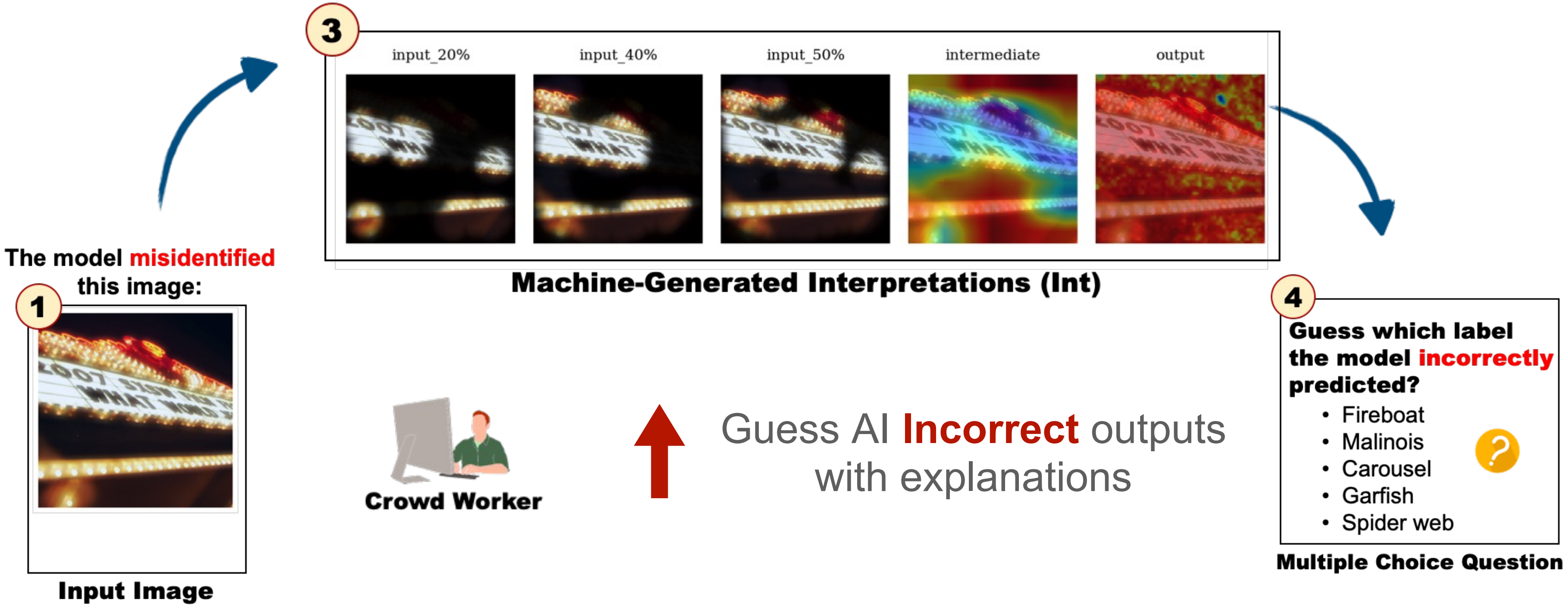
Background & Motivation

RQ1: Are XAI Useful for Humans?

NLP Interpretability

Vision Interpretability HCOMP 2020

Useful XAI for Humans



The model **misidentified** this image:

Input Image

Machine-Generated Interpretations (Int)

Guess AI **Incorrect** outputs with explanations

Guess which label the model incorrectly predicted?

- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web

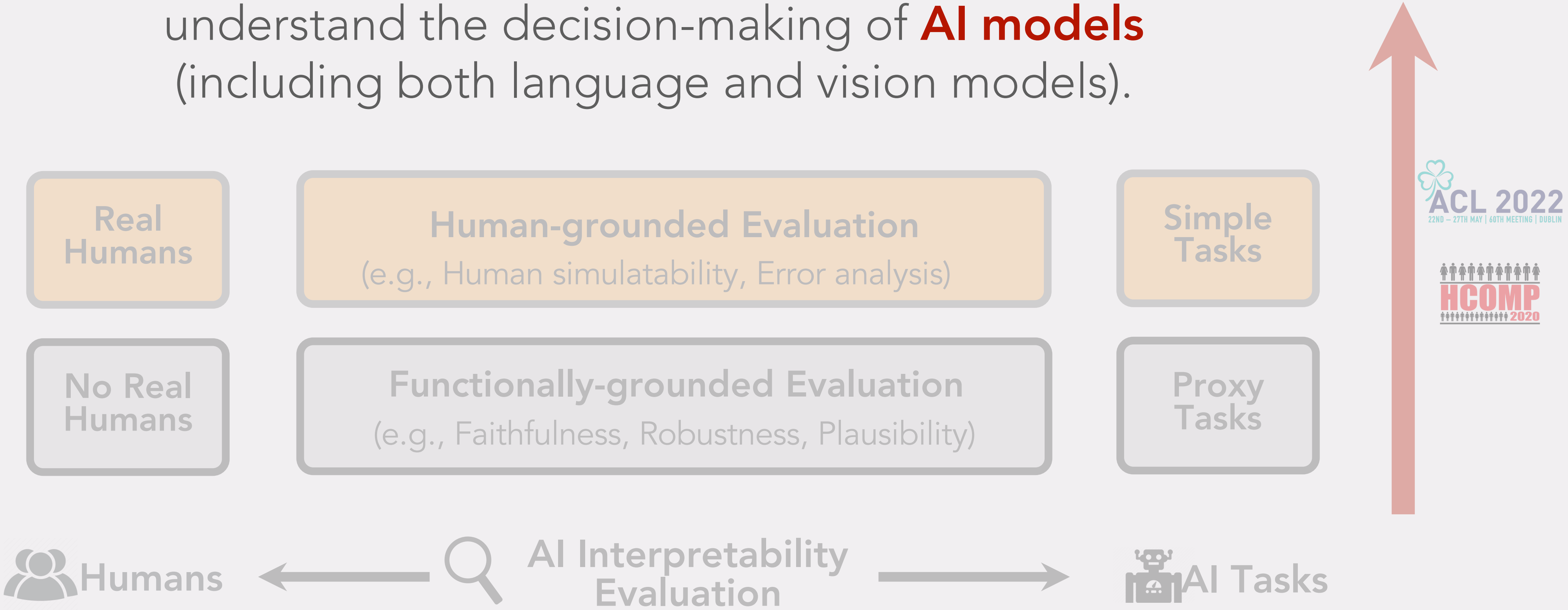
Multiple Choice Question

Visual AI explanations **did not increase,** but rather **decreased**, the **human's accuracy** in guessing the AI's **incorrect** decision-making.

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.

XAI is NOT always Useful for Humans

AI explanations are **NOT always useful** for **humans** to understand the decision-making of **AI models** (including both language and vision models).



Real Humans

Human-grounded Evaluation

(e.g., Human simulatability, Error analysis)

