

Towards Useful Al Interpretability for Humans via Interactive Al Explanations

Hua Shen

♥ @huashen218 M huashen218@psu.edu

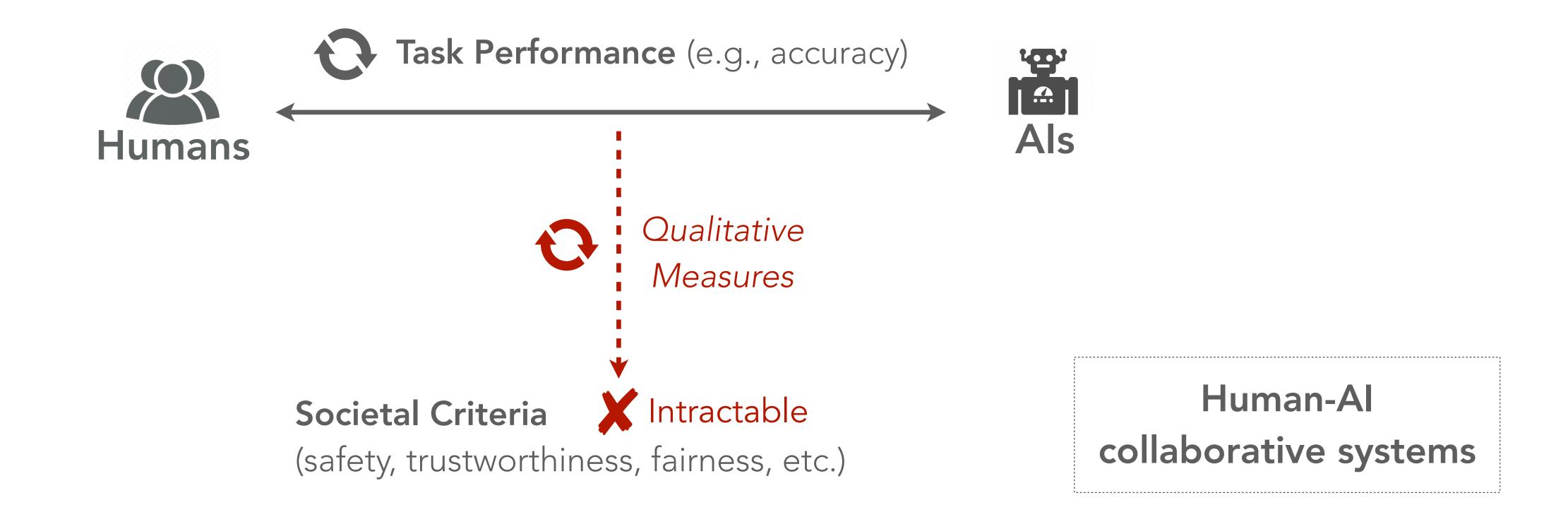
Penn State University

Background & Motivation



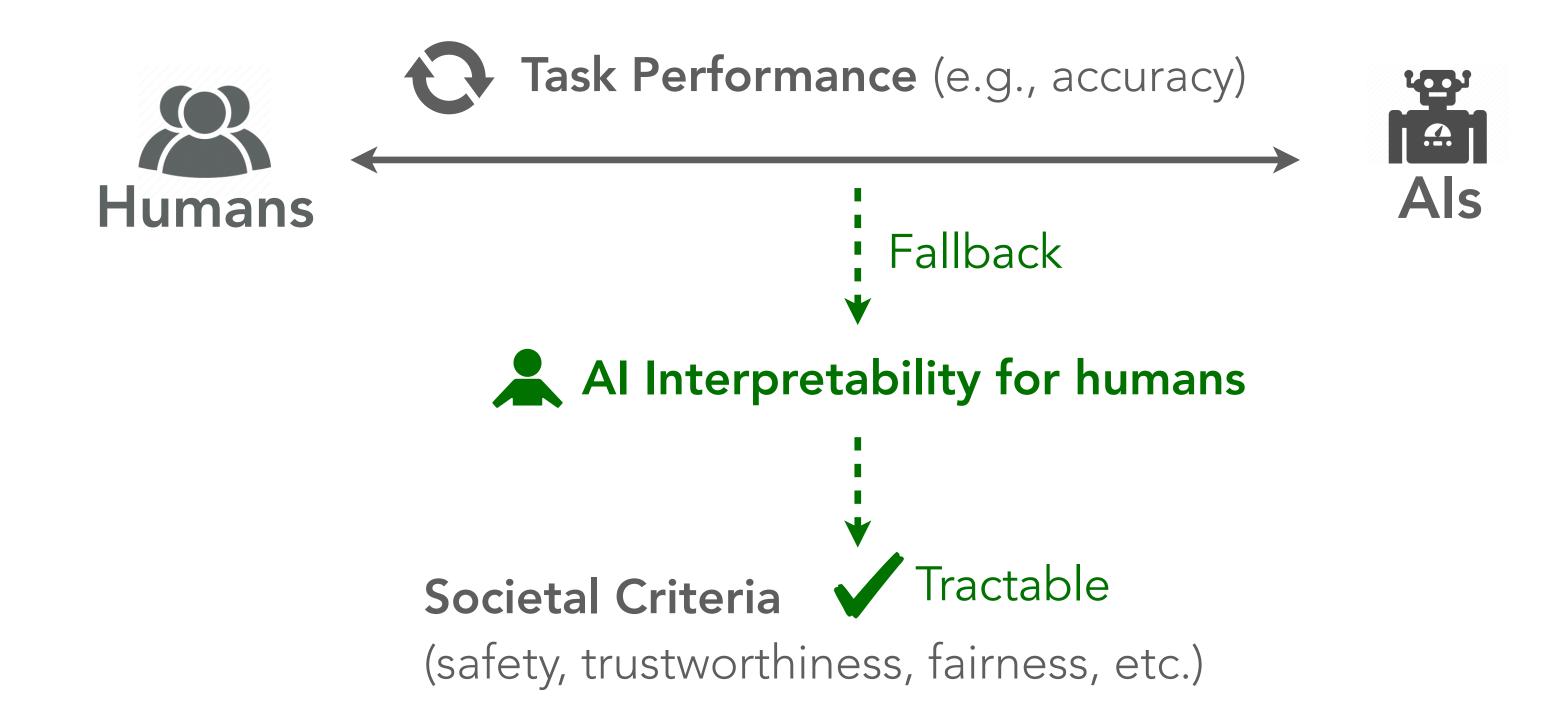
Why do we need Al interpretability?

Human-Al 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.).

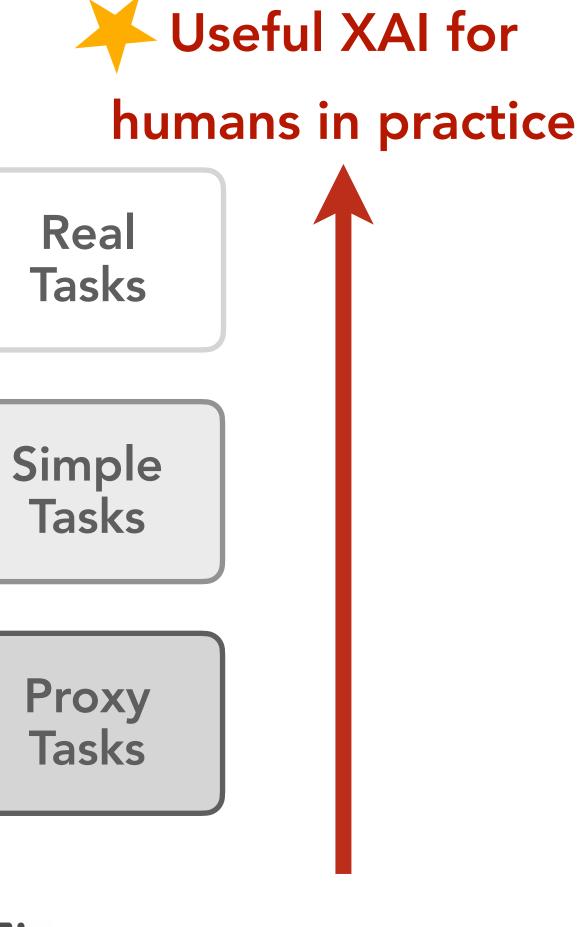


The usefulness of XAI for humans is crucial

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



Evaluation of XAI usefulness



Real Humans

Real

Humans

Humans

Application-grounded Evaluation

(e.g., Collaboration usefulness, Performance gain)

Human-grounded Evaluation

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

No Real Functionally-grounded Evaluation

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

Humans

Al Interpretability Evaluation

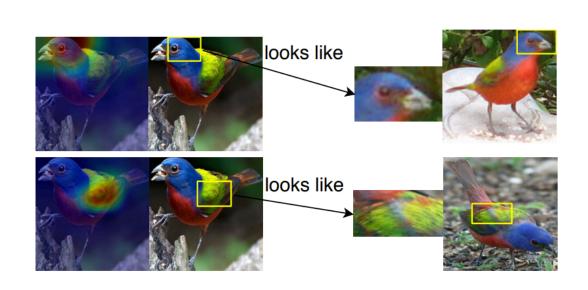
(Doshi-Velez, Finale, & Kim, Been. (2017))

Al Tasks

Trajectory of XAI algorithms



Attribution Heatmaps



(e.g. ProtoPNet) **Prototype Explanations**



(e.g. Influence Functions, Representor Point) **Training Examples**

NEGATIVE

train id4642

cow predicted as ox





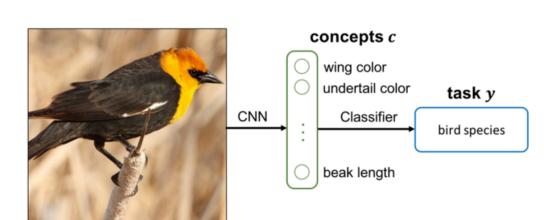
(e.g. SCOUT, Triggers)

Counterfactuals

2012

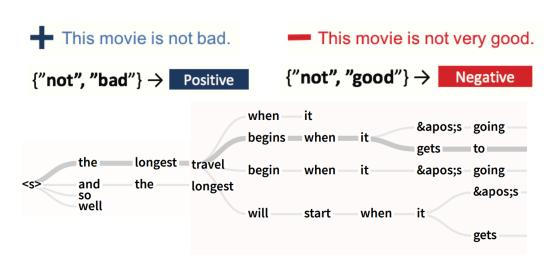
Concept-based Explanations

(e.g. TCAV, Concept Bottleneck)



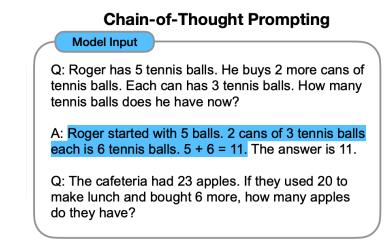
Decision Rules / Graphs

(e.g. Anchors)



Free-text Rationales

(e.g. Chain-of-thought)

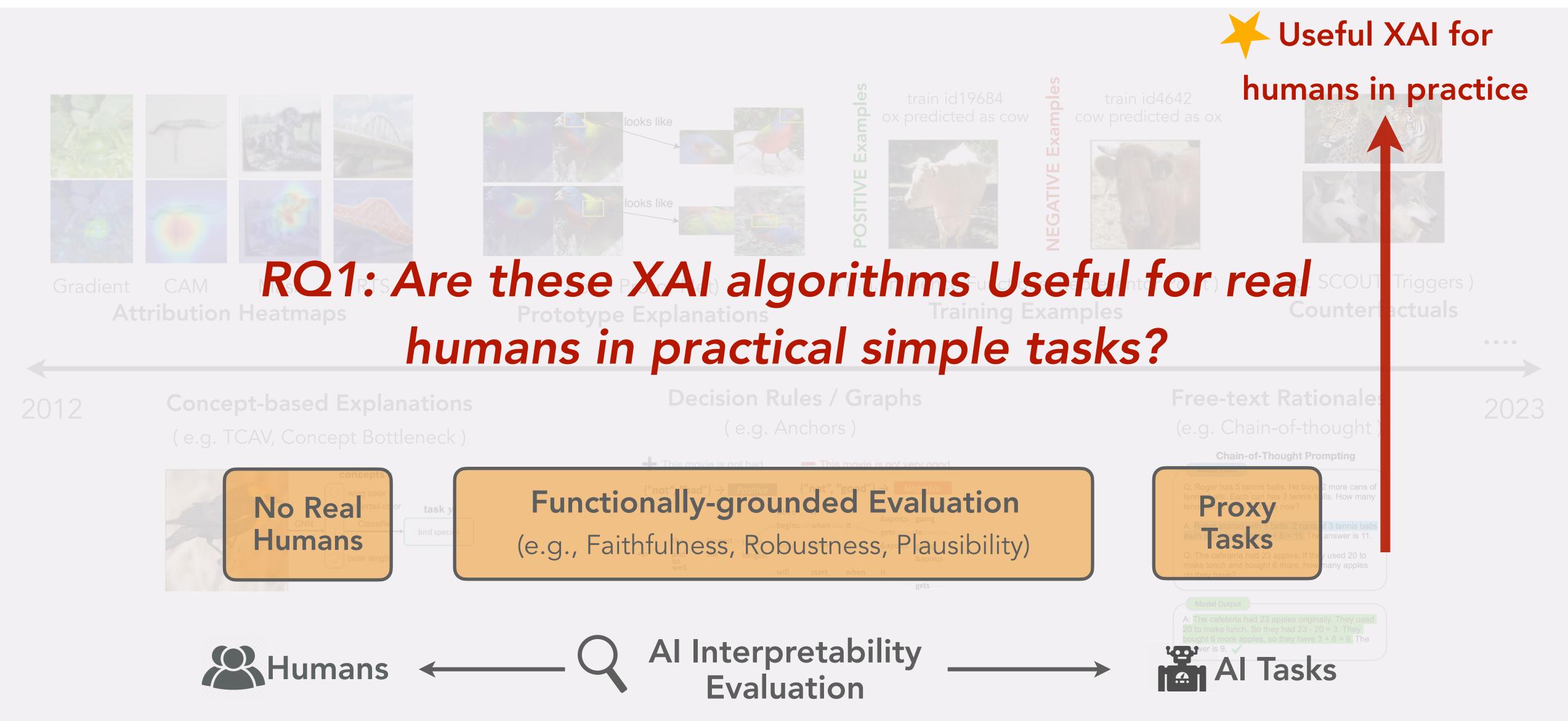


A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

2023

 \bullet

Under-Explored: human evaluation of XAI usefulness



Background & Motivation

Useful XAI

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:



Input Image



Guess which label the model incorrectly predicted?