Simple Tasks

No Real Humans

Functionally-grounded Evaluation

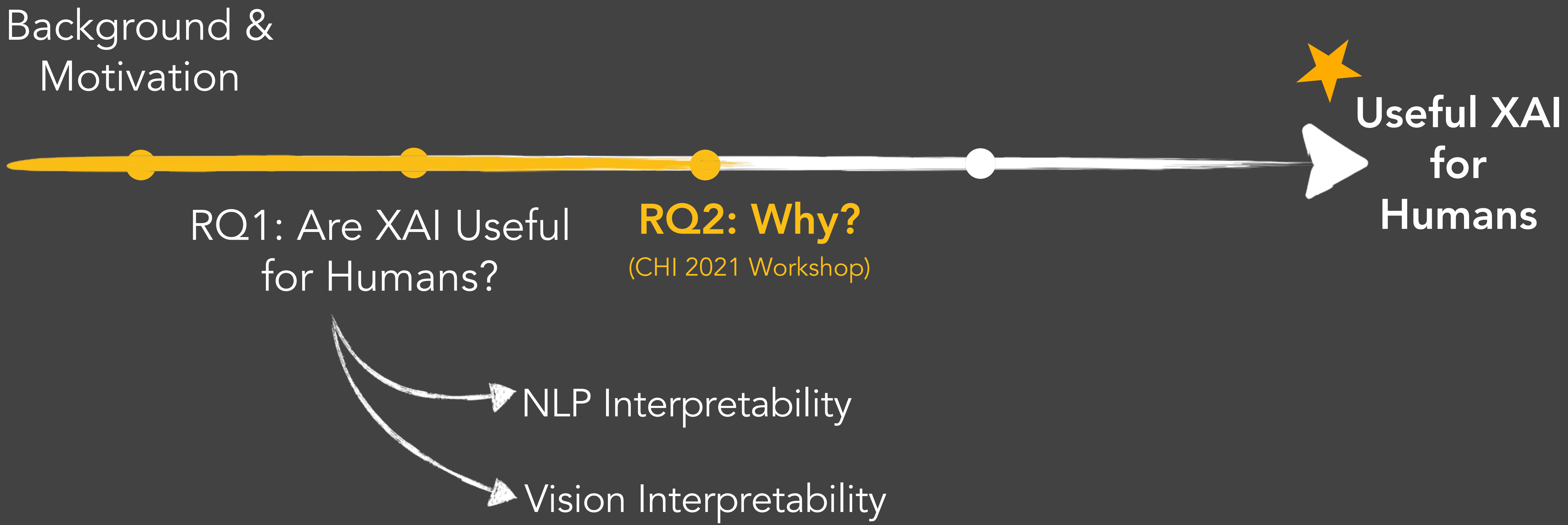
(e.g., Faithfulness, Robustness, Plausibility)

Proxy Tasks

Humans

AI Interpretability Evaluation

AI Tasks



Background & Motivation

RQ1: Are XAI Useful for Humans?

NLP Interpretability

Vision Interpretability

RQ2: Why?

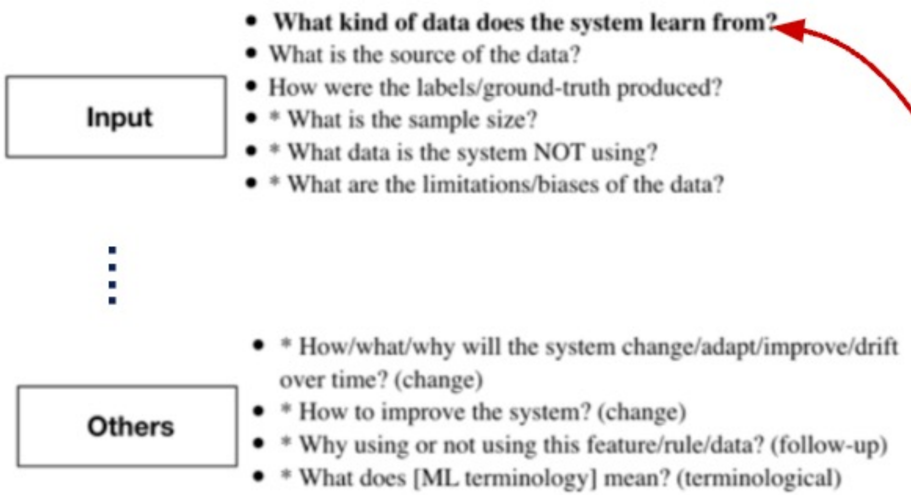
(CHI 2021 Workshop)

Useful XAI for Humans

Disparity between XAI with Humans?

43 User Questions in Practice

(Liao, Q. V., Gruen, D., & Miller, S. 2020)



Input

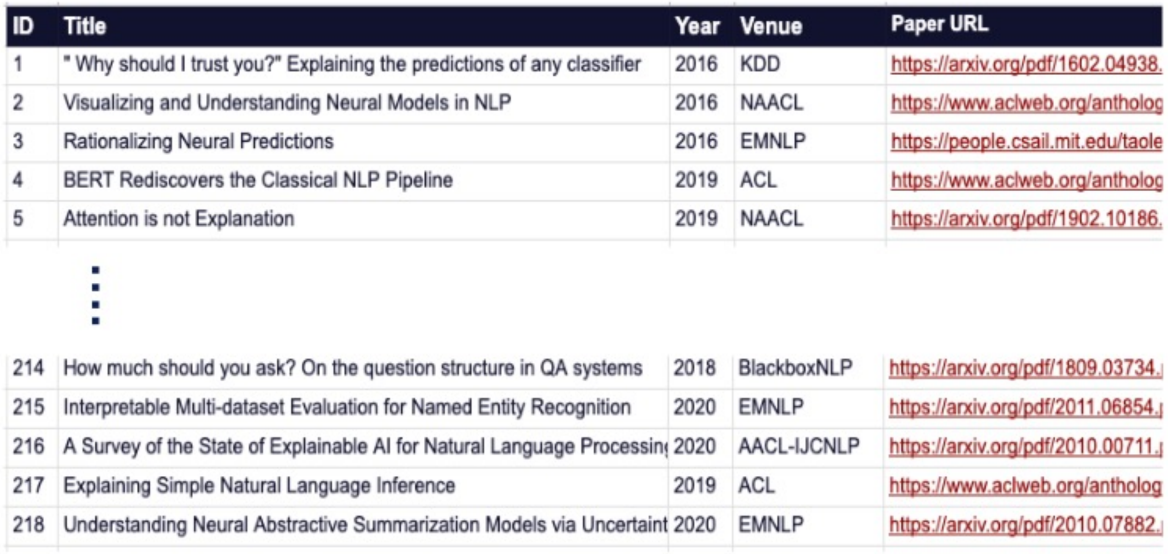
- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?

⋮

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 in NLP

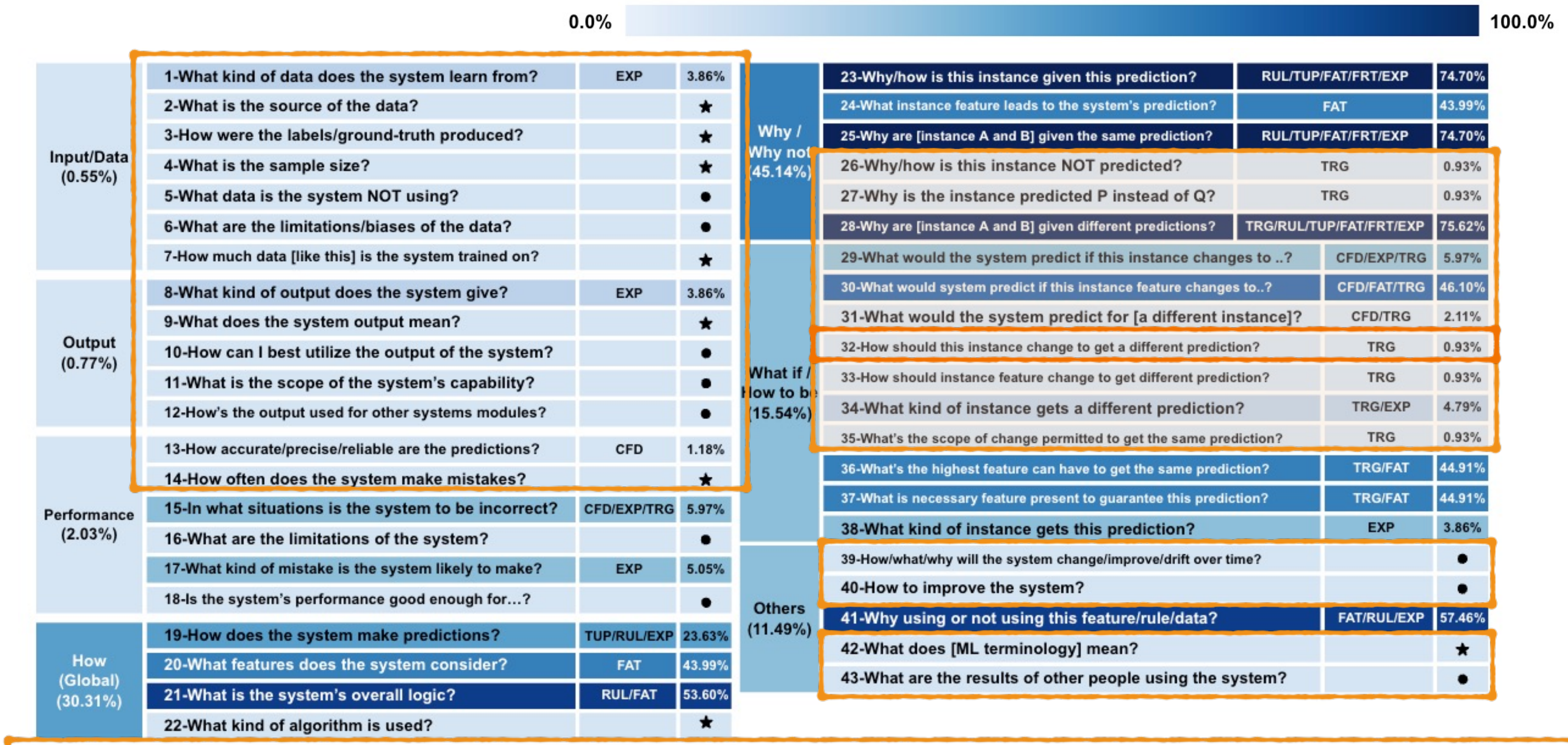


ID	Title	Year	Venue	Paper URL
1	" Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938.
2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/antholog
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
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.

We match the **disparity** between the existing **200+ XAI papers** with **43 practical user questions**!

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

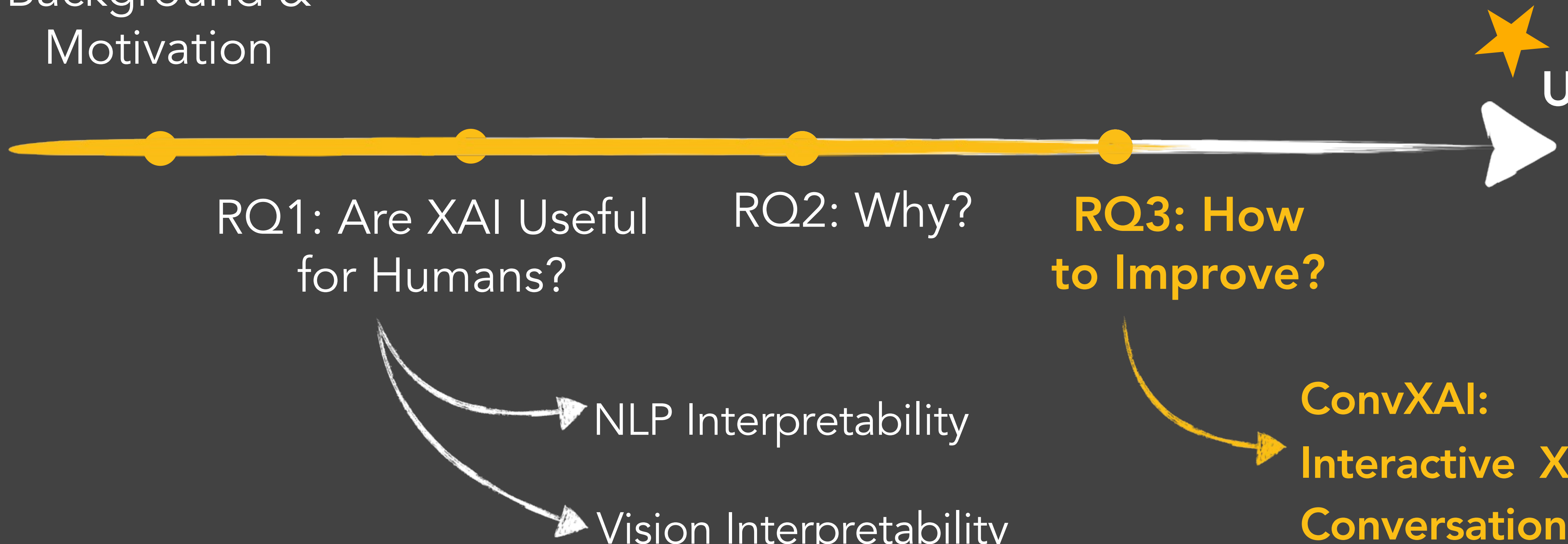
Existing XAIs largely Ignored…



	Question	Method	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?		★
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?		●

→ *What AI systems **CANNOT** achieve (e.g., counterfactuals).*

→ *Diverse information across the whole AI lifecycle (data, model, deployment, etc.)*



Background & Motivation

RQ1: Are XAI Useful for Humans?

NLP Interpretability

Vision Interpretability

RQ2: Why?

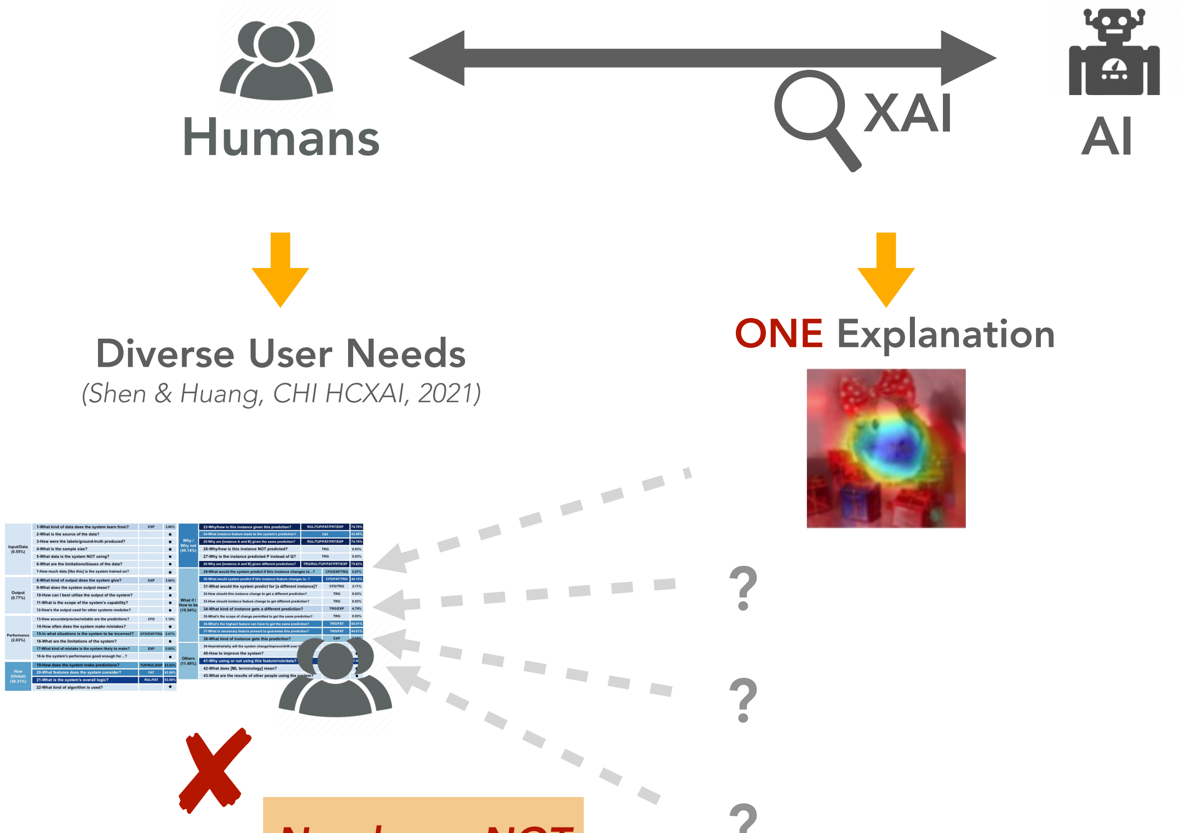
RQ3: How to Improve?