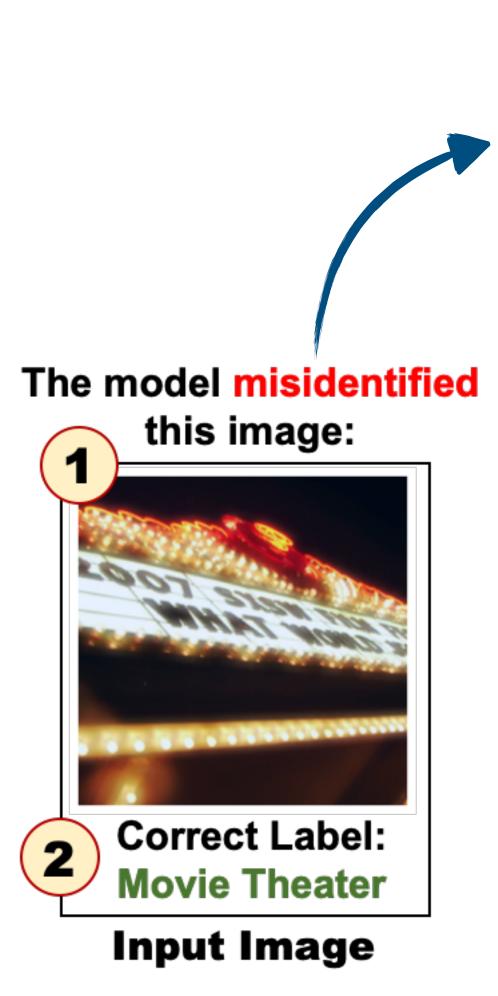
- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web

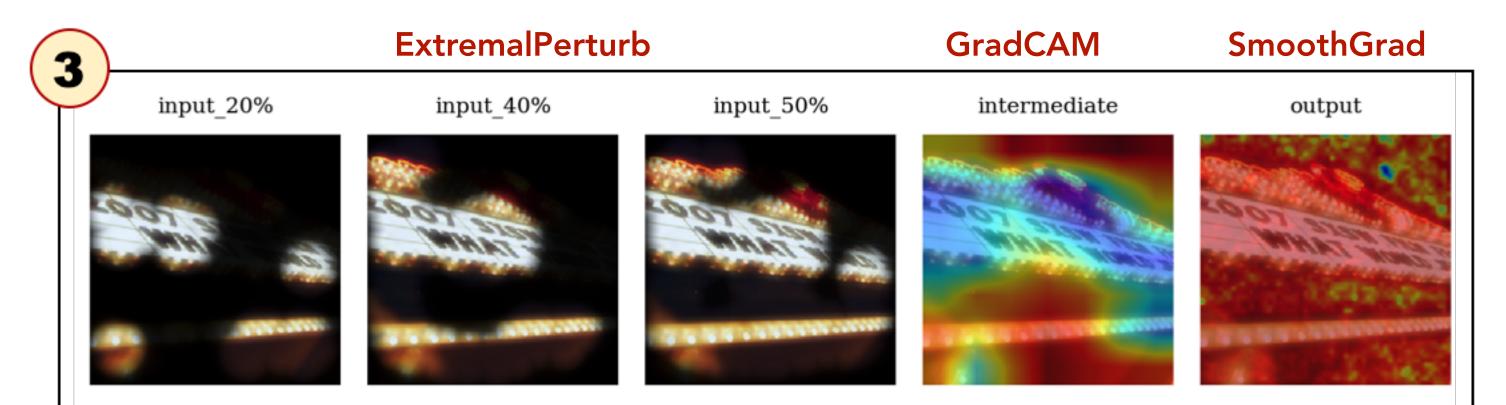


Multiple Choice Question



What Al explanations are used?





Machine-Generated Interpretations (Int)

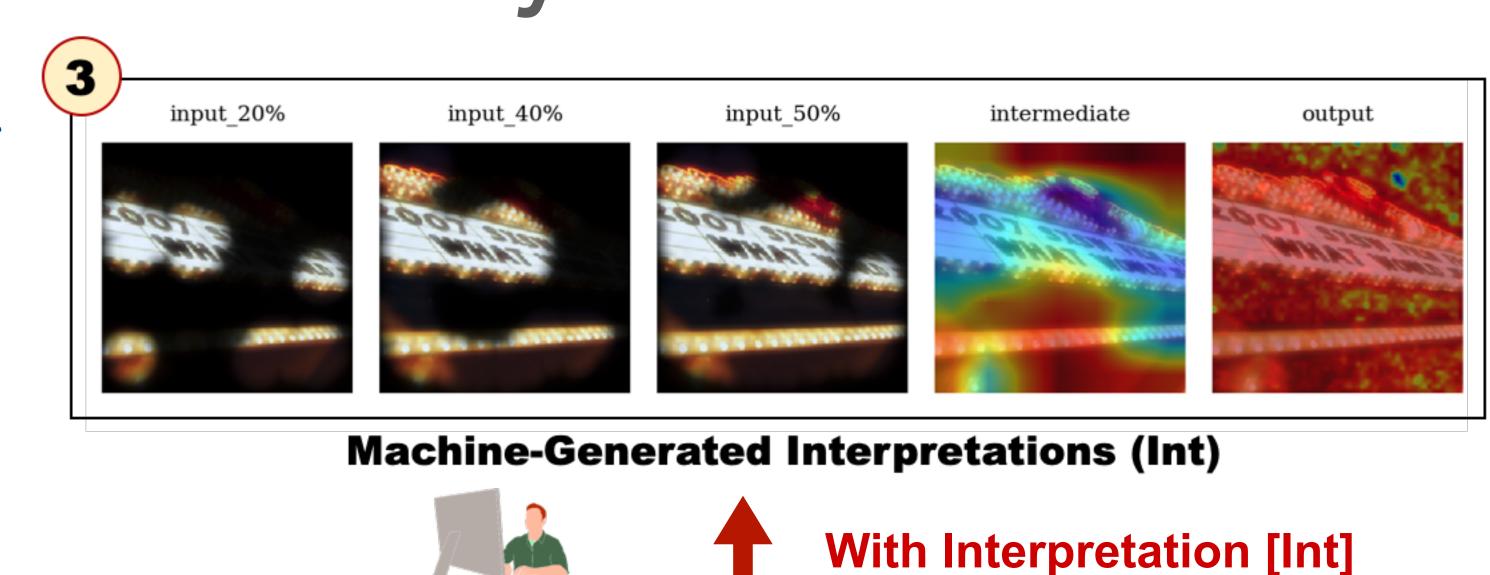
Widely-adopted Saliency Maps as Al Explanations

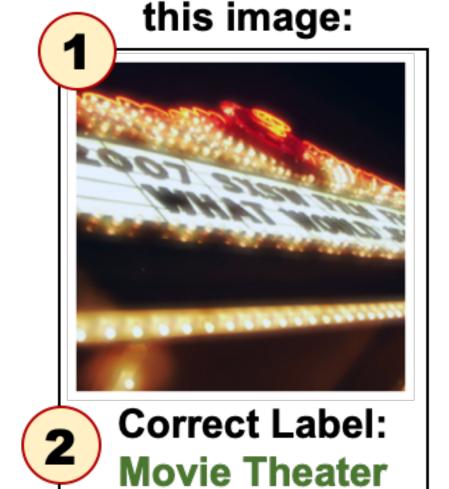


- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web

Multiple Choice Question

Design of Human Study





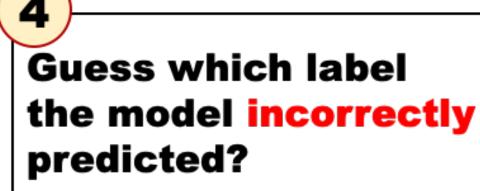
The model misidentified

Input Image



Crowd Worker

1000 submissions



- Fireboat
- Malinois
- Carousel
- Garfish
- Spider web

Multiple Choice Question





Without Interpretation [No-Int]

Results

	C 1	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)

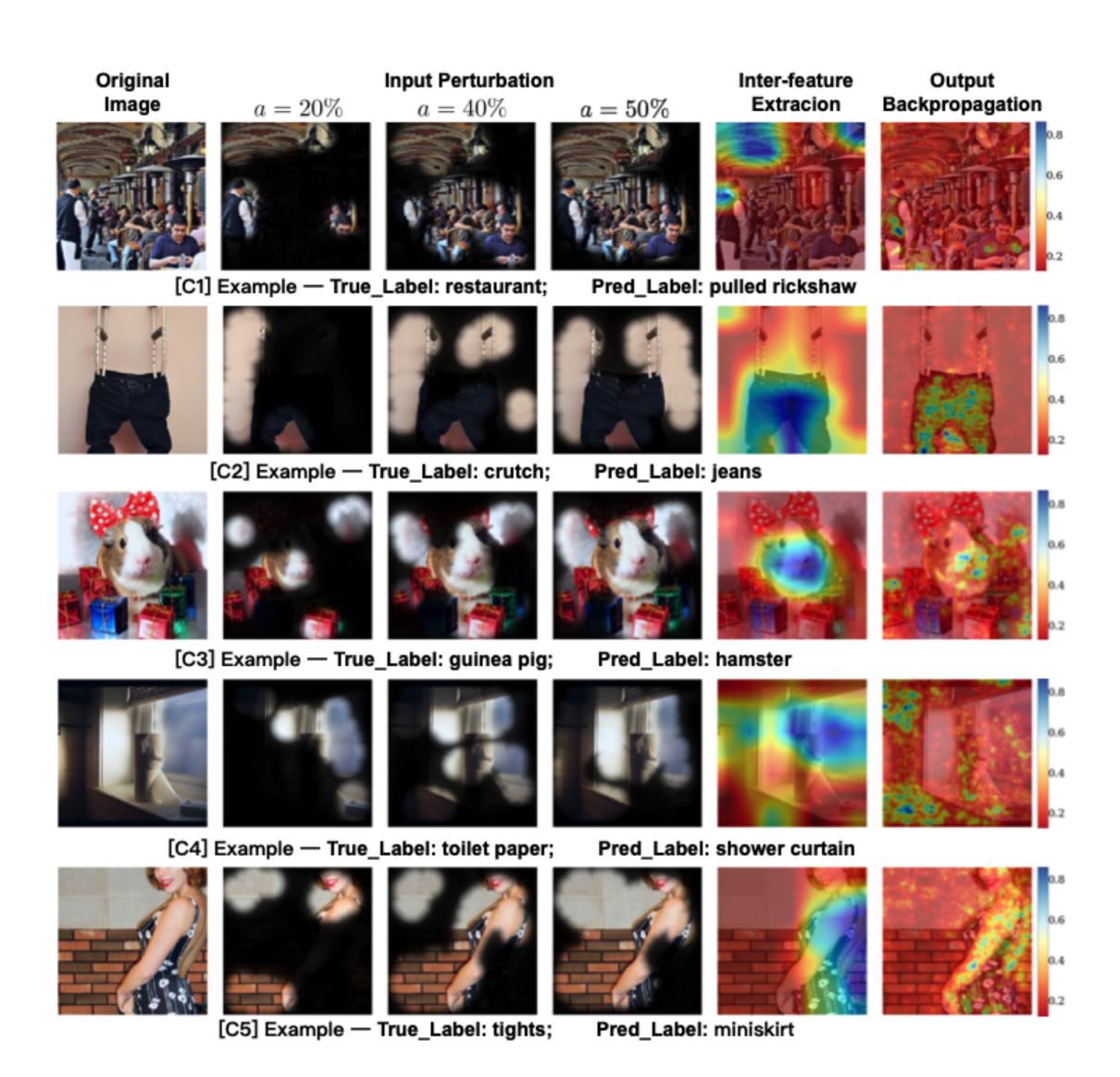
10% Accuracy Drop

Key Findings

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



Model error categories for fine-grained analysis



•C1: Local Character Inference

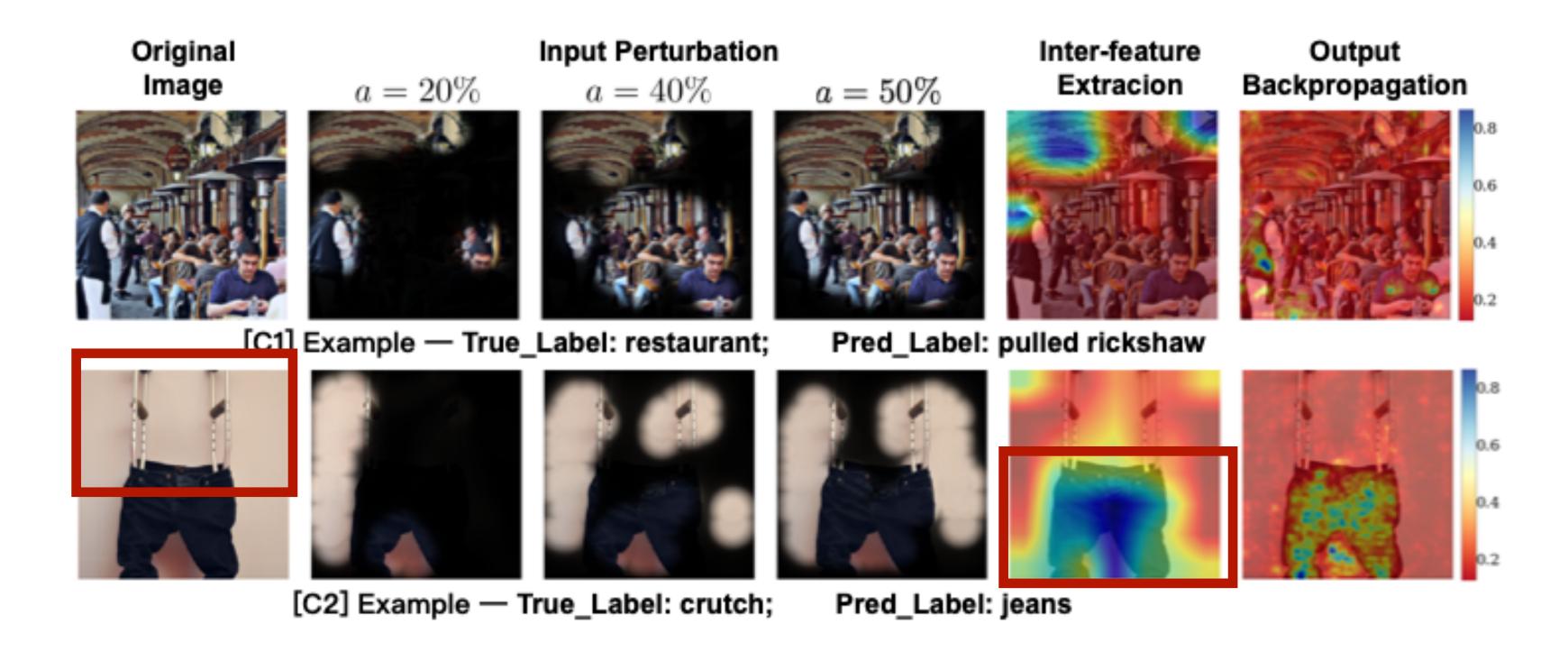
C2: Multiple Objects Selection

•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).