ConvXAI: Interactive XAI via Conversations

CSCW 2023 🏆 Best Demo

Useful XAI for Humans

Challenges of Existing XAI



Diverse User Needs
(Shen & Huang, CHI HCXAI, 2021)

ONE Explanation

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	[illegible]
Others (11.49%)	39-How/what/why will the system change/improve/drift over ti[illegible]	[illegible]	[illegible]
	40-How to improve the system?	[illegible]	●
	41-Why using or not using this feature/rule/data?	[illegible]	[illegible]
	42-What does [ML terminology] mean?	[illegible]	★
	43-What are the results of other people using the system?	[illegible]	●

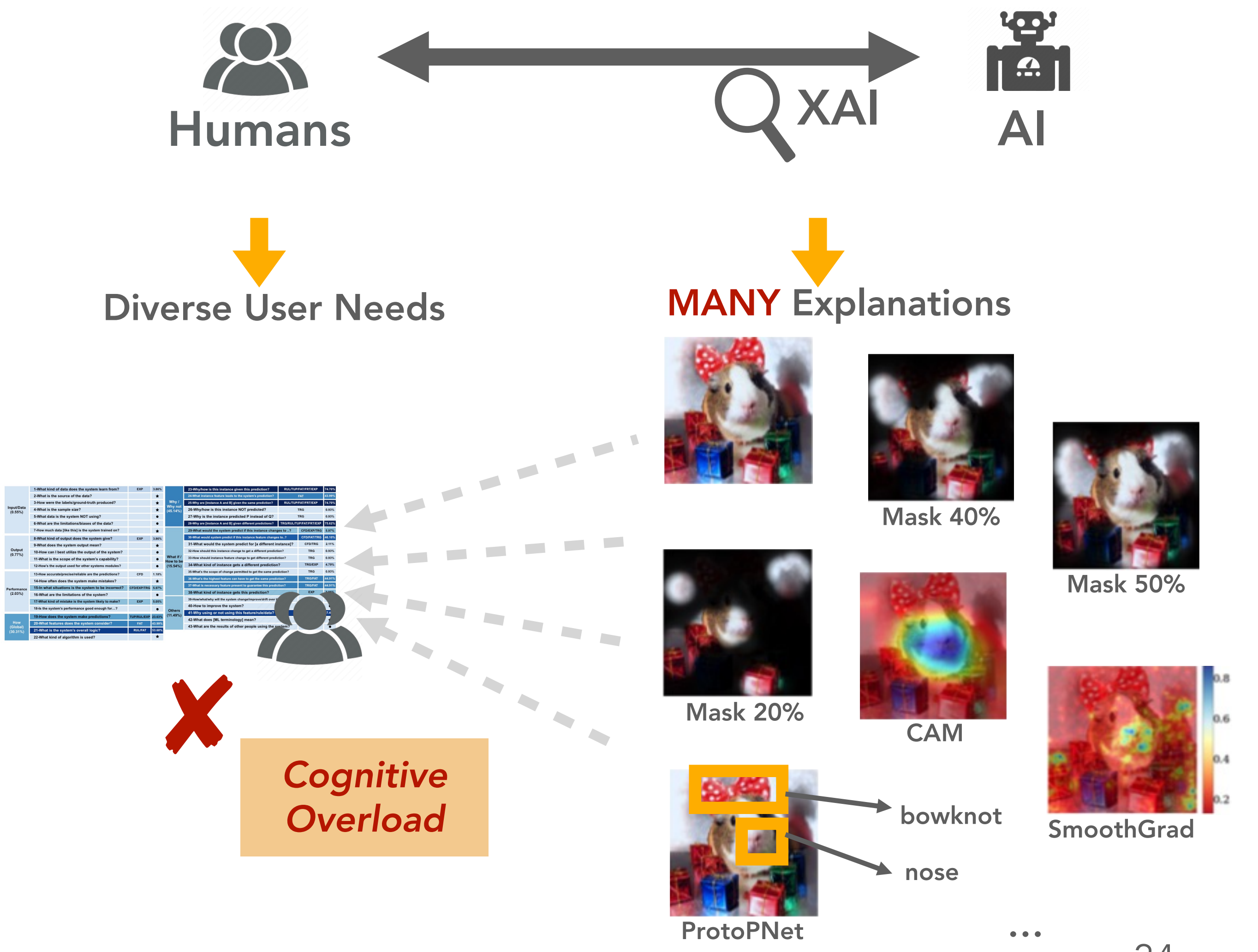
✗

Needs are NOT satisfied

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

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

Challenges of Existing XAI

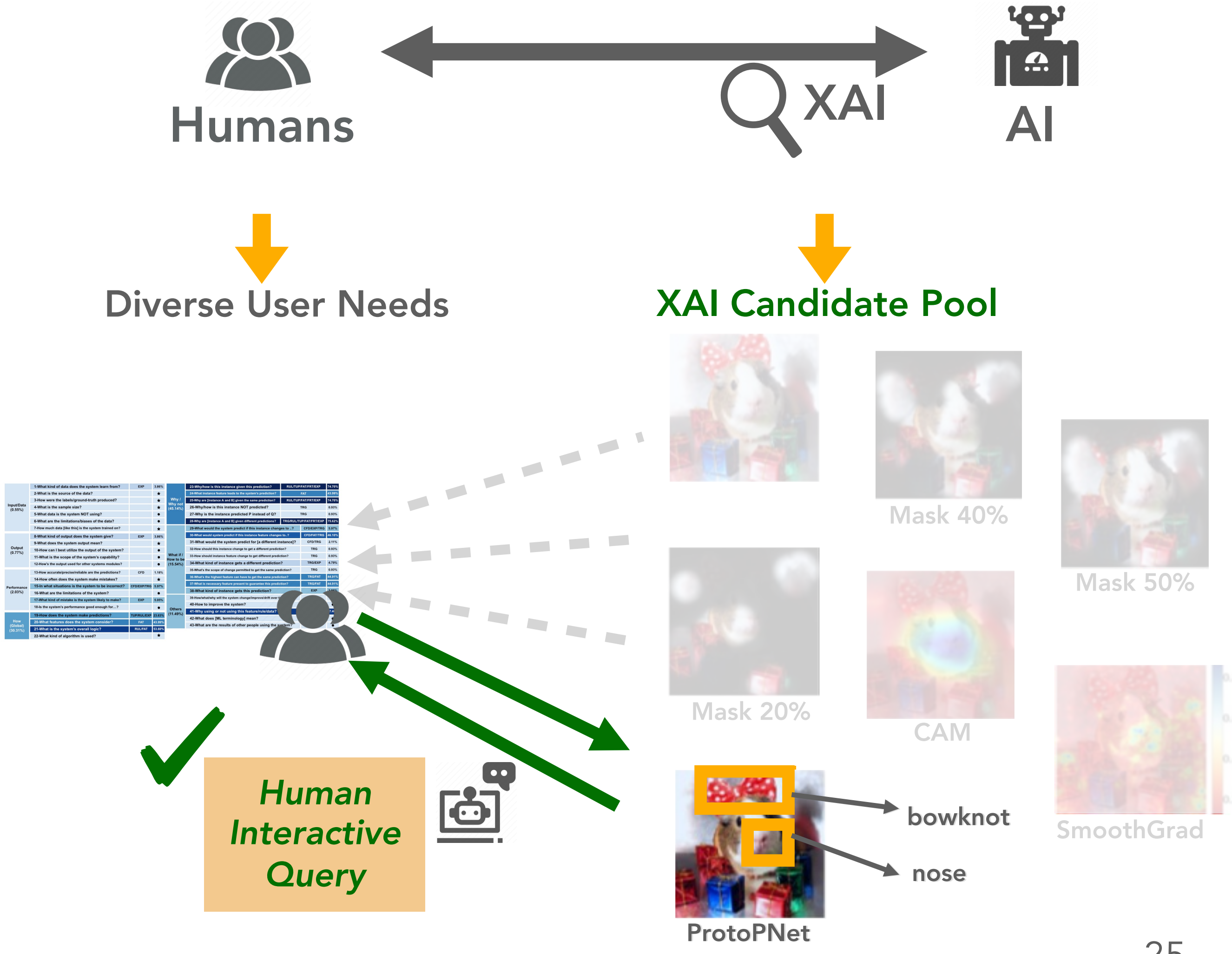


Input/Data (0.55%)	1-What kind of data does the system learn from?	EXP	3.86%
	2-What is the source of the data?		★
	3-How were the labels/ground-truth produced?		★
	4-What is the sample size?		★
	5-What data is the system NOT using?		●
	6-What are the limitations/biases of the data?		●
	7-How much data [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	[illegible]
Others (11.49%)	39-How/what/why will the system change/improve/drift over [illegible]	[illegible]	[illegible]
	40-How to improve the system?	[illegible]	[illegible]
	41-Why using or not using this feature/rule/data?	[illegible]	[illegible]
	42-What does [ML terminology] mean?	[illegible]	[illegible]
	43-What are the results of other people using the system?	[illegible]	●