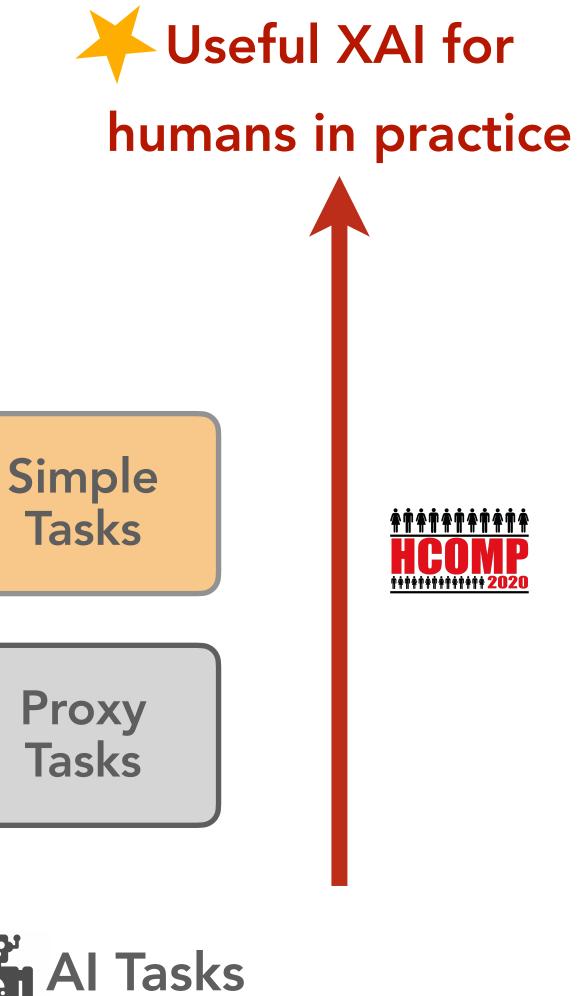
C1: Local Character Inference

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





Human-grounded Evaluation

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

No Real Humans Functionally-grounded Evaluation

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





Al Interpretability **Evaluation**



Background & Motivation

for RQ1: Are XAI Useful Humans



for Humans?



Are Shortest Rationales the Best Explanations for Human Understanding?



Hua Shen



Sherry Wu



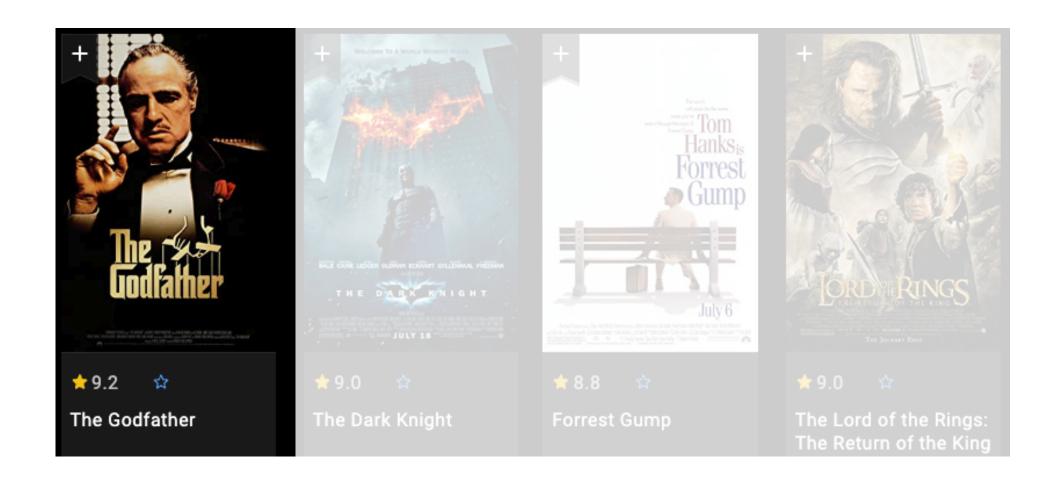
Wenbo Guo



Kenneth Huang

Rationales in Text Classification Tasks

Sentiment Analysis For Movie Reviews



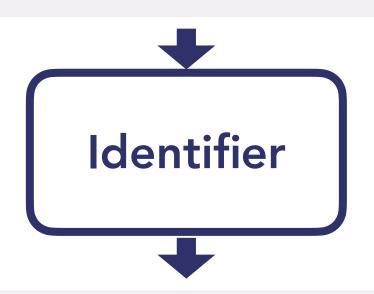


Explanations:

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

Self-Explaining Models

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



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

Explanations



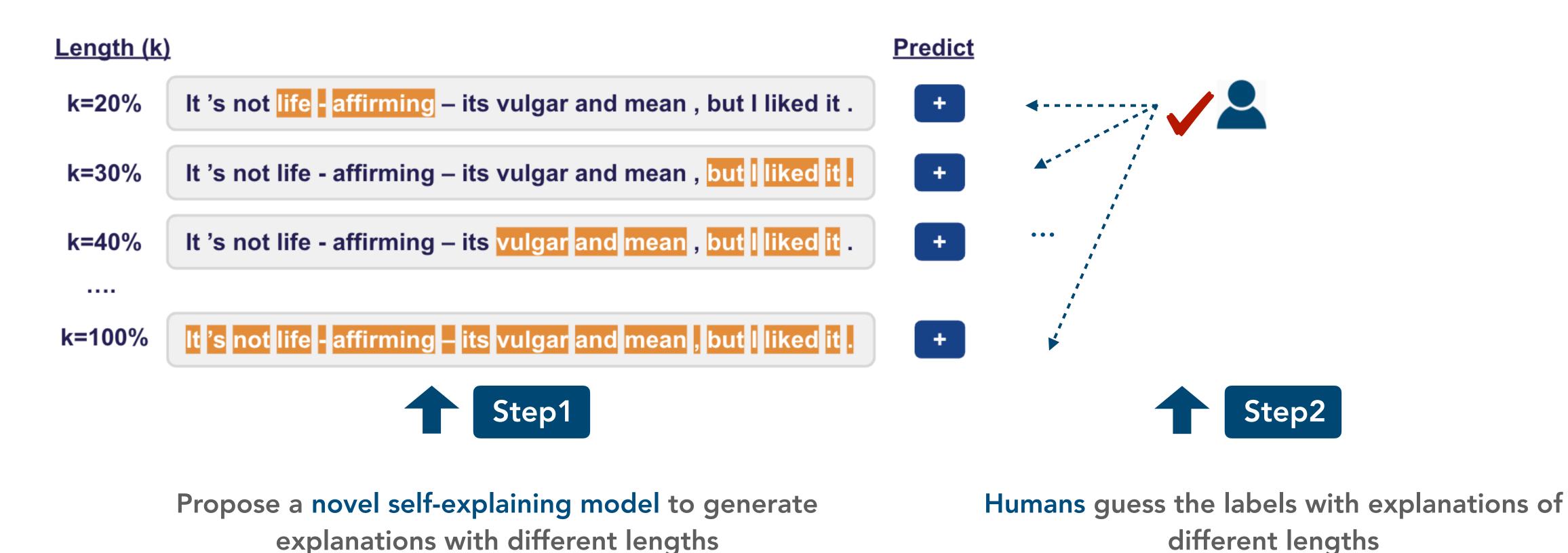
Shorter Explanations are Better.

? Yet to be validated by human studies!

Are Shortest Explanations the Best for Human Understanding?

Overview of Study Design

Goal: the impact of explanation lengths on human understanding:

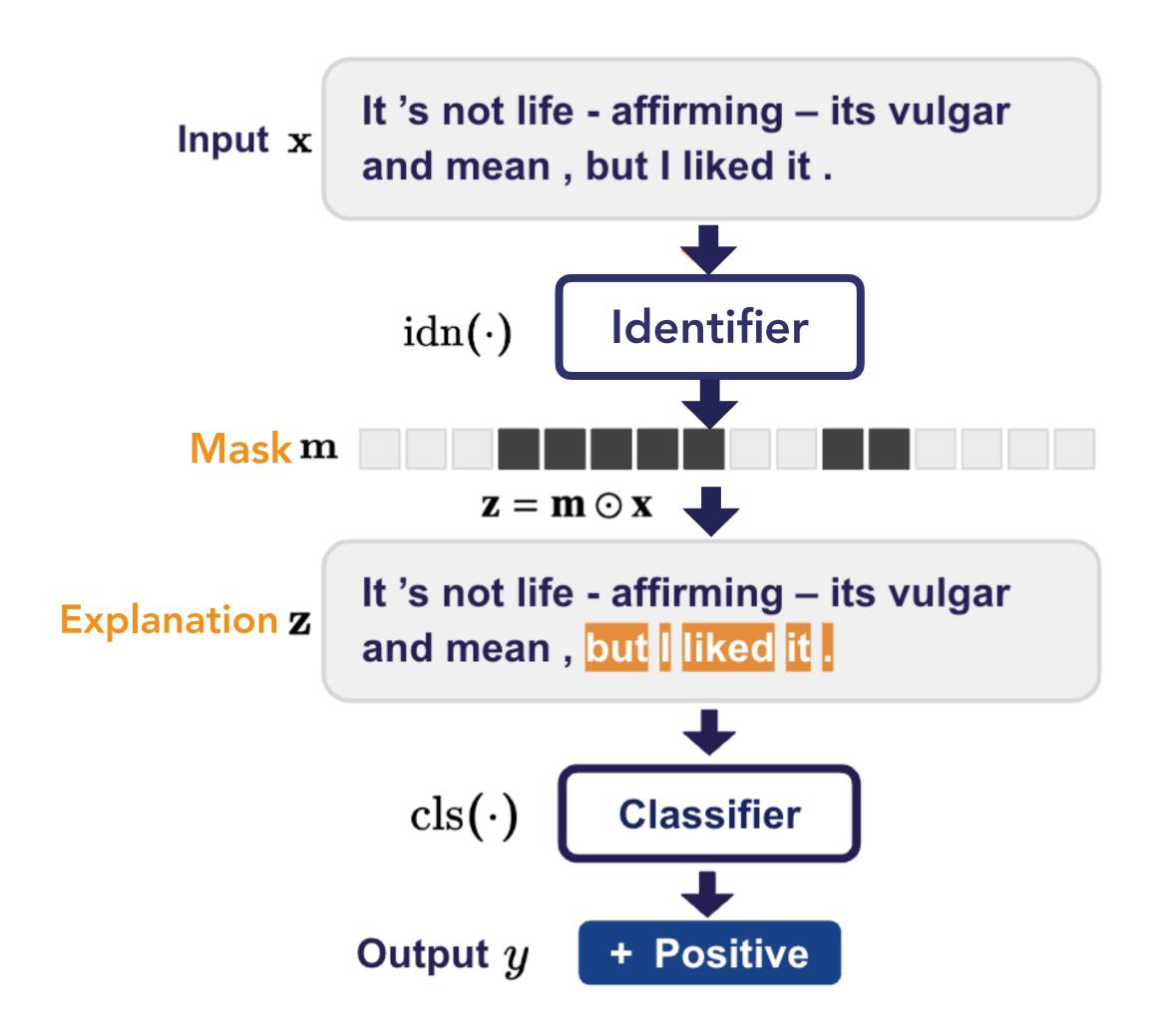


Contribution

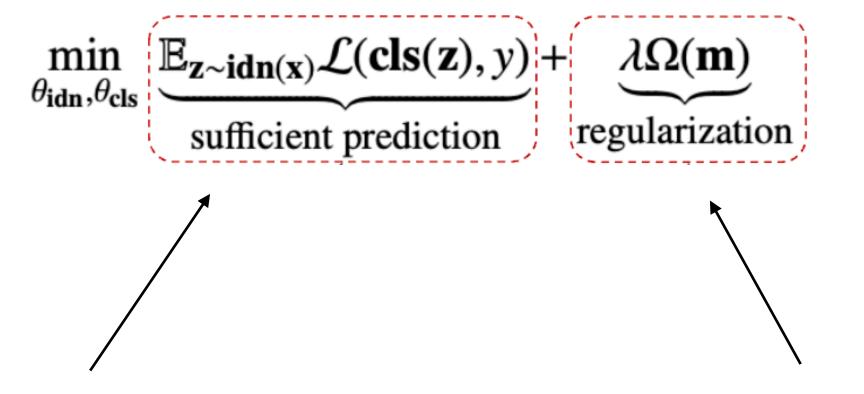
A novel self-explaining model

human evaluation on intrinsic interpretability usefulness

LimitedInk model generates rationals with different length



Optimization Objective



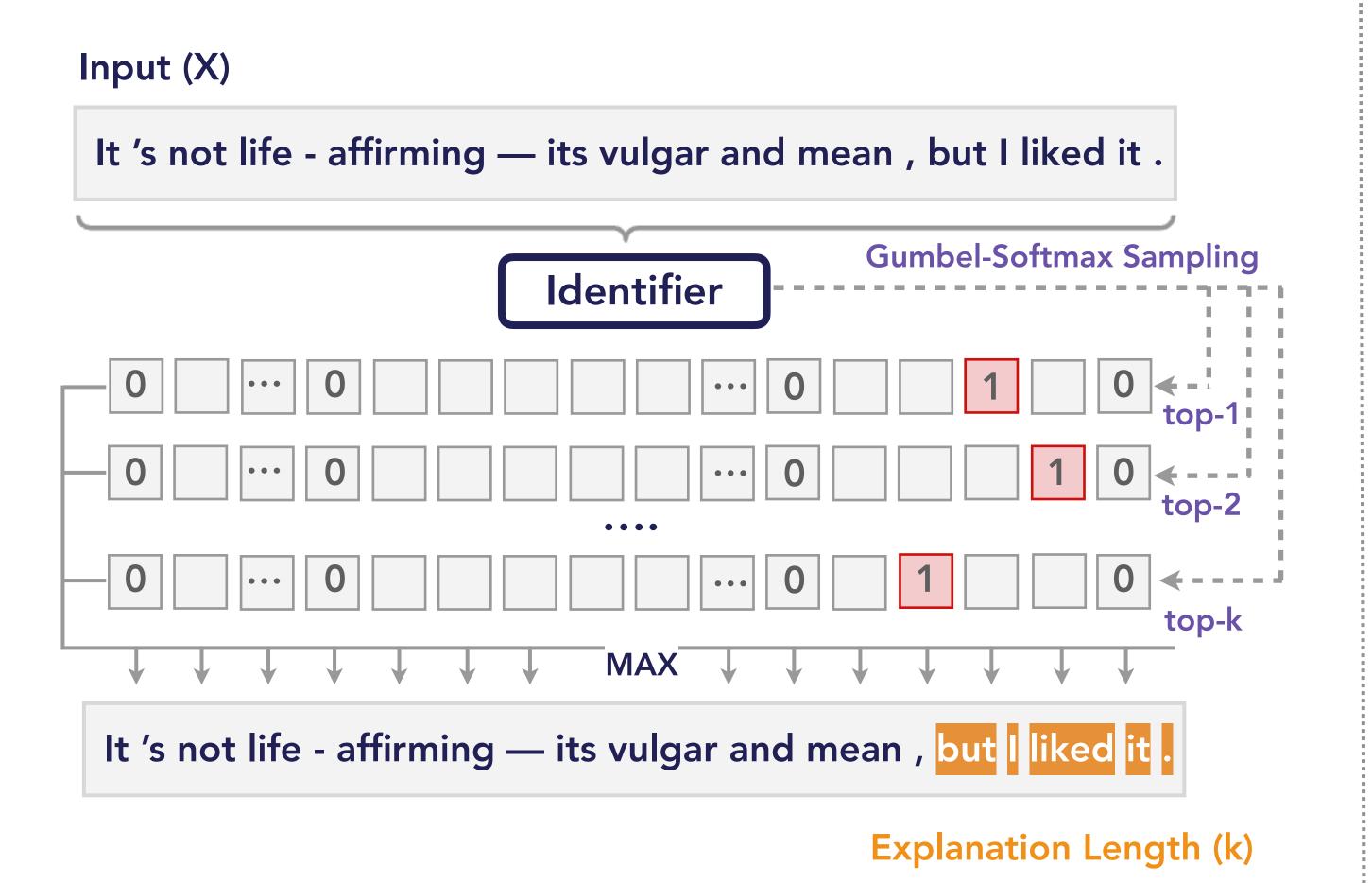
1. Gumbel-Softmax Sampling

2. Vector and Sort Regularization

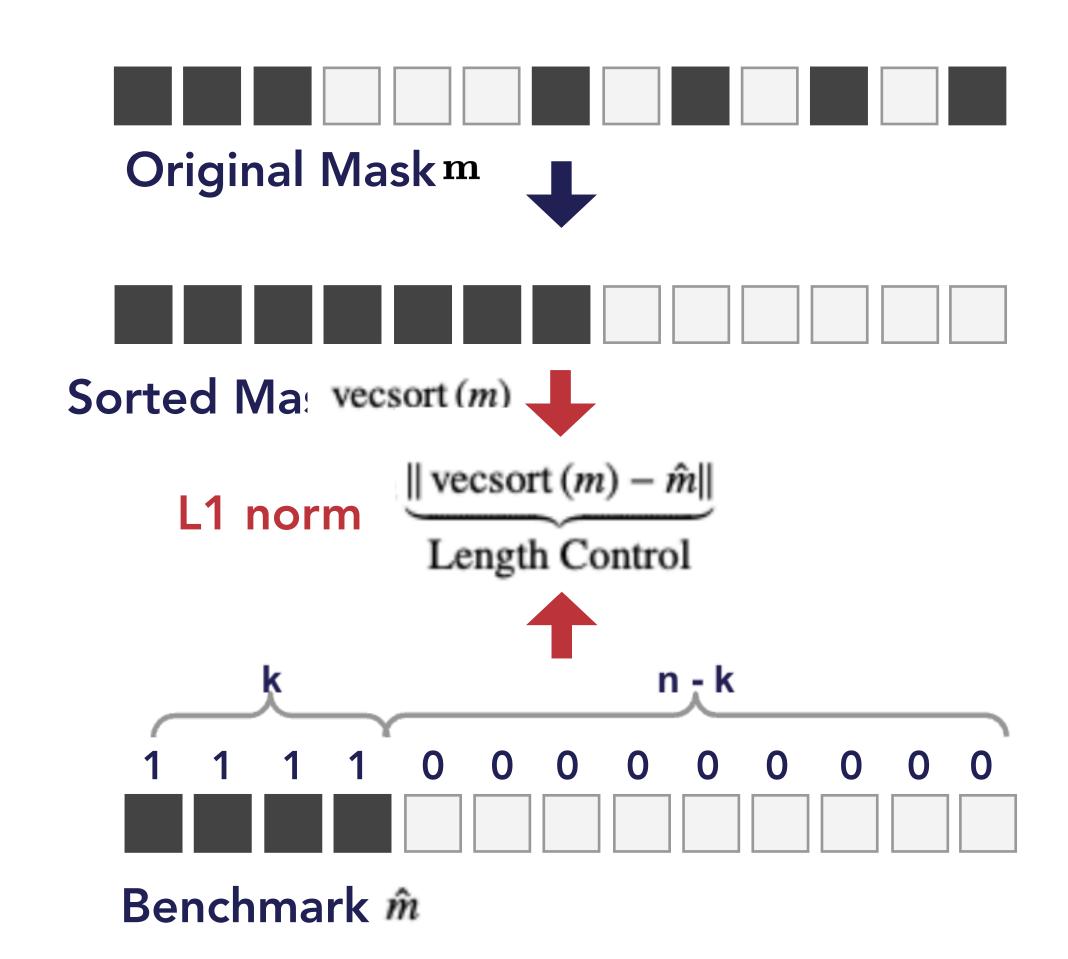


How to control explanation length in LimitedInk

1. Gumbel-Softmax Sampling



2. Vector and Sort Regularization



Can LimitedInk perform well on classification?

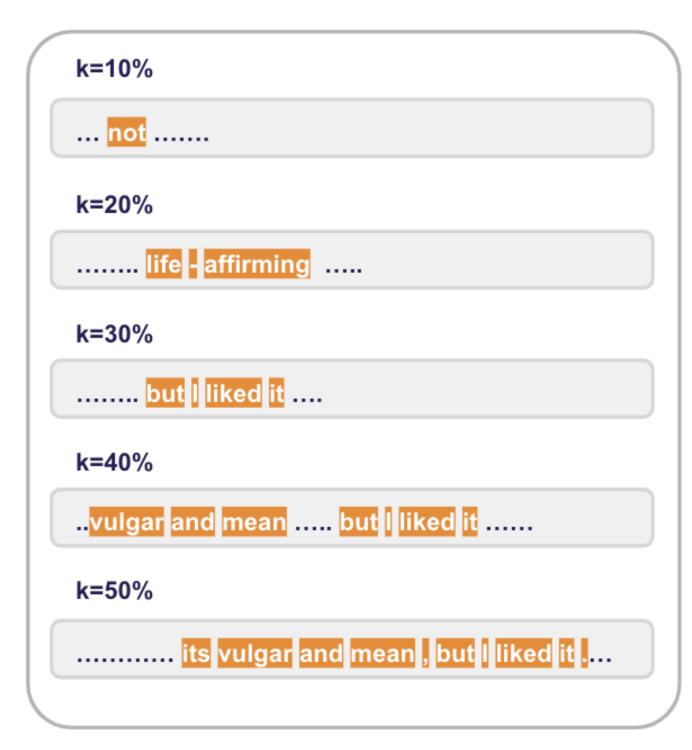
- End-task classification: Task, weighted average F1
- Human Plausibility with annotated dataset: Precision, Recall, Token-level F1

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	.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	Length Level 50%				30%			50%			50%				40%					

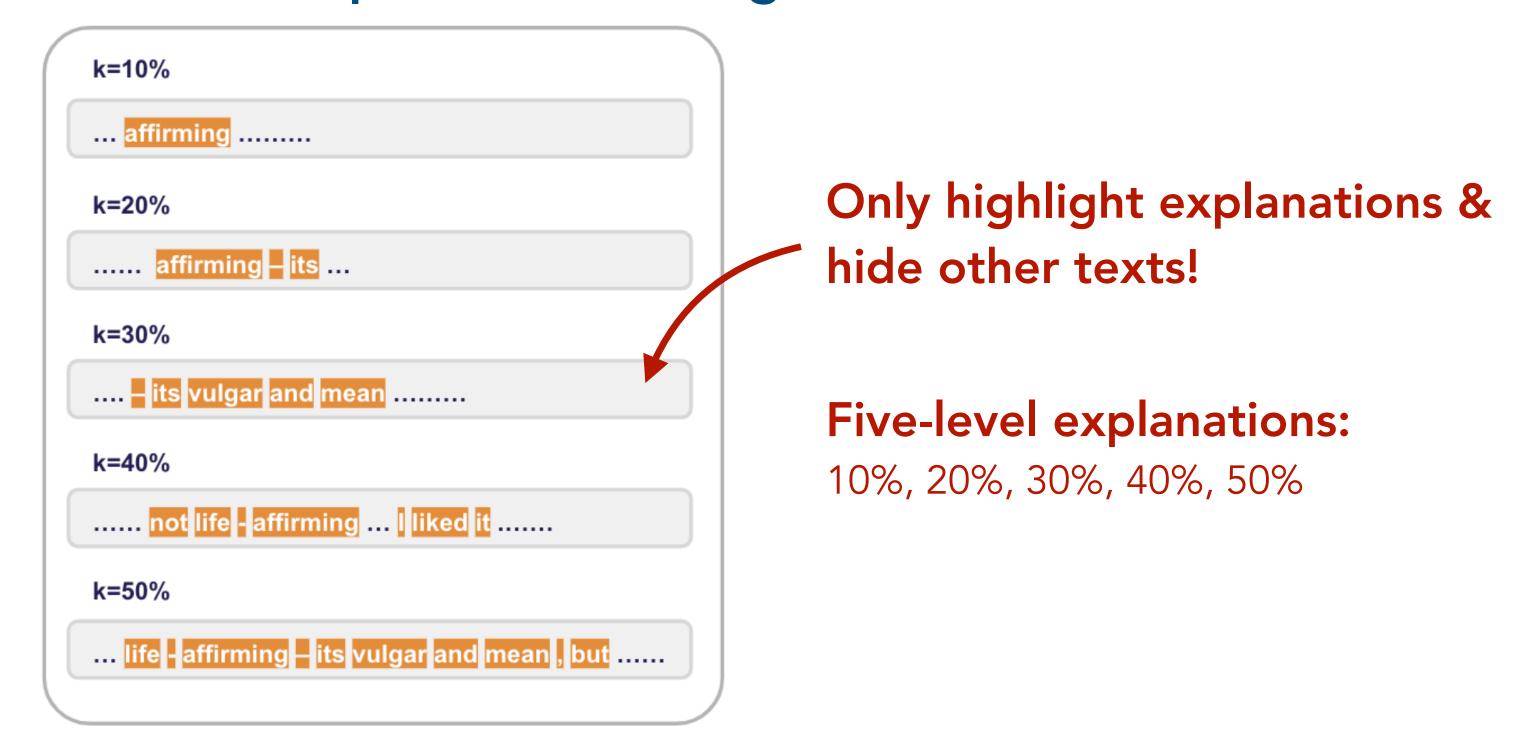
LimitedInk performed compatible with three SOTA baselines on the two common rationale metrics in five ERASER text classification benchmark datasets.

Step2 - Human Study Setups

LimitedInk Explanations

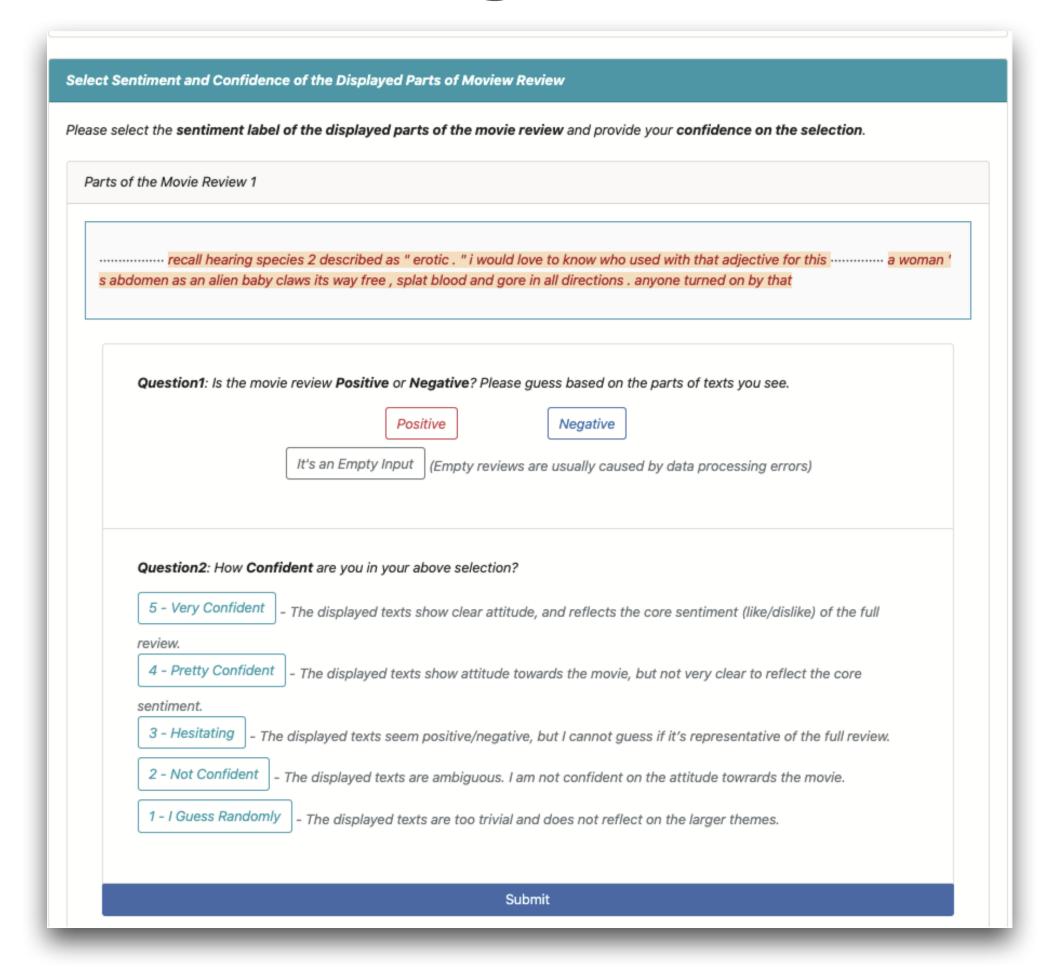


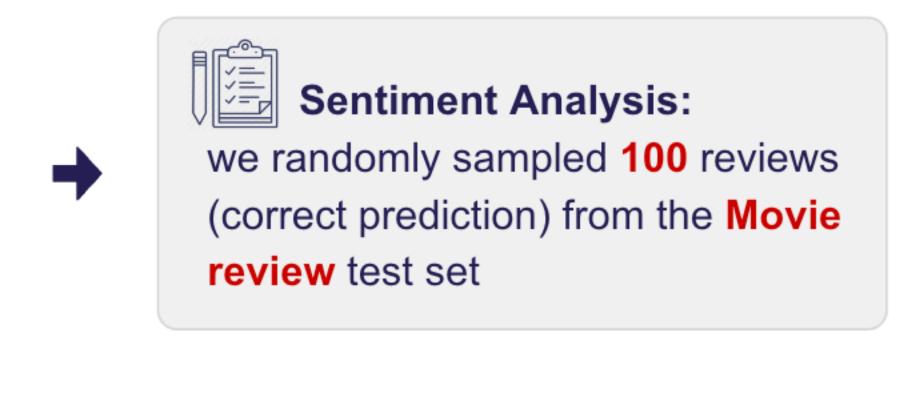
Random text spans (similar length)