- 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

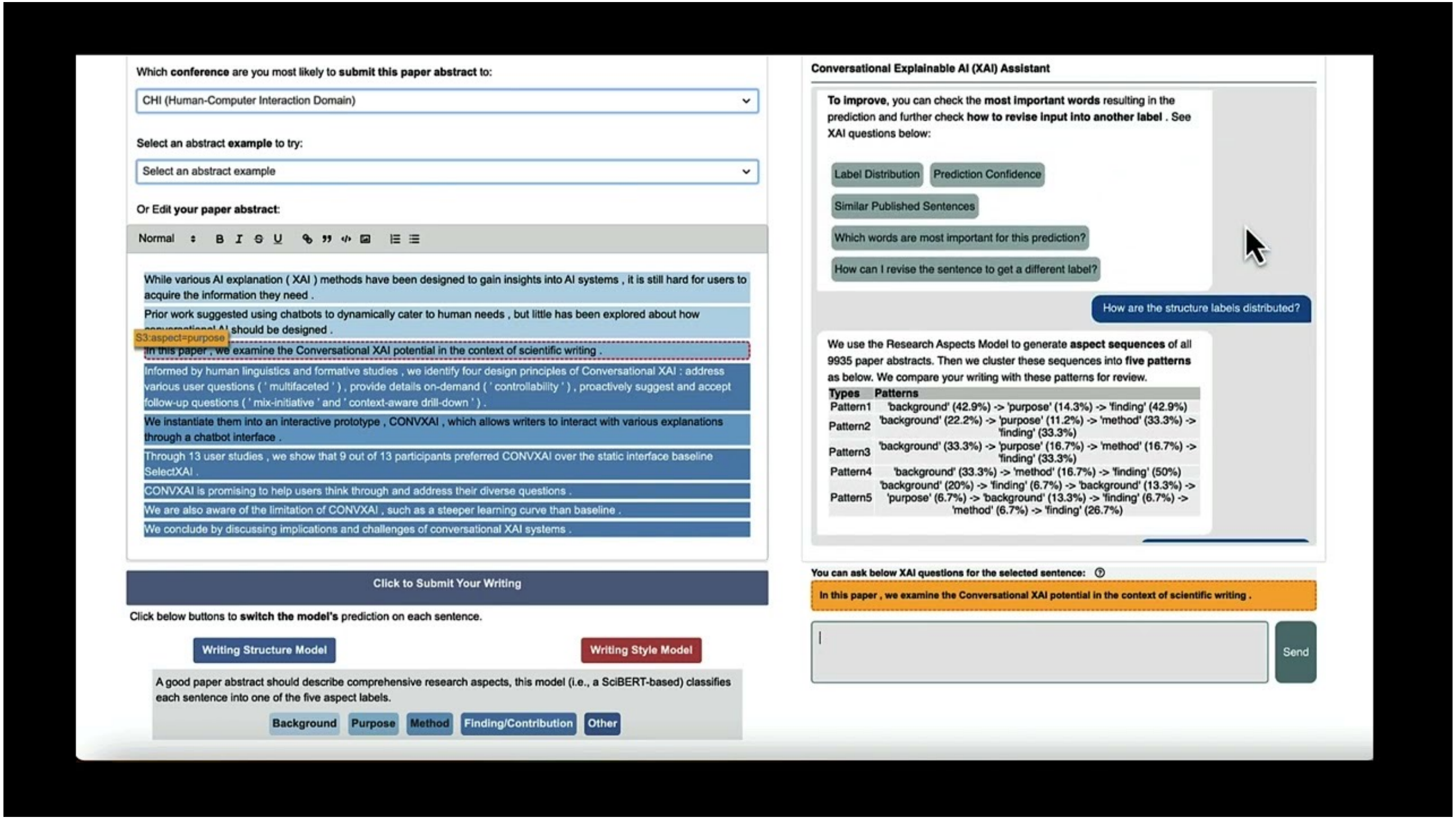


Input/Data (0.55%)	1-What kind of data does the system learn from?	EXP	3.86%
	2-What is the source of the data?		★
	3-How were the labels/ground-truth produced?		★
	4-What is the sample size?		★
	5-What data is the system NOT using?		●
	6-What are the limitations/biases of the data?		●
	7-How much data [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 t		
	40-How to improve the system?		●
	41-Why using or not using this feature/rule/data?		7.4
	42-What does [ML terminology] mean?		★
	43-What are the results of other people using the system?		●

- 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.

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**:

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

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

S3:aspect=purpose

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

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

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

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

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

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

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

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

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.

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 .

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

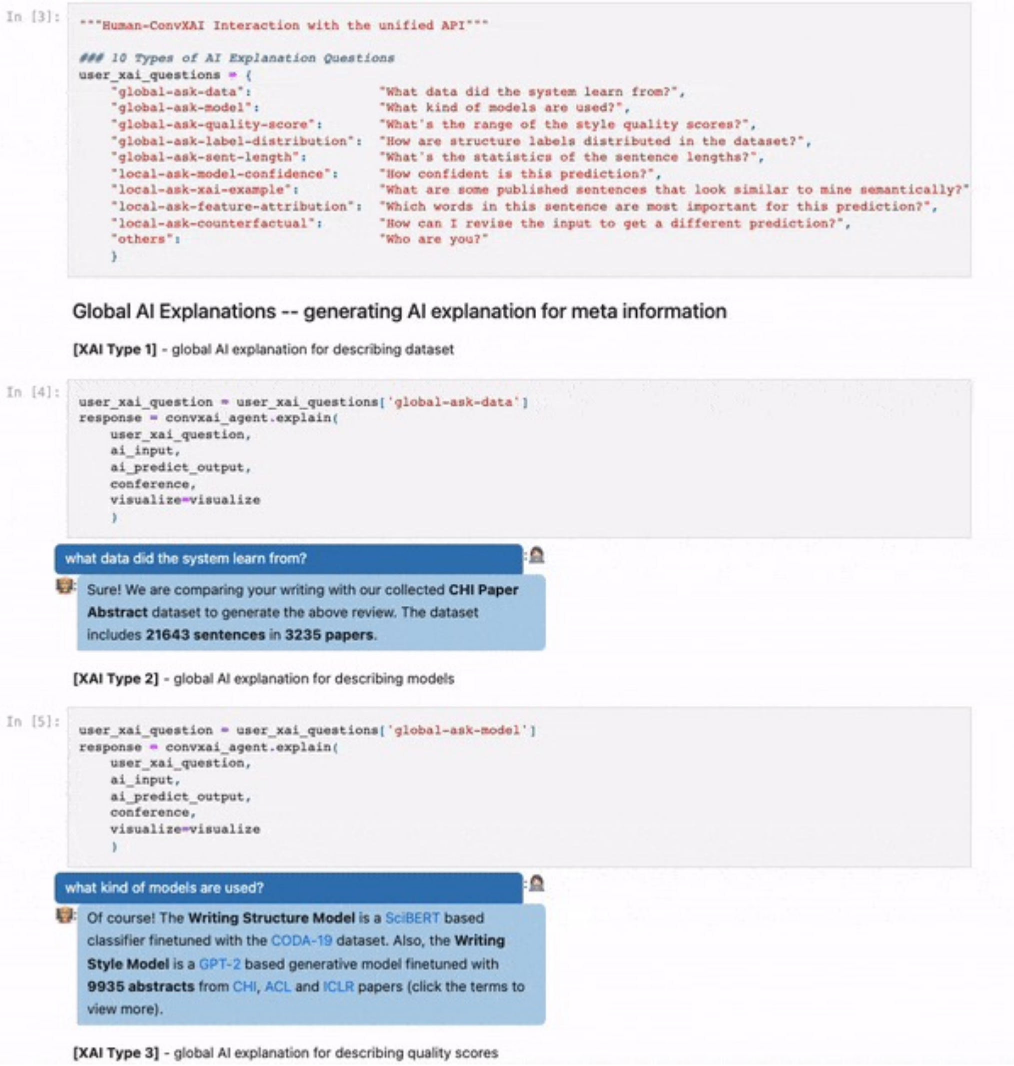
Technical Challenges & Contributions

Challenges:

1. **No unified approach** for **various XAI**
2. No **dialog system** to parse **XAI** user **questions** and **customization**

Technical Contribution

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



In [3]:
```
"""Human-ConvXAI Interaction with the unified API"""

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

Global AI Explanations -- generating AI explanation for meta information

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

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

what data did the system learn from?

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

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

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

what kind of models are used?

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

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

Evaluate ConvXAI with real human studies

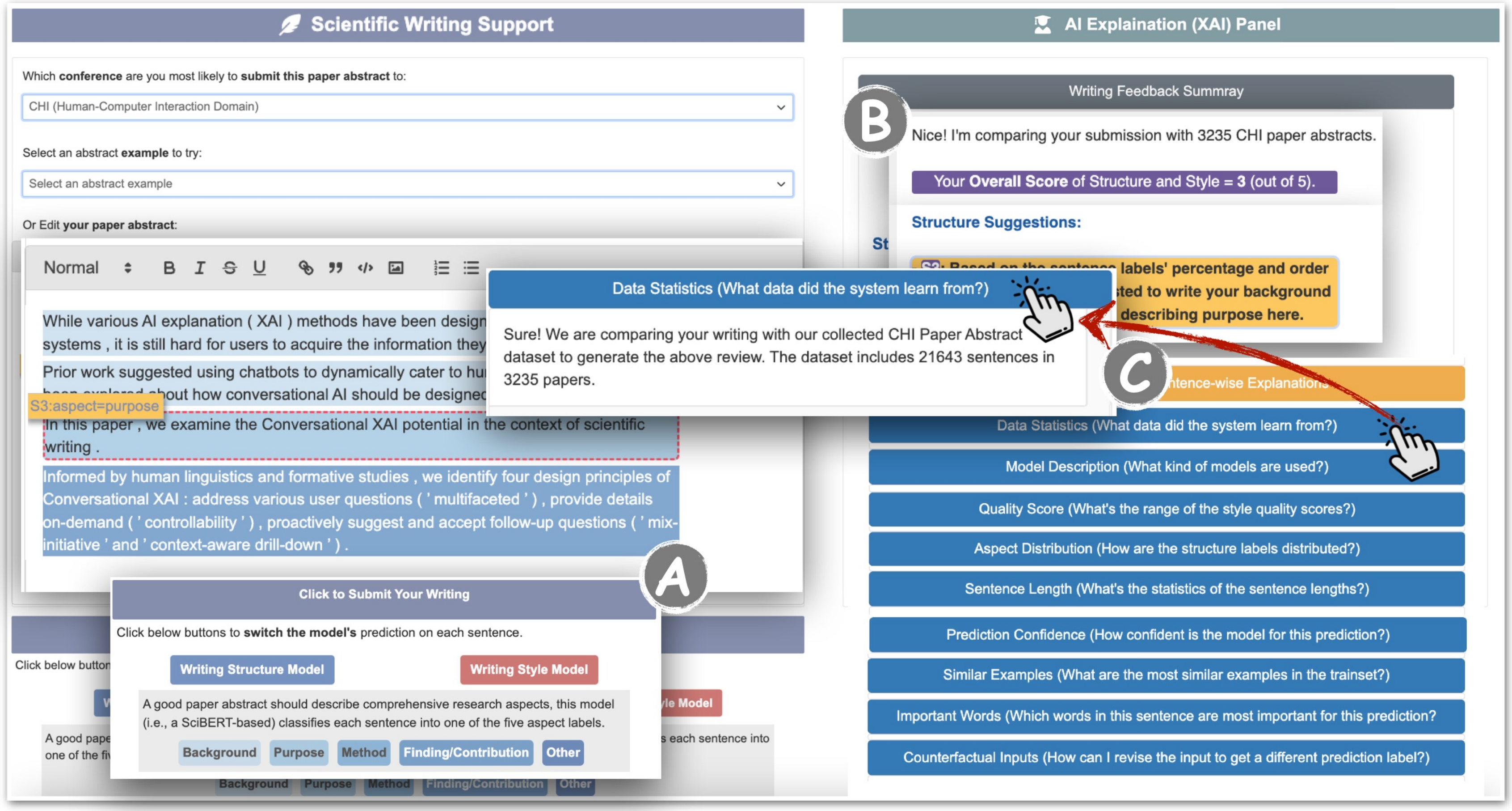
	Task1	Task2
Who is studies	**13** graduate researchers	**8** researchers
When	09/2022 (90min)	12/2022 (90min) (**rejoin**)

How it's studied

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**

Baseline System (SelectXAI)

Within-Subjects Study Design



Scientific Writing Support

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

CHI (Human-Computer Interaction Domain)

Select an abstract **example** to try:

Select an abstract example

Or Edit **your paper abstract:**

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

AI Explaination (XAI) Panel

Writing Feedback Summray

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 … suggested to write your background … describing purpose here.

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

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

…ntence-wise Explanations

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

Model Description (What kind of models are used?)

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

Aspect Distribution (How are the structure labels distributed?)

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

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

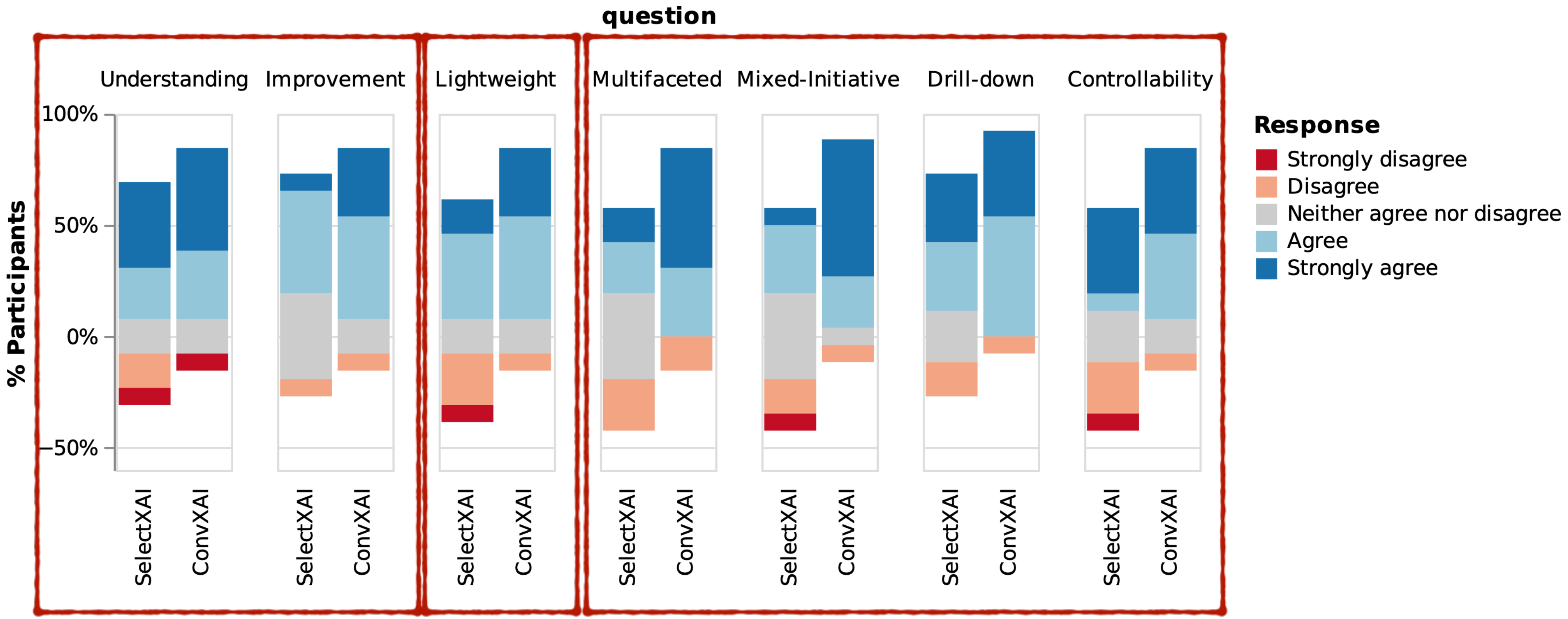
Similar Examples (What are the most similar examples in the trainset?)