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

Human Task Design







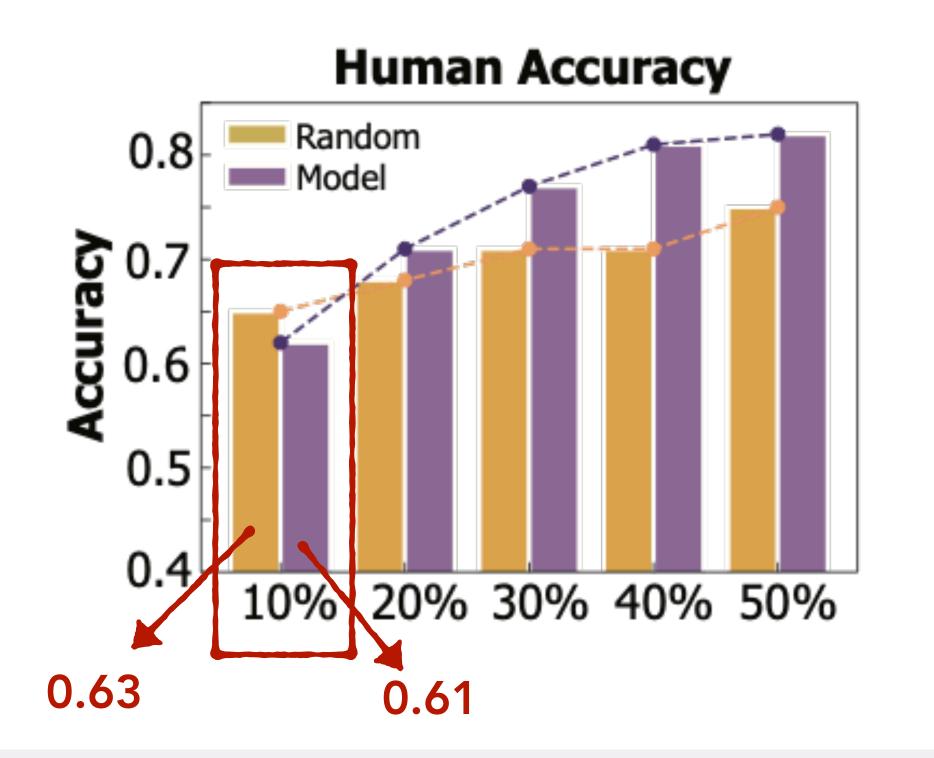


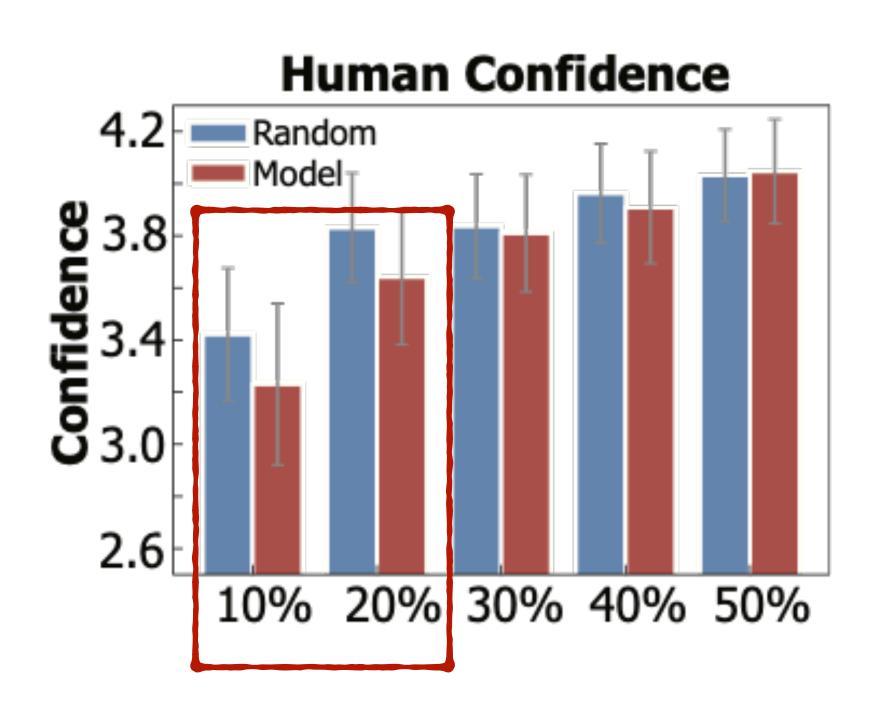
confidence

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.

29

Key Findings





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

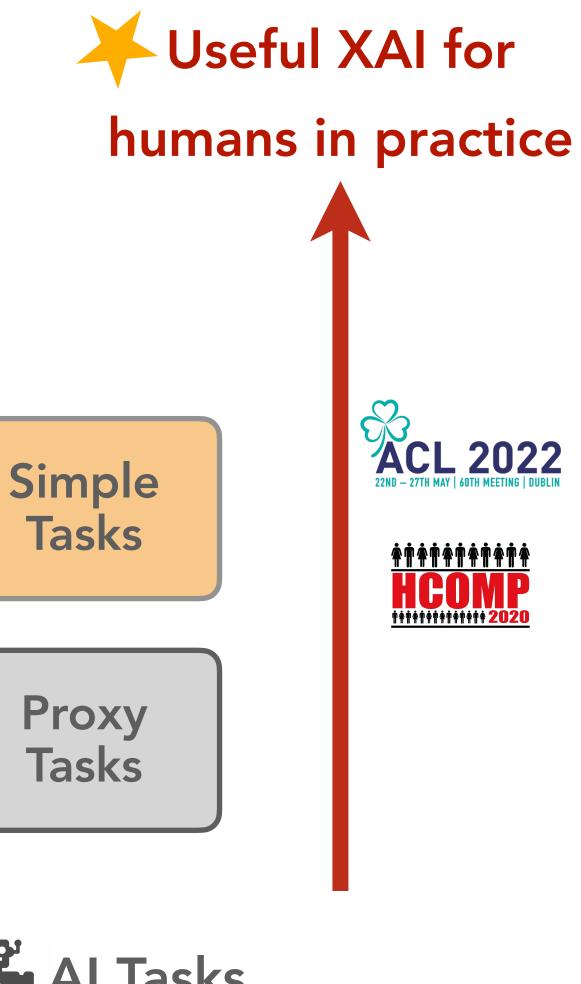
The shortest rationales are NOT always the best for human understanding.

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



Real Humans

No Real

Humans

Human-grounded Evaluation

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

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

Functionally-grounded Evaluation

Humans

Al Interpretability **Evaluation**



Background & Motivation

RQ1: Are XAI Useful RQ2: Why?

for Humans?

Useful XAI







Explaining the Road Not Taken



Hua Shen

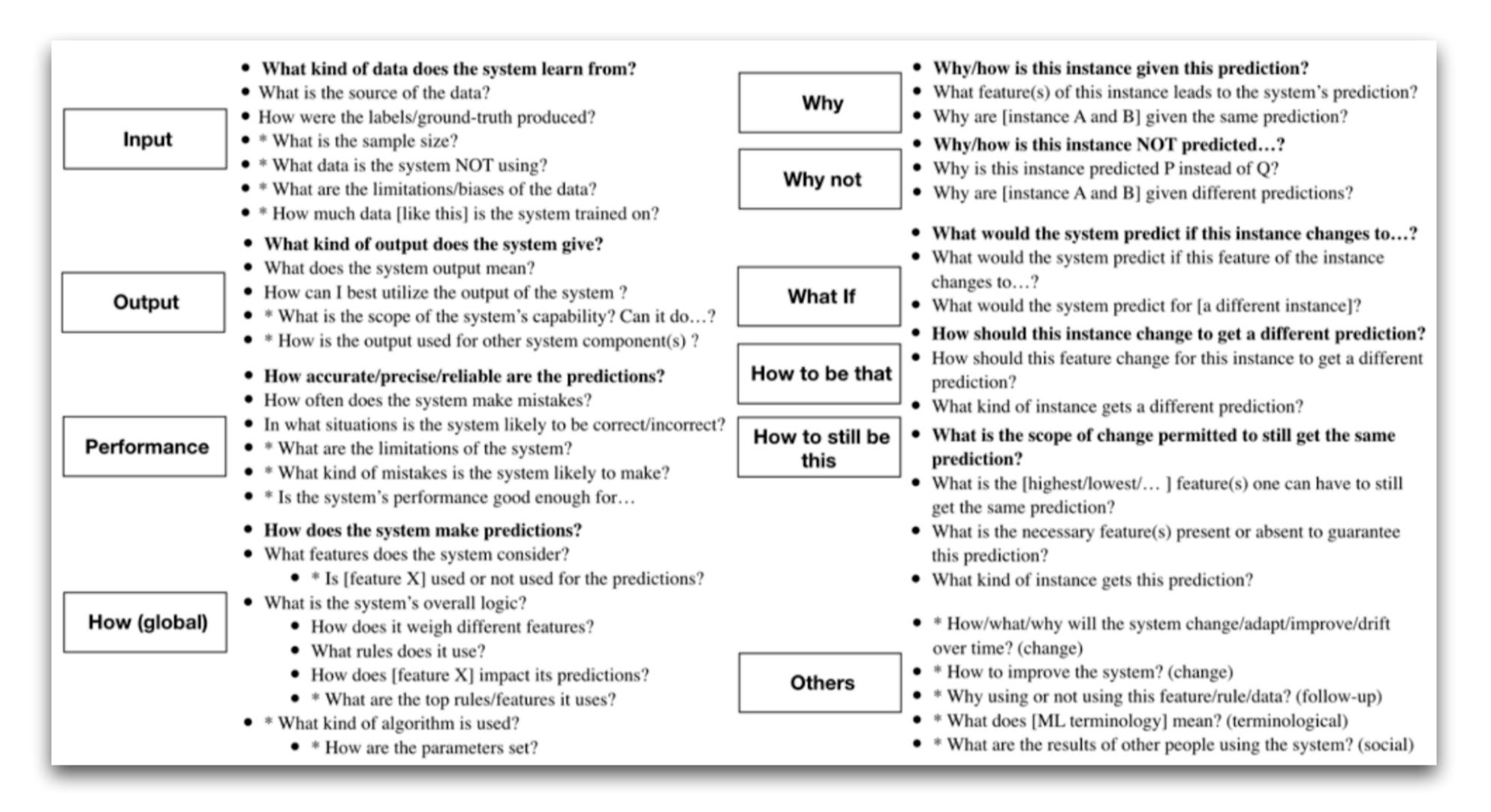


Ting-Hao 'Kenneth' Huang

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

INV	Vane	Venue	Paper URL	Title	Vess	Venue	Paper URL		ID Title Yes	r Venue	Paper URL
Why should I trust you?" Explaining the predictions of an			https://enriv.org/pdf/1602.04938.pdf	74 FIND: 1 wan in the Loop Debugging Deep Text Classifiers		EMNLP	https://www.aclweb.org/s	(anthology/2020 ample o	147 Multimodal language analysis in the wild: Cmu-mosei dataset and interpretat 201		https://www.aclweb.org/anthology/P18-1
causal framework for explaining the predictions of black		EMNLP	https://erxiv.org/pdf/1707.01943.pdf	75 Fine-grained aris, of sentence embeddings using auxiliary pre		ICLR	https://anxiv.org/pdf/1608		148 Multimodal Routing: Improving Local and Global Interpretability of Multimoda 202		https://erxiv.org/pdf/2004.14198.pdf
Nagnostic Study of Explainability Techniques for Text C		EMNLP	https://arxiv.org/pdf/2009.13295.pdf	76 Generating Fact Checking to Livertions		ACL	https://arxiv.org/pdf/2004		149 Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels 202		https://andv.org/pdf/2010.07526.pdf
A figuring-based English Math Word Problem Solver with		COLING	https://www.aclweb.org/anthology/C16-2032.pd	77 Generating question relevant captions : 14 visual question answ		ACL	https://arxiv.org/pdf/1906		150 Neural vector conceptualization for word vector space interpretation 201		https://www.aclweb.org/anthology/W19-;
A priver in bertology; What we know about how bert work		TACL	https://arxiv.org/pdf/2002.12327.pdf	78 Generating Token-Level Explanations for Natural - squage Infere		NAACL.	https://arxiv.org/pdf/1904		151 No Explainability without Accountability: An Empirical Study of Explanations (202)		https://homes.cs.washington.edu/~weld/
A Shar of Attention Mechanism for Interpretation of Neura	al Automatic Post-Ei 2018	ACL	https://www.aclweb.org/arthology/W18-2702.pd	79 GEval: Tool for Debugging NLP Datasets and Models	2019	BlackboxNLP	https://www.aclweb.org/s		152 Obtaining Faithful Interpretations from Compositional Neural Networks 202		https://www.aclweb.org/anthology/2020.
A structural probe for finding syntax in word representation	ns 2019	NAACL	https://www.aclweb.org/anthology/N19-1419.pd	80 Slobal model interpretation via recursive partitioning	2018	DSS	https://arxiv.org/pdf/1802		153 Open Sesame: Getting Inside BERT's Linguistic Knowledge 201:	BlackboxNLP	https://www.aclweb.org/anthologyW19-
A Survey of the State of Explainable All for Natural Langua	age Processing 2020	AACL-IJONLP	https://enxiv.org/pdf/2010.00711.pdf	81 GLUC SE: General.ized and COntextualized Story Explanations	2020	EMNLP	https://arxiv.org/pdf/2005	9.07758.pdf	154 OpenDialKG: Explainable Conversational Reasoning with Attention-based W 201	ACL.	https://www.aclweb.org/anthology/P19-1
Allennip into pret: A framework for explaining predictions	of nip models 2019	EMNLP	https://www.aclweb.org/anthology/D19-3002.pd	82 Guiding the Finning of Semantics: Interpretable Video Captioning	via POS Ti 2019	EM: C	https://pdfs.semanticsch	holar.org/7ad5/4b109a05	155 Pathologies of Neural Models Make Interpretations Difficult 201	EMNLP	https://enxiv.org/pdf/1804.07781.pdf
An Information Bottleneck Approach for Controlling Conci	iseness in Rationals 2020	EMNLP	https://enriv.org/pdf/2005.00652.pdf	83 HEIDL: Learning Ling httic Expressions with Deep Learning and H	Human-in-ti 2019	ACL	https://www.aclweb.org/	anthology/P19-3023.pdf	156 Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BI 202	ACL	https://erxiv.org/pdf/2004.14786.pdf
n Interpretabli Knowledge Transfer Model for Knowledg		ACL	https://www.aclweb.org/anthology/P17-1088.pdf	84 HotpotQA: A Dataset for Div. se, Explainable Multi-hop Question		EMNLP	https://www.aclweb.org/s	(anthology/D18-1259.pd)	157 Predicting and interpreting embeddings for out of vocabulary words in downs 201	BlackboxNLP	https://www.aclweb.org/anthologyW18-
Interpretable Reasoning Network for Multi-Relation Qu		COLING	https://www.aclweb.org/anthology/C18-1171.pdf	85 How contextual are contextualized and representations? Compa		EMNLP-IJCNLP	https://arxiv.org.s. **190)	9.00512.pdf	158 Principles of Explanatory Debugging to Personalize Interactive Machine Lear 201		https://openaccess.city.ac.uk/id/eprint/1
Analysing the potential of seq-to-seq models for incremen	ntal interpretation in 2018	BlackboxNLP	https://www.aclweb.org/anthology/W18-5419.pd	86 How do Decisions Emerge across Layers in Vaural Models? Inter		EMNLP	https://arxiv.org/pdf/2004	4. hand pdf	159 Probing Emergent Semantics in Predictive Agents via Question Answering 202		https://enxiv.org/pdf/2006.01016.pdf
vial sis methods in neural language processing: A surve		TACL	https://ensiv.org/pdf/1812.08951.pdf	87 How Important is a Neuron	2019	ICLR	https://arxiv.org/pdf/1805		160 Probing for semantic evidence of composition by means of simple classificati 201		https://www.aclweb.org/anthology/W16
Inalytical methods for interpretable ultradense word emb		EMNLP	https://www.aclweb.org/anthology/D19-1111.pdf	88 How much should you ask? On the question structure in Grayste	ms 2018	BlackboxNLP	https://enxlv.org/pdf/1805		161 Probing Neural Dialog Models for Conversational Understanding 202		
Inalyzing the Structure of Attention in a Transformer Lang		BlackboxNLP	https://www.aclweb.org/anthology/W19-4808.pd	89 How Useful Are the Machine-Generated Interpretations to General		HCOMP	https://arxiv.org/abs/200		162 - gram Induction by Rationale Generation: Learning to Solve and Explain A 201		https://www.aclweb.org/anthology/P17-
nchors High-Precision Vodel-Agnostic Explanations	2018	AAAI	https://homes.cs.washington.edu/~marcotor/aas	90 Human Attention in Visual Question Answering: Do Humans and I		EMNLP	https://www.actweb.org/		163 PROVER of Generation for Interpretable Reasoning over Rules 202		https://www.aclweb.org/anthology/2020
re side in heads really bitter than one?		NeuIPS	https://enxiv.org/pdf/1905.10650.pdf	91 Human Attention Maps for Text Classification: Do Humans and No.		ACL	https://www.aclweb.org/		164 Quick and (not so) by I Insupervised Selection of Justification Sentences f 202		https://www.aclweb.org/anthology/D19
ssessing rocial and interestional biases in contextualization		NeuIPS	https://enxiv.org/pdf/1911.01485.pdf	92 Human-grounded Evaluations of Explanation Methods for Text Cla		STANLP-LICNLP	https://www.actweb.org/		165 Quint: Interpretable question and using over knowledge bases. 201		https://www.eclweb.org/enthology/D17-
Utention interpretability acrost rilp tasks	2019	Andv	https://arxiv.org/pd/1909.11218.pdf	93 Identification, interpretability, and Bayesian word embeddings	2019	NAACE	https://www.aclweb.org/		166 Rationalizing Neural Predictions 201		https://people.csail.mit.edu/taolei/paper
Ittertion is not Explanation		NAACL.	https://arxiv.org/pdf/1902.10186.pdf	94 Identifying and Controlling Important Neurons in Neural Machine 1		ICLR	https://arxiv.org/pdf/1811		167 Rethinking Cooperative Rationalization: Introspective Araction and Comple 201		https://exxiv.org/pdf/1910.13294.pdf
Utention is not not Explanation	2019 matic MeSH Indexer 2018	EMNLP BioASQ	https://www.aclweb.org/anthology/D19-1002.pd	95 Imparting Interpretability to Word Embeddings while Preserving S		TASLP	https://arxiv.org/pdf/1807		168 Saliency-driven word alignment interpretation for neural macro acceleration 201 169 Self-Assembling Modular Networks for Interpretable Networks for Interpretable Modular Networks for Inte	ACL EMNLP	https://www.aclweb.org/anthology/W19
	matic MeSH Indexer 2018 explanatory models, 2019	EMNLP-UCNLP	https://www.acleeb.org/arthology/W18-5306/	96 Improving Abstractive Document Summarization with Salient Infor 97 Interpretable emoji prediction via label-wise attention LSTMs		ACL EMNLP		anthology/P19-1205.pdf	169 Self-Assembling Modular Networks for Interpretable Multi-Hop Reasoning 201 170 Self-Critical Reasoning for Robust Visual Question Answering 201	New-1010	https://www.aclweb.org/anthology/D19
			https://www.actweb.org/anthology/D19-1415.pdl			EMNLP	https://www.active.org/s	lanthology/D18-1508.pdl		Appli	https://enviv.org/pdf/1905.09998.pdf
utomatic rule extraction from long shift term memory ne ERT Rediscovers ne Classical NLP Pueline	2019	ICLR ACL	https://enxiv.org/pdf/1702.02540.pdf	98 Interpretable Entity Representations through Large-Scale Typing 99 Interpretable Multi-dataset Evaluation for Named Entity Recognition		EMNLP	https://ansiv.org/pdf/2005	1.0000	171 Self-Explaining Structures Improve NLP Models 202 172 Seq2seq-vis: A visual debugging tool for sequence-to-sequence models 201:		https://arxiv.org/pdf/2012.01786.pdf
		101.0	https://www.actweb.org/anthology/P19-1452.pdf	400 Interpretable Mount Sentiment for State of the Sentiment of Sentiment Se		District P	https://enxiv.org/pdf/2011	Indicate and the second	172 degated vis. A visual debugging loa for sequence-to-sequence models 201	1700	Fins://enxiv.org/pdf/1804.09299.pdf v.org/pdf/1502.03044.pdf
nature: A unified an insperie model inter											corg/pdf/1711.08792.pdf
hains-of-Reasoning a TextGraphs 2019	Title				Year	Venue	F	Paper URL			org/pdf/1801.09041.pdf
NM: An Interpretable Complex-valued Ne											eacheb.org/anthology/202
DGS: A Compositional eneralization Ct	H 147	1116	OIL E		0040	LADE					org/pdf/2008.05122.pdf
id-Start and Interpretabley: Turning Reg	" Why shot	uld I trus	st you?" Explaining the	predictions of any classifier	2016	KDD	l h	nttps://arxiv.o	org/pdf/1602.04938.pdf		.org/pdf/1711.07414v1.pdf
mparing Automatic and It man Evaluat	,		,	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,			_				org/pdf/2007.10018.pdf
ompositional Explanations Neurons	3 C P-1		d	data ta NILD	0040	NIA A OI	T.		L L		corg/pdf/1809.07424.pdf
Constructing Interpretive Spatt Temporal 2	Visualizing	and Un	derstanding Neural Mo	dels in NLP	2016	NAACL	l n	nttps://www.a	aclweb.org/anthology/N16-1082.pdf		activeb.org/anthology/P19
Deconfounded lexicon induction for interpri-	0		9					-			acleeb.org/anthology/2020
Deconfounded Lexicon Induction or Intere	Detionalisi	a a Niacca	al Dradiations		2046	ENANH D	L.		a annil mit adviltaninilanananilananint		
Designing and Interpreting Probed with Co	Rationalizi	ng iveur	al Predictions		2016	EMNLP	<u>n</u>	nups://peopi	<u>e.csail.mit.edu/taolei/papers/emnlp16</u>	<u>_rationale</u>	2.DQI (org/pdf/2004.14243.pdf
Detecting and Explaining Causes From Te-											
Detecting Linguistic Characteristics Alzh	DEDT Dod										.aclweb.org/anthology/N19-
	DER LEM	ISCOVERS	the Classical NI P Pin	eline	2019	ACI	h	ttne://www	aclweb org/anthology/P19-1452 ndf		edweb.org/anthology/N19- eadweb.org/anthology/P18-
old the Model Understand the Questa ?	DEKT Ked	iscovers	the Classical NLP Pip	eline	2019	ACL	<u>h</u>	nttps://www.	aclweb.org/anthology/P19-1452.pdf		cackeeb.org/anthology/N19- cackeeb.org/anthology/P18- corg/adf/2002.05867.pdf
Dissonance Between Human and Machie				eline							cacheeb.org/anthology/N19- cacheeb.org/anthology/P18- corg/adf/2002.05867.pdf corg/adf/1809.02701.pdf
Dissonance Between Human and Machine Do Human Rationales Improve Machine				eline							cacheeb.org/anthology/N19 cacheeb.org/anthology/P18- corg/adf/2002.05867.pdf corg/adf/1809.02701.pdf corg/adf/1808.08079.pdf
Dissonance Between Human and Machine Do Human Rationales Improve Machine Do Multi-hop Readers Dream of Reasonin	Attention is			eline	2019 2019	ACL NAACL			aclweb.org/anthology/P19-1452.pdf org/pdf/1902.10186.pdf		cackweb.org/arthology/N19- cackweb.org/anthology/P18- corg/adf/1809.02701.pdf corg/adf/1809.02701.pdf corg/adf/1808.08079.pdf corg/adf/1703.04730.pdf
Dissonance Between Human and Machine Do Human Rationales Improve Machine Do Multi-hop Readers Dream of Reasonin Do Neural Dialog Systems Use the Conver	Attention is	not Exp	planation		2019	NAACL	<u>h</u>	nttps://arxiv.o	org/pdf/1902.10186.pdf		cackweb.org/artihology/N19- cackweb.org/artihology/P18- corg/adf/1809.02701.pdf corg/adf/1808.08079.pdf corg/adf/1703.04730.pdf corg/adf/1809.08037.pdf
issonance Between Human and Machine to Human Rationales Improve Machine to Multi-hop Readers Dream of Reasona to Neural Dialog Systems Use the Conver- toes it care what you asked? Understanding Importance	Attention is	not Exp	planation https://www.oru/sed/1809.03740.sef	119 KERMIT: Complementing Transformer Architectures with Encoder	2019 s of Explici 2020	NAACL	https://www.sclweb.org/s	nttps://arxiv.o	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202) EMNLP	Lackweb.org/artihology/N19- Lackweb.org/artihology/P18- Lorg/adf/2002.05867.adf Lorg/adf/1808.08079.adf Lorg/adf/1703.04730.adf Lorg/adf/1809.08037.adf Lorg/adf/2010.07882.adf
issonance Between Human and Machine to Human Rationales Improve Machine to Human Rationales Improve Machine to Multi-hop Readers Dream of Reasonin to Neural Dialog Systems Use the Conver- tions it care what you asked? Understanding Importance to String-Based Neural MT Learn Source Syntax?	Attention is	BlackbookEP	planation https://arxiv.org/pdf/1809.03740.pdf https://www.isi.edu/naturai-language/m/lemnip1	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea	2019 s of Exp8ci 2020 soning ove 2019	NAACL EMALP	https://www.aclweb.org/i	nttps://arxiv.o	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 193 Understanding neural networks through representation erasure 201	Antiv	Lackeeb, org/anthology/N19 Lackeeb, org/anthology/N19 Lorg/adt/2002, 05867, pdf Lorg/adt/1809, 02701, pdf Lorg/adt/1809, 08079, pdf Lorg/adt/1703, 04730, pdf Lorg/adt/1809, 08037, pdf https://arxiv.org/adt/2010, 07882, pdf https://arxiv.org/adt/1612, 08220, pdf
sonance Between Human and Machine of Human Rationales Improve Machine of Human Rationales Improve Machine of Multi-hop Readers Dream of Reasona of Neural Dialog Systems Use the Conver- oes it care what you asked? Understanding Importance oes String-Based Neural MT Learn Source Syntax? oes String-Based Neural MT Learn Source Syntax?	Attention is	BlackbookEP EMNLP EMNLP	planation https://arxiv.org/pdf/1809.03740.pdf https://www.isi.edu/natural-language/m//emnip1 https://www.aciweb.org/arthology/D16-1159.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention	2019 s of Explici 2020 soning ove 2019 2018	NAACL EMNLP EMNLP NeuIPS	https://www.ackweb.org/i https://www.ackweb.org/i https://arxiv.org/pdf/1807	nttps://arxiv.o	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 193 Understanding neural networks through representation erasure 194 Universal adversarial triggers for attacking and analyzing NLP 201	Aniv EMNLP-IJCNLP	Lackeeb, org/anthology/N19 Lackeeb, org/anthology/P18 Lorg/adf/2002, 05867, adf Lorg/adf/1809, 02701, adf Lorg/adf/1809, 08079, adf Lorg/adf/1809, 08037, adf https://arxiv.org/adf/2010, 07882, adf https://arxiv.org/adf/1612, 08220, adf https://arxiv.org/adf/1612, 08220, adf https://arxiv.org/adf/1612, 08220, adf
sonance Between Human and Machine of Human Rationales Improve Machine of Human Rationales Improve Machine of Multi-hop Readers Dream of Reasonal of Neural Dialog Systems Use the Converses it care what you asked? Understanding Importance ses String-Based Neural MT Learn Source Syntax? Des String-Based Neural MT Learn Source Syntax?	Attention is of Werbs in Deep Li 2018 2016 2016 aplainable Claim Ve 2020	BlackbookEP EMNLP EMNLP ACL	planation https://arxiv.org/pdf/1909.03740.pdf https://www.isi.edu/natural-language/m/l/emnlp1 https://www.acleeb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/2004.13455.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivis	2019 s of Explici 2020 soning ove 2019 2018 al Explanat 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings	https://www.ackweb.org/ https://www.ackweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/2016	nttps://arxiv.o lanthology/2020.amvlaus lanthology/2019-1187.adl 7.03756.pdf 0.04119.pdf	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202	Aniv EMNLP-IJCNLP ACL	Lackeeb, org/anthology/N19 Lackeeb, org/anthology/N19 Lorg/adf/2002, 05867, adf Lorg/adf/1809, 02701, adf Lorg/adf/1809, 08079, adf Lorg/adf/1809, 08037, adf https://arxiv.org/adf/2010, 07882, adf https://arxiv.org/adf/1612, 08220, adf https://arxiv.org/adf/1612, 08220, adf https://arxiv.org/adf/2010, 07882, adf https://arxiv.org/adf/2010, 07882, adf
sonance Between Human and Machine of Human Rationales Improve Machine of Human Rationales Improve Machine of Multi-hop Readers Dream of Reasonal of Neural Dialog Systems Use the Converses it care what you asked? Understanding Importance os String-Based Neural MT Learn Source Syntax? TCA: Decision Tree-based Co-Attention Networks for Exercise Natural language inference with natural language exercises.	Attention is of Werbs in Deep L/ 2018 2016 2016 explainable Claim Ve 2020 explanations 2018	BlackboxNLP EMNLP EMNLP ACL NeuIPS	planation https://arxiv.org/pdf/1909.03740.pdf https://www.isi.edu/natural-language/mi/emnlp1 https://www.aciweb.org/pdf/2004.13455.pdf https://arxiv.org/pdf/2004.13455.pdf https://arxiv.org/pdf/1812.01193.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent elignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivis 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea	2019 s of Explici 2020 soning ove 2019 2018 al Explanat 2020 uning from 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL	https://www.ackweb.org/ https://www.ackweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/2016 https://www.ackweb.org/	https://arxiv.o lanthology/2020.amr/la-r lanthology/2020.amr/la-r (anthology/2020.ad-dem	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201	Arxiv EMNLP-IJCNLP ACL ACL	Lackweb, org/anthology/N19 Lackweb, org/anthology/P18 Lorg/pdf/2002.05867.pdf Lorg/pdf/1809.02701.pdf Lorg/pdf/1809.08079.pdf Lorg/pdf/1703.04730.pdf Lorg/pdf/1809.08037.pdf https://arxiv.org/pdf/2010.07882.pdf https://arxiv.org/pdf/1612.08220.pdf https://arxiv.org/pdf/1612.08220.pdf https://arxiv.org/pdf/1612.08220.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/pdf/2005.01218.pdf