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

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

Survey results of human study in Task1

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



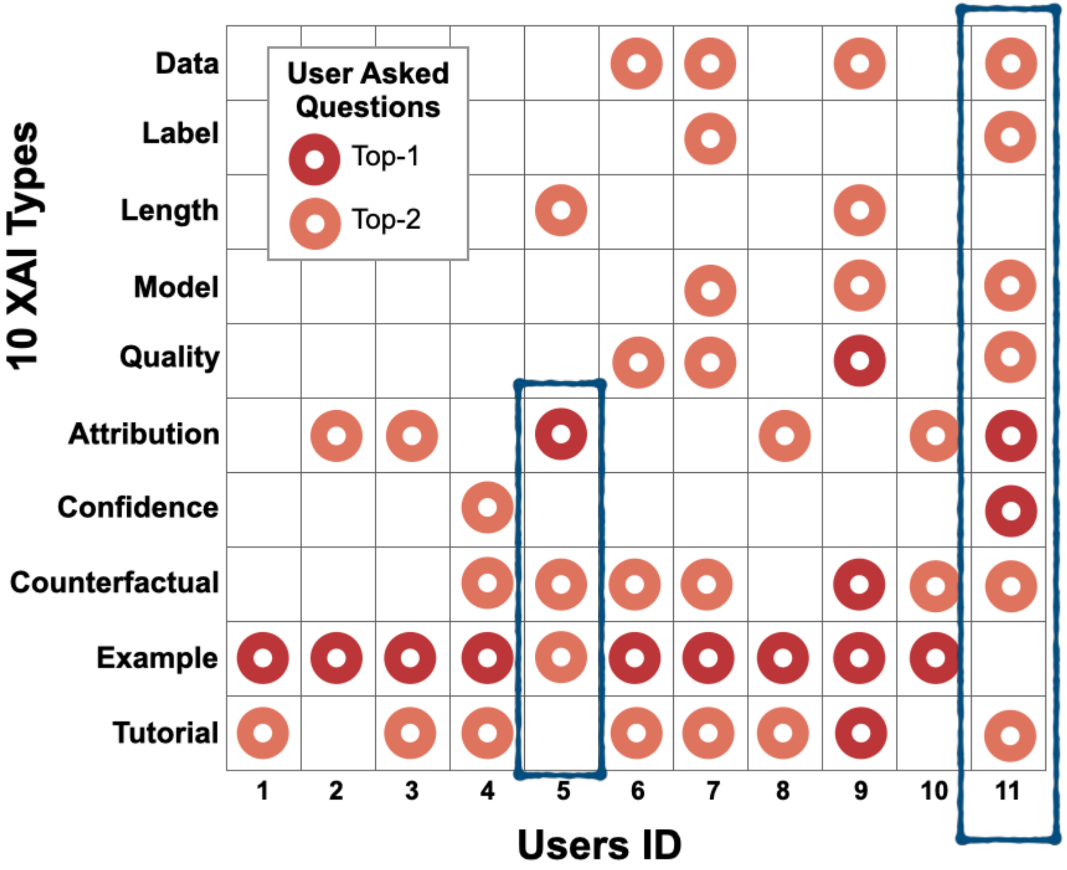
Useful in Understanding & Improving Writing

Less Cognitive Load

More aligned with human-centered design rationales

Humans’ XAI usage patterns in Task1

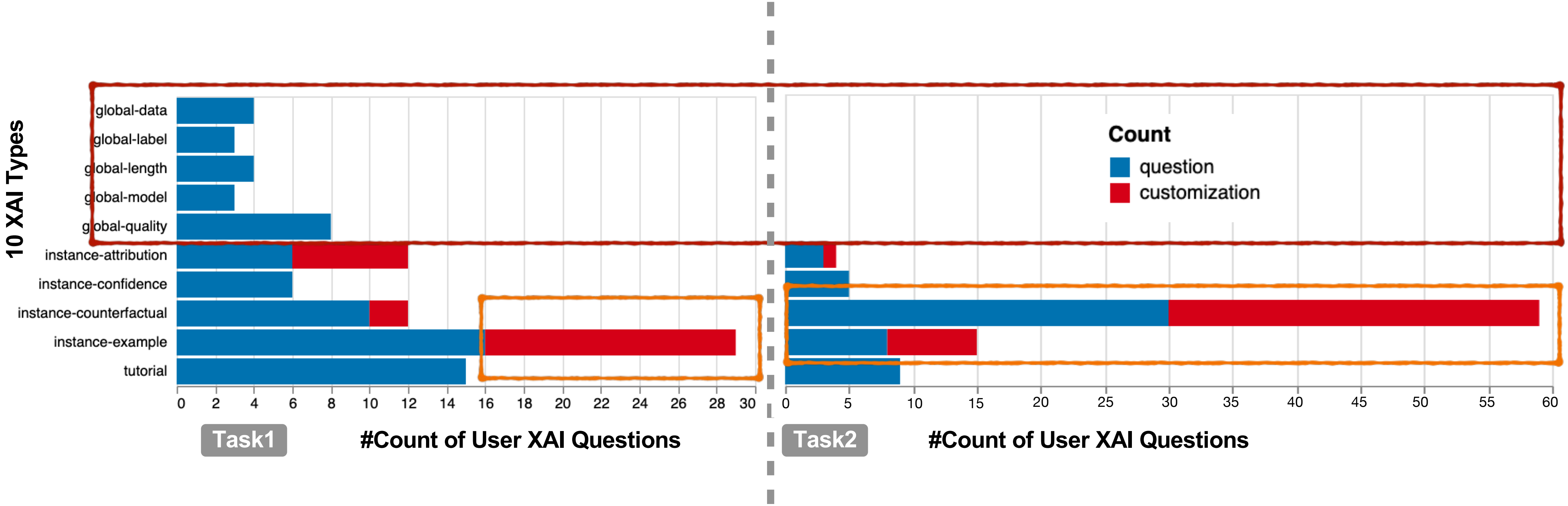
Finding#2: Different users prefer to use different XAI formats in the real-world tasks.



10 XAI Types	1	2	3	4	5	6	7	8	9	10	11
Data						○	○		○		○
Label							○				○
Length					○				○		
Model							○		○		○
Quality						○	○		●		○
Attribution		○	○		●			○		○	●
Confidence				○							●
Counterfactual				○	○	○	○		●	○	○
Example	●	●	●	●	○	●	●	●	●	●	
Tutorial	○		○	○		○	○	○	●		○

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

Finding#3: Users XAI needs changed along time and converged to instance-wise XAIs.

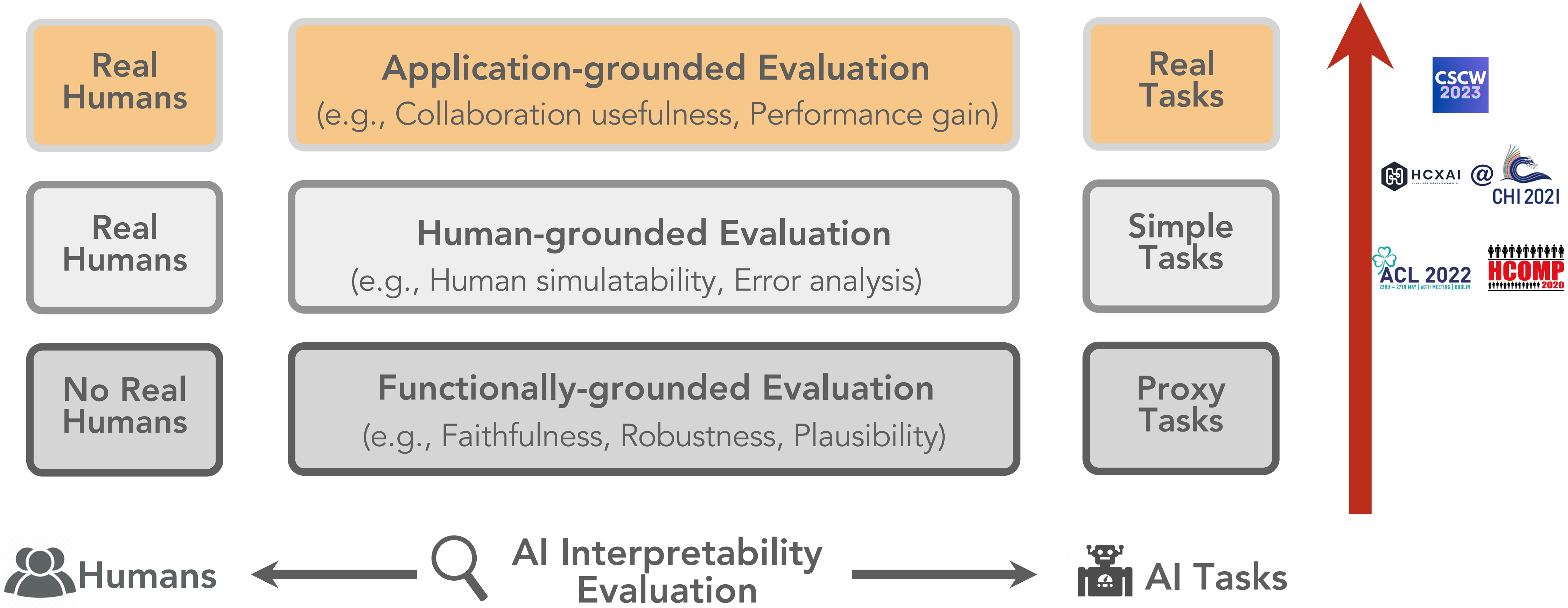


Finding#4: User-oriented XAI Customization is important in many XAI types.

Takeaway

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

Human-Centered XAI Usefulness

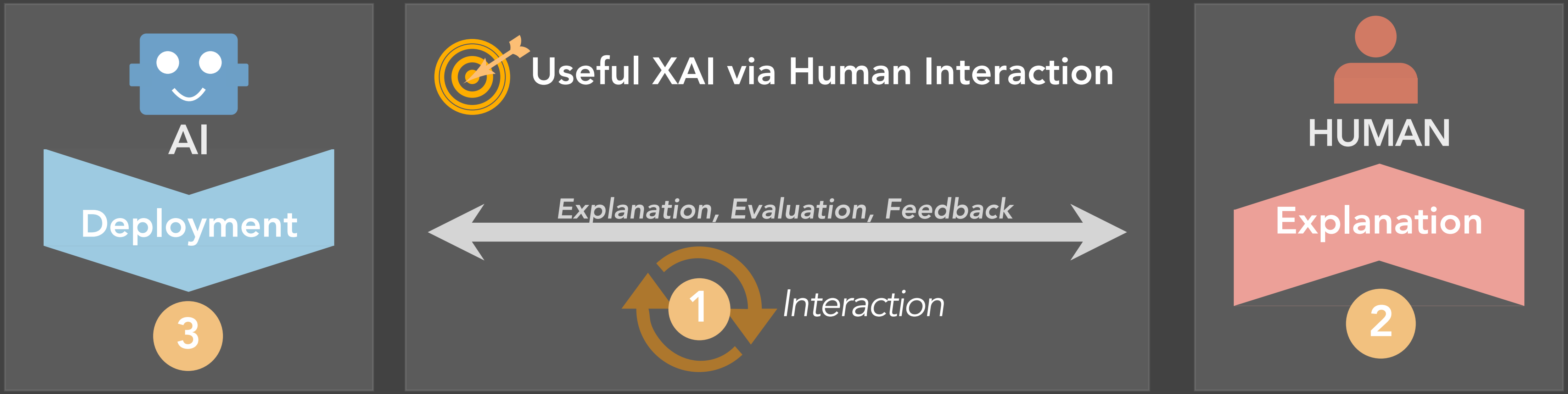


ConvXAI for Human-Centered Useful XAI

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

What’s *Next* …


AI
Deployment
3
Useful XAI via Human Interaction
Explanation, Evaluation, Feedback
1
Interaction
HUMAN
Explanation
2

ConvXAI: *A Start of Useful XAI for Humans*

1. ***Tools*** — *How to construct scalable interactive/conversational XAI tools for a wider range of human-AI collaboration tasks?*
2. ***Useful for Humans*** — *How to measure usefulness for humans and tailor interactive XAI to improve human performance?*
3. ***Useful for AIs*** — *How to collect human feedback from interactive XAI to improve AI model performance?*

Other **Human-centered AI** papers (2020 - 2023)

1. **Hua Shen**, Vicky Zayats, Johann Rocholl, Dan Walker, and Dirk Padfield. MultiTurnCleanup: A Benchmark for Multi-Turn Spoken Conversational Transcript Cleanups. EMNLP 2023 🏆 **Google Research Scholarships**
2. **Hua Shen**, Chieh-Yang Huang, Tongshuang Wu, Ting-Hao (Kenneth) Huang. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing. CSCW 2023 Demo.🏆 **Best Demo Award**