sonance Between Human and Macure b Human Rationales Improve Machine b Human Rationales Improve Machine b Multi-hop Readers Dream of Reasonin b Neural Dialog Systems Use the Conver- bes it care what you asked? Understanding Importance bes String-Based Neural MT Learn Source Syntax? bes String-Based Neural MT Learn Source Syntax? ICA: Decision Tree-based Co-Attention Networks for Earni: Natural language inference with natural language e stNTS: An Neural Programmer-Interpreter Model for Se	Attention is of Werbs in Deep L/ 2018 2016 2016 explainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019	BlackboxNLP EMNLP EMNLP ACL NeuiPS ACL	planation https://www.isi.edu/nstural-language/millemnip1 https://www.isi.edu/nstural-language/millemnip1 https://www.aclweb.org/arthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aclweb.org/arthology/P19-1331.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficient statements.	2019 s of Explici 2020 soning ove 2019 2018 at Explanat 2020 rming from 2020 int bag-of-< 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS	https://www.aclweb.org/ https://www.aclweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1907 https://www.aclweb.org/ https://arxiv.org/pdf/1707	https://arxiv.orgines.com/per/ lanthology/2020.amn/per/ lanthology/2020.amn/per/ 7.03756.pdf 0.04118.pdf lanthology/2020.acl-dem 2.03342.pdf	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Ox 201	Arxiv EMNLP-IJCNLP ACL BlackboxNLP	cackesb.org/anthology/N1/ cackesb.org/anthology/P18/ corg/pdf/2002.05887.pdf corg/pdf/1809.02731.pdf corg/pdf/1809.08079.pdf corg/pdf/1809.08037.pdf corg/pdf/1809.08037.pdf https://arxiv.org/pdf/2010.07882.pdf https://arxiv.org/pdf/1812.08220.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/pdf/2005.01218.pdf https://www.ackeeb.org/anthology/P18/ https://www.ackeeb.org/anthology/P18/ https://www.ackeeb.org/anthology/W1
sonance Between Human and Machine of Human Rationales Improve Machine of Multi-hop Readers Dream of Reasonal of Neural Dialog Systems Use the Converses it care what you asked? Understanding Importance oes String-Based Neural MT Learn Source Syntax? TCA: Decision Tree-based Co-Attention Networks for Exertil: Natural language inference with natural language estiNTS: An Neural Programmer-Interpreter Model for Seduce: Explaining model decisions through unsupervised	Attention is of Werbs in Deep L/ 2018 2016 2016 explainable Claim Ve 2020 explanations 2018 entence Simplificatic 2019 concepts extraction 2019	BlackboxNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv	bitps://enviv.org/pdf/1909.03740.pdf https://www.isi.edu/natural-languaga/m/iemnlp1 https://www.aclweb.org/arithology/D16-1159.pdf https://enviv.org/pdf/1812.01193.pdf https://enviv.org/pdf/1812.01193.pdf https://enviv.org/pdf/1812.01193.pdf https://enviv.org/pdf/1905.11652.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivis 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficiency 125 Learning Corresponded Rationales for Text Matching	2019 s of Explici 2020 soning ove 2019 2018 al Explanat 2020 rming from 2020 ant bag-of< 2019 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR	https://www.aclweb.org/ https://www.aclweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1703 https://arxiv.org/pdf/1703 https://apenreview.net/p	https://arxiv.orgines.com/per/ lanthology/2020.amn/per/ lanthology/2020.amn/per/ 2.03756.pdf 0.04119.pdf lanthology/2020.acl-dem/ 2.03342.pdf pdf/hdr-yk/Qas096n	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Ou 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Ox 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201	Arxiv EMNLP-IJCNLP ACL ACL BlackboxNLP EMNLP	Lackeeb, org/anthology/N19 Lackeeb, org/anthology/P18 Lorg/pdf/2002.05867.pdf Lorg/pdf/1809.02701.pdf Lorg/pdf/1809.08037.pdf Lorg/pdf/1809.08037.pdf Lorg/pdf/1809.08037.pdf https://arxiv.org/pdf/1612.08220.pdf https://arxiv.org/pdf/1612.08220.pdf https://arxiv.org/pdf/2005.01218.pdf https://www.ackeeb.org/anthology/D19 https://www.ackeeb.org/anthology/P18 https://www.ackeeb.org/anthology/P18 https://www.ackeeb.org/anthology/P18
sonance Between Human and Macure between Human Rationales Improve Machine between Human Rationales Importance between Human Bource Syntax? between String-Based Neural MT Learn Source Syntax? ICA: Decision Tree-based Co-Attention Networks for Earn!: Natural language inference with natural language e stNTS: An Neural Programmer-Interpreter Model for Se twee: Explaining model decisions through unsupervised sembling Visual Explanations for VQA	Attention is of Werbs in Deep Li 2018 2016 2016 xplainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019 concepts extraction 2019 2017	BlackboxNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W	bitps://enviv.org/pdf/1909.03740.pdf https://www.isi.edu/natural-language/mi/lemnlp1 https://www.aclweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/1802.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://enviv.org/pdf/1812.01193.pdf https://enviv.org/pdf/1905.11852.pdf https://enviv.org/pdf/1905.11852.pdf https://www.cs.utexas.edu/-mi/papers/rajani.vig	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivis 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization	2019 s of Explici 2020 soning ove 2019 2018 al Explanat 2020 sming from 2020 smt bag-of-< 2019 2019 on 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICOM	https://www.aclweb.org/ https://www.aclweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/12016 https://arxiv.org/pdf/1703 https://arxiv.org/pdf/1703 https://arxiv.org/pdf/1908	https://arxiv.orgines.com/per/ lanthology/D19-1187.pdl 7.03756.pdf 0.04119.pdf lanthology/2020.acl-dem 2.03342.pdf pdf/hdr-nk/Qas096m 8.05601.pdf	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Ou 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Oc 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categoric 200	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP T EMNLP NAACL	c. ackeeb. org/anthology/N1/ c. ackeeb. org/anthology/P18/ corg/pdf/1809.02701.pdf corg/pdf/1809.08079.pdf corg/pdf/1809.08037.pdf corg/pdf/1809.08037.pdf https://arxiv.org/pdf/1812.08220.pdf https://arxiv.org/pdf/1812.08220.pdf https://arxiv.org/pdf/1812.08220.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/pdf/2005.01218.pdf https://arxiv.org/anthology/P18 https://arxiv.org/anthology/P18 https://arxiv.org/anthology/P18 https://arxiv.org/anthology/P18 https://arxiv.org/anthology/P18
sonance Between Human and Machine Human Rationales Improve Machine Human Rationales Improve Machine Multi-hop Readers Dream of Reasonin Neural Dialog Systems Use the Conver- es it care what you asked? Understanding Importance es String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? ICA: Decision Tree-based Co-Attention Networks for Es init: Natural language inference with natural language e itNTS: An Neural Programmer-Interpreter Model for Se tuce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA aser: A benchmark to evaluate rationalized nip models	Attention is of Werbs in Deep Li 2018 2016 2016 xplainable Claim Ve 2020 explanations 2018 entence Simplificatic 2019 concepts extraction 2019 2017 2020	BlackboxNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W ACL	bitps://enviv.org/pdf/1909.03740.pdf https://www.isi.edu/natural-language/millemnip1 https://www.aclweb.org/artihology/D16-1159.pdf https://enviv.org/pdf/1812.01193.pdf https://enviv.org/pdf/1812.01193.pdf https://enviv.org/pdf/1912.01193.pdf https://enviv.org/pdf/1905.11852.pdf https://enviv.org/pdf/1905.11852.pdf https://www.cs.utexas.edu/-mil/papers/rejani.vig https://enviv.org/pdf/1911.03429.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivis 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Matchine 128 Learning Dynamics of Attention: Human Prior for Interpretable Matchine 129 Learning Dynamics of Attention: Human Prior for Interpretable Matchine 129 Learning Dynamics of Attention: Human Prior for Interpretable Matchine 129 Learning Dynamics of Attention:	2019 s of Explici 2020 soning ove 2019 2018 al Explanat 2020 sming from 2020 smt bag-of-c 2019 an 2019 chine Reas 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICOM NeuIPS	https://www.aclweb.org/ https://www.aclweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1703 https://arxiv.org/pdf/1905 https://arxiv.org/pdf/1905 https://arxiv.org/pdf/1905	https://arxiv.org/lands/per/lands/pe	org/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Ou 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Oc 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categoric 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP F EMNLP NAACL J LJCAI-W	c. ackesb. org/anthologa/N1/ c. ackesb. org/anthologa/P1/ corg/pdf/1809.02791.pdf corg/pdf/1809.02791.pdf corg/pdf/1809.08037.pdf corg/pdf/1809.08037.pdf corg/pdf/1809.08037.pdf https://arxiv.org/pdf/2010.07882.pdf https://arxiv.org/pdf/1612.0820.pdf https://arxiv.org/pdf/1612.0820.pdf https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.ackesb.org/anthologa/D1/ https://www.cs.utexas.edu/users/mi/p/
sonance Between Human and Machine b Human Rationales Improve Machine b Human Rationales Improve Machine b Multi-hop Readers Dream of Reasonin b Neural Dialog Systems Use the Conver- bes it care what you asked? Understanding Importance bes String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Ex- snit: Natural language inference with natural language e- stNTS: An Neural Programmer-Interpreter Model for Se tuce: Explaining model decisions through unsupervised seembling Visual Explanations for VQA aser: A benchmark to evaluate rationalized nip models reluating and Characterizing Human Rationales	Attention is of Werbs in Deep Li 2018 2016 2016 xplainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019 2017 2020 2020	BlackboxNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W ACL EMNLP	bitps://enviv.org/pdf/1909.03740.pdf https://www.isi.edu/natural-languaga/m/iemnlp1 https://www.aciweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/2004.13455.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://enviv.org/pdf/1905.11852.pdf https://enviv.org/pdf/1911.03429.pdf https://enviv.org/pdf/1911.03429.pdf https://enviv.org/pdf/2010.04736.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivi 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning oredible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive	2019 s of Explici 2020 soning ove 2019 2018 al Explanat 2020 sming from 2020 smt bag-of-c 2019 2019 on 2019 chine Reas 2019 Logic Prog 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP	https://www.aclweb.org/ https://www.aclweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1807 https://www.aclweb.org/ https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://www.aclweb.org/	https://arxiv.org/person/perso	brg/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categoric 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP F EMNLP NAACL F LICAL-W T ICASSP	cackeb.org/anthologe/N1 cackeb.org/anthologe/P1/corg/pdf/2002.05867.pdf corg/pdf/1809.02791.pdf corg/pdf/1809.082791.pdf corg/pdf/1809.08037.pdf corg/pdf/1809.08037.pdf https://arxiv.org/pdf/2010.07882.pdf https://arxiv.org/pdf/2010.07882.pdf https://arxiv.org/pdf/2005.01218.pdf https://www.ackeeb.org/anthologe/D1/https://www.ackeeb.org/anthologe/D1/https://www.ackeeb.org/anthologe/D1/https://www.ackeeb.org/anthologe/D1/https://www.ackeeb.org/anthologe/D1/https://www.ackeeb.org/anthologe/N0/https://www.cs.utexas.edu/users/mi/phttps://www.cs.utexas.edu/users/mi/phttps://web.eecs.utexas.edu/users/
sonance Between Human and Machine Human Rationales Improve Machine Human Rationales Improve Machine Multi-hop Readers Dream of Reasonin Neural Dialog Systems Use the Conver- es it care what you asked? Understanding Importance es String-Based Neural MT Leam Source Syntax? es String-Based Neural MT Leam Source Syntax? CA: Decision Tree-based Co-Attention Networks for Exitic Natural language inference with natural language et INTS: An Neural Programmer-Interpreter Model for Se tuce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA aser: A benchmark to evaluate rationalized nip models aluating and Characterizing Human Rationales aluating Explainable Al; Which Algorithmic Explanation	Attention is of Werbs in Deep Li 2018 2016 2016 xplainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019 2017 2020 2020 as Help Users Predi 2020	BlackboxNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W ACL EMNLP ACL	bitps://www.isi.edu/natural-languaga/millemnip1 https://www.isi.edu/natural-languaga/millemnip1 https://www.aciweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://arxiv.org/pdf/1905.11852.pdf https://www.cs.utexas.edu/-millpapers/rajani.vig https://arxiv.org/pdf/1911.03429.pdf https://arxiv.org/pdf/2010.04736.pdf https://arxiv.org/pdf/2010.04736.pdf https://www.aciweb.org/anthology/2020.adi-mais	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivi 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explainations from Language Data	2019 s of Explici 2020 soning ove 2019 2018 al Explanat 2020 sming from 2020 smt bag-of-c 2019 an 2019 chine Rese 2019 Logic Prog 2020 2018	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICOM NeuIPS EMNLP BlackboxNLP	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/	https://arxiv.org/person/perso	brg/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguat 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categoric 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualisation and diagnostic classifiers' reveal how recurrent and recursive no 201	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP F EMNLP NAACL F LICAI-W T ICASSP JAIR	cackesb.org/anthology/N1) cackesb.org/anthology/P18 corg/pdf/1809.02791.pdf corg/pdf/1809.02791.pdf corg/pdf/1809.08079.pdf corg/pdf/1809.08037.pdf https://arxiv.org/pdf/2010.07882.pdf https://arxiv.org/pdf/2010.07882.pdf https://arxiv.org/pdf/2010.07882.pdf https://www.ackeeb.org/anthology/D13
nonance Between Human and Machine Human Rationales Improve Machine Multi-hop Readers Dream of Reasonin Neural Dialog Systems Use the Converse at care what you asked? Understanding Importance as String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Exhit: Natural language inference with natural language ethTS: An Neural Programmer-Interpreter Model for Seuce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA user: A benchmark to evaluate rationalized nip models aluating and Characterizing Human Rationales aluating Explainable Al: Which Algorithmic Explanation aluating neural network explanation methods using hyting supervised using su	Attention is of Werbs in Deep Li 2018 2016 2016 xplainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019 2017 2020 2020 as Help Users Predi 2020 brid documents and 2018	BlackboxNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W ACL EMNLP ACL ACL	bitps://www.ini.edu/natural-language/milemnip1 https://www.ini.edu/natural-language/milemnip1 https://www.aciweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1332.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivi 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree	2019 s of Explicit 2020 soning ove 2019 2018 all Explanet 2020 sming from 2020 smt bag-of-c 2019 2019 on 2019 chine Reas 2019 Logic Prog 2020 2018 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR	https://www.aclweb.org/ https://www.aclweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1807 https://www.aclweb.org/ https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908	https://arxiv.org/lanthology/2020.emolp-r/lanthology/2020.emolp-r/lanthology/2020.emolp-r/lanthology/2020.acl-dem/2.03342.pdf idnihology/2020.acl-dem/2.03342.pdf idnihology/2020.acl-dem/8.05601.pdf 5.11666.pdf ianthology/2020.emolp-r/lanthology/2	the properties of the content of the	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL IJCAI-W ICASSP JAIR ACL	cackeb, orgianthology/N1 cackeb, orgianthology/N1 corgiadt/2002.05867.pdf corgiadt/1809.02791.pdf corgiadt/1809.02791.pdf corgiadt/1809.08079.pdf corgiadt/1809.08037.pdf https://arxiv.orgiadt/2010.07882.pdf https://arxiv.orgiadt/2010.07882.pdf https://arxiv.orgiadt/2010.07882.pdf https://arxiv.orgiadt/2005.01218.pdf https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/N0 https://www.cs.utexas.edu/users/mi/p https://www.cs.utexas.edu/users/mi/p https://www.ackeeb.org/anthology/N0 https://www.ackeeb.org/anthology/N0 https://www.ackeeb.org/anthology/N0 https://www.ackeeb.org/anthology/N1