3. Tongshuang Wu, **Hua Shen**, Daniel S Weld, Jeffrey Heer, Marco Tulio Ribeiro. ScatterShot: Interactive In-context Example Curation for Text Transformation. IUI 2023. 🏆 **Best Paper Honorable Mention**
4. **Hua Shen***, Adaku Uchendu*, Jooyoung Lee*, Thai Le, Ting-Hao'Kenneth'Huang, and Dongwon Lee. Does Human Collaboration Enhance the Accuracy of Identifying Deepfake Texts? AAAI HCOMP 2023
5. **Hua Shen**, Tongshuang Wu. Parachute: Evaluating Interactive Human-LM Co-writing Systems. CHI 2023 In2Writing Workshop
6. **Hua Shen**, Tongshuang Wu, Wenbo Guo, Ting-Hao (Kenneth) Huang. Are Shortest Rationales the Best Explanations For Human Understanding? ACL 2022
7. Binfeng Xu, Xukun Liu, **Hua Shen**, Zeyu Han, Yuhan Li, Murong Yue, Zhiyuan Peng, Yuchen Liu, Ziyu Yao, Dongkuan Xu. Gentopia.AI: A Collaborative Platform for ToolAugmented LLMs. EMNLP 2023 Demo
8. **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.
9. Shih-Hong Huang, Chieh-Yang Huang, Yuxin Deng, **Hua Shen**, Szu-Chi Kuan, and TingHao'Kenneth'Huang. Too Slow to Be Useful? On Incorporating Humans in the Loop of Smart Speakers. AAAI HCOMP 2022 WiP/Demo
10. **Hua Shen**, Ting-hao (Kenneth) Huang. Explaining the Road Not Taken. CHI 2021 HCXAI Workshop
11. **Hua Shen**, Ting-hao (Kenneth) Huang. How Useful Are the Machine-Generated Interpretations? A Human Evaluation on Guessing the Wrongly Predicted Labels. AAAI HCOMP 2020
12. Xinyang Zhang, Ningfei Wang, **Hua Shen**, Shouling Ji, Ting Wang. Interpretable Deep Learning under Fire. USENIX 2020
13. 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

14. 15. CHI 2024 Under Review……

Keywords

Human-annotated AI dataset

Conversational XAI for Human

Human-AI Interactive System

Human Evaluation on LLM

Human-AI Co-writing Eval

Human Eval on NLP XAI

Human-AI Agent Interact Tool

Fairness on Speaker Verification

Human-in-the-loop Speech

Survey of 200+ XAI Papers

Human Eval on CV XAI

XAI Robustness

AI Adversarial & Security

Acknowledgment!!

Advisors & Ph.D. Committee

RisingStar Mentor & Organizers

Mentors & Collaborators

thank you!

Hua Shen

huashen@umich.edu @huashen218

University of Michigan