sonance Between Human and Machine Human Rationales Improve Machine Deaders Dream of Reasonin Neural Dialog Systems Use the Converse at care what you asked? Understanding Importance as String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Enril: Natural language inference with natural language ethTS: An Neural Programmer-Interpreter Model for Seuce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA user: A benchmark to evaluate rationalized nip models aluating and Characterizing Human Rationales aluating Explainable Al: Which Algorithmic Explanation aluating neural network explanation methods using hybert: A visual analysis tool to explore learned represent	Attention is of Werbs in Deep Li 2018 2016 2016 2016 xplainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019 2017 2020 2020 as Help Users Predi 2020 and documents and 2018 tations in transforms 2020	BlackboxNLP EMNLP EMNLP ACL Areiv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL	bitps://www.ini.edu/natural-language/milemnip1 https://www.ini.edu/natural-language/milemnip1 https://www.aciweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://arxiv.org/pdf/1911.03429.pdf https://arxiv.org/pdf/1911.03429.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivit 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning interpretable negation rules via weak supervision at door	2019 s of Explicit 2020 soning ove 2019 2018 all Explanat 2020 sming from 2020 smit bag-of-c 2019 2019 on 2019 chine Reas 2019 b Logic Prog 2020 2018 2020 ament level 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL	https://www.aclweb.org/ https://www.aclweb.org/ https://arxiv.org/pdf/1807 https://arxiv.org/pdf/1807 https://www.aclweb.org/ https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908 https://arxiv.org/pdf/1908	https://arxiv.org/lanthology/2020.emn/p-r/lanthology/2020.emn/p-r/lanthology/2020.emn/p-r/lanthology/2020.acl-dem/2.03342.pdf/lanthology/2020.acl-dem/2.03342.pdf/lanthology/2020.emn/p-r/lanthology/2	the properties of the content of the	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL IJCAI-W ICASSP JAIR ACL NAACL	c acleeb.org/anthologe/N1/ c acleeb.org/anthologe/P1/ corg/pdf/1809.02791.pdf corg/pdf/1809.02791.pdf corg/pdf/1809.08037.pdf corg/pdf/1809.08037.pdf torg/pdf/1809.08037.pdf https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/D1/ https://www.acleeb.org/anthologe/P1/ https://www.acleeb.org/anthologe/P1/ https://www.acleeb.org/anthologe/P1/ https://www.acleeb.org/anthologe/P1/ https://www.acleeb.org/anthologe/P1/ https://www.acleeb.org/anthologe/P1/ https://www.acleeb.org/anthologe/P1/ https://www.acleeb.org/anthologe/P1/
sonance Between Human and Machine Human Rationales Improve Machine Death Particular Disting Systems Use the Converse at care what you asked? Understanding Importance as String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Exit. Natural language inference with natural language ethTS: An Neural Programmer-Interpreter Model for Seuce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA user: A benchmark to evaluate rationalized nip models eluating and Characterizing Human Rationales eluating Explainable Al: Which Algorithmic Explanation eluating neural network explanation methods using hybert: A visual analysis tool to explore learned represent dERT: Representation Engineering with Natural Language.	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 concepts extraction 2019 2017 2020 2020 es Help Users Predi 2020 ord documents and 2018 tations in transform 2020 uage Explanations 2020	BlackbowNLP EMNLP EMNLP ACL Arriv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL	bitps://www.ini.edu/natural-language/milemnip1 https://www.ini.edu/natural-language/milemnip1 https://www.aciweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1805.11852.pdf https://arxiv.org/pdf/1905.11852.pdf https://arxiv.org/pdf/1911.03429.pdf https://arxiv.org/pdf/2010.04736.pdf https://arxiv.org/pdf/2010.04736.pdf https://arxiv.org/pdf/1911.03429.pdf https://arxiv.org/pdf/1910.05276.pdf https://arxiv.org/pdf/1910.05276.pdf https://arxiv.org/pdf/1910.05276.pdf https://arxiv.org/pdf/1910.05276.pdf https://arxiv.org/pdf/1910.05276.pdf https://arxiv.org/pdf/1910.05276.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivit 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning Interpretable Relationships between Entities, Relations of	2019 s of Explicit 2020 soning ove 2019 2018 all Explanat 2020 sming from 2020 smit bag-of-c 2019 2019 on 2019 chine Reas 2019 b Logic Prog 2020 2018 2020 sment level 2019 and Concep 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL ACL	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.research-co https://www.research-co https://www.research-co	https://arxiv.org/parkips/ lanthology/2020.emn/p-r lanthology/2020.em/p-r (anthology/2020.acl-dem 2.03342.pdf (anthology/2020.acl-dem 8.05601.pdf 5.11666.pdf (anthology/2020.emn/p-r 8.04127.pdf (anthology/2020.emn/p-r 8.04127.pdf (anthology/2020.emn/p-r 8.04127.pdf (anthology/2020.emn/p-r 8.04127.pdf (anthology/2020.emn/p-r 8.04127.pdf	the properties of the second s	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL IJCAI-W ICASSP JAIR ACL NAACL EMNLP-IJCNLP	cackeb, orgianthology/N1 cackeb, orgianthology/P1 corgipdf/2002.05867.pdf corgipdf/1809.02791.pdf corgipdf/1809.02791.pdf corgipdf/1809.08079.pdf corgipdf/1809.08037.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2005.01218.pdf https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/N0 https://www.ackeeb.org/anthology/N0 https://www.ackeeb.org/anthology/N1 https://www.ackeeb.org/anthology/P1 https://www.ackeeb.org/anthology/P1 https://www.ackeeb.org/anthology/P1 https://www.ackeeb.org/anthology/P1 https://www.ackeeb.org/anthology/P1 https://www.ackeeb.org/anthology/P1 https://www.ackeeb.org/anthology/N1 https://arxiv.org/pdf/1908.05620.pdf
sonance Between Human and Machine Human Rationales Improve Machine Multi-hop Readers Dream of Reasonin Neural Dialog Systems Use the Converse is it care what you asked? Understanding Importance is String-Based Neural MT Learn Source Syntax? String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Exit: Natural language inference with natural language entry. A Neural Programmer-Interpreter Model for Secoil Explaining model decisions through unsupervised learning Visual Explanations for VQA ser. A benchmark to evaluate rationalized nip models illusting and Characterizing Human Rationales illusting Explainable Al: Which Algorithmic Explanation illusting neural network explanation methods using hyter: A visual analysis tool to explore learned represent BERT: Representation Engineering with Natural Language Information Foursett Leveraging Language Models for Committee	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 retence Simplificatic 2019 2017 2020 2020 es Help Users Predi 2020 erid documents and 2018 tations in transform 2020 euage Explanations 2020 econsense Reasonin 2019	BlackbowNLP EMNLP EMNLP ACL Arriv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL	bitps://www.ini.edu/natural-language/milemnip1 https://www.ini.edu/natural-language/milemnip1 https://www.aciweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1332.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P2020.aci-main https://www.aciweb.org/arithology/2020.aci-main https://www.aciweb.org/arithology/2020.aci-main https://www.aciweb.org/arithology/2020.aci-main https://www.aciweb.org/arithology/2020.aci-main	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivit 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tiree 131 Learning Interpretable Relationships between Entities, Relations a 132 Learning to Explain Entity Relationships in Knowledge Graphs	2019 se of Explicit 2020 soning over 2019 2018 all Explanat 2020 sming from 2020 smit bag-of-c 2019 2019 on 2019 schine Reac 2019 Logic Prog 2020 2018 2020 sment level 2019 and Concep 2020 2015	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboaNLP ICLR NAACL ACL ACL	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.research-co https://www.research-co https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/	https://arxiv.org/person/perso	prg/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categori. 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualization and diagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 205 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding the Effectiveness of BERT 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL IJCAI-W ICASSP JAIR ACL NAACL EMNLP-IJCNLP EMNLP EMNLP ICASSP	cackeb, orgianthology/N1 cackeb, orgianthology/P1 corgipdf/2002.05887.pdf corgipdf/1809.02791.pdf corgipdf/1809.02791.pdf corgipdf/1809.08079.pdf corgipdf/1809.08037.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2005.01218.pdf https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/N0 https://www.ackeb.org/anthology/P1
sonance Between Human and Machine in Human Rationales Improve Machine in Multi-hop Readers Dream of Reasonin Neural Dialog Systems Use the Converse is it care what you asked? Understanding Importance is String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Explicit Natural language inference with natural language ethTS: An Neural Programmer-Interpretor Model for Seluce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA isser. A benchmark to evaluate rationalized nip models aluating and Characterizing Human Rationales aluating Explainable Al: Which Algorithmic Explanation aluating neural network explanation methods using hybert: A visual analysis tool to explore learned represent plainable Active Learning (KAL): An Empirical Study of	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 2016 2016 2018 2018 2018 2017 2020 2020 2020 2020 2020 2020 2020	BISICKBOWNLP EMINLP EMINLP ACL Arriv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL ACL ACL	bitps://www.isi.edu/natural-language/mt/emnlp1 https://www.isi.edu/natural-language/mt/emnlp1 https://www.isi.edu/natural-language/mt/emnlp1 https://www.aclweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explainations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning Interpretable Relationships between Entities, Relations of 132 Learning to Explain Entity Relationships in Knowledge Graphs 133 Learning to Explain: Answering Why-Questions via Rephrasing	2019 se of Explicit 2020 soning over 2019 2018 all Explanatt 2020 sming from 2020 smit bag-of-c 2019 2019 on 2019 schine React 2019 chine React 2020 2018 2020 sment level 2019 and Concep 2020 2015 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboaNLP ICLR NAACL ACL ACL ACL	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/	https://arxiv.org/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.ed-dem/203342.pdf//articles/2020.em/particles/203342.pdf//articles/2020.em/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.em/particles/2020.ed-main/particles/2020.ed-ma	192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categoric 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualisation and diagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 204 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding the Effectiveness of BERT 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vigue: Explaining, elaborating, and enhancing your answers for visual questi 201	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL IJCAI-W ICASSP JAIR ACL NAACL EMNLP-IJCNLP EMNLP EMNLP EMNLP ICASSP	cackeb, orgianthology/N1 cackeb, orgianthology/P1 corgipdf/2002.05867.pdf corgipdf/1809.02791.pdf corgipdf/1809.02791.pdf corgipdf/1809.08079.pdf corgipdf/1809.08037.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/1812.08220.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2005.01218.pdf https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/D1 https://www.ackeb.org/anthology/P1
to a Model Understand the Cuestory isonance Between Human and Machine I Human Rationales Improve Machine I Multi-hop Readers Dream of Reasonin I Neural Dialog Systems Use the Conver- es it care what you asked? Understanding Importance es String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? ICA: Decision Tree-based Co-Attention Networks for Estric: Natural language inference with natural language estings: An Neural Programmer-Interpreter Model for Seluce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA seer: A benchmark to evaluate rationalized nip models aluating and Characterizing Human Rationales aluating Explainable Al: Which Algorithmic Explanation aluating repairal network explanation methods using hyb- bert: A visual analysis tool to explore learned represent pBERT: Representation Engineering with Natural Lang- plain Yourself Leveraging Language Models for Comm plainable Active Learning (XAL): An Empirical Study of plainable Automated Fact-Checking for Public Health C	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 retence Simplificatic 2019 2017 2020 2020 es Help Users Predi 2020 erid documents and 2018 tations in transform 2020 euage Explanations 2020 consense Reasonin 2019 (How Local Explana 2020 Claims 2020	BlackbowNLP EMNLP EMNLP ACL Arriv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL	bitps://www.isi.edu/natural-language/mt/emnlp1 https://www.isi.edu/natural-language/mt/emnlp1 https://www.isi.edu/natural-language/mt/emnlp1 https://www.aclweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concesponded Rationales for Text Matching 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning interpretable Relationships between Entities, Relations of 132 Learning to Explain: Answering Why-Questions via Rephrasing 133 Learning to Explain: Datasets and Models for Identifying Valid Res	2019 se of Explicit 2020 soning over 2019 2018 all Explanatt 2020 sming from 2020 smit bag-of-c 2019 2019 on 2019 schine React 2019 chine React 2020 2018 2020 sment level 2019 and Concep 2020 2015 2019 assoning Ch 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboaNLP ICLR NAACL ACL ACL ACL EMNLP	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.research-co https://www.research-co https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/	https://arxiv.orgical lanthology/2020.emnlp-r (anthology/2020.emnlp-r (anthology/2020.acl-dem 2.03342.pdf (anthology/2020.acl-dem 8.05601.pdf 5.11666.pdf (anthology/2020.emnlp-r 8.04127.pdf (anthology/2020.emnlp-r (anthology/2020.emnlp-r (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main (anthology/2020.acl-main)	192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categori. 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualisation and diagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 204 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding the Effectiveness of BERT 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vigue: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT is not: Lessons from a new suite of psycholinguistic diagnostica 202	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL ICAI-W ICASSP JAIR ACL NAACL EMNLP-IJCNLP EMNLP-IJCNLP EMNLP EMNLP TACL	cackeeb, orgianthology/N1 cackeeb, orgianthology/P1 corgipdf/1909, 05687, pdf corgipdf/1909, 05791, pdf corgipdf/1909, 05079, pdf corgipdf/1909, 05079, pdf corgipdf/1909, 05077, pdf torgipdf/1909, 05037, pdf https://arxiv.orgipdf/2010, 07892, pdf https://arxiv.orgipdf/2010, 07892, pdf https://arxiv.orgipdf/2005, 01218, pdf https://arxiv.orgipdf/2005, 01218, pdf https://www.ackeeb.org/anthology/D1/ https://www.ackeeb.org/anthology/D1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://www.ackeeb.org/anthology/N1/ https://arxiv.org/pdf/1908,05620.pdf https://arxiv.org/pdf/1908,05620.pdf https://orgiv.org/pdf/2010,06775.pdf https://orgiv.org/pdf/1907,13528.pdf
sonance Between Human and Mac not Human Rationales Improve Machine Delation Readers Dream of Reasons Neural Dialog Systems Use the Converse at care what you asked? Understanding Importance as String-Based Neural MT Learn Source Syntax? as String-Based Neural MT Learn Source Syntax? GA: Decision Tree-based Co-Attention Networks for Enri: Natural language inference with natural language of INTS: An Neural Programmer-Interpreter Model for Secoe: Explaining model decisions through unsupervised sembling Visual Explanations for VQA user: A benchmark to evaluate rationalized nip models slusting and Characterizing Human Rationales slusting reural network explanation methods using hybert: A visual analysis tool to explore learned represent (BERT: Representation Engineering with Natural Langulain Yourself! Leveraging Language Models for Commissionable Active Learning (VAL): An Empirical Study of Islainable Automated Fact-Checking for Public Health Colsinable Clinical Decision Support from Text	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 2016 2016 2016 2018 explainable Claim Ve 2020 explanations 2018 reterico Simplificatic 2019 2017 2020 2020 es Help Users Predi 2020 erid documents and 2018 tations in transform 2020 usage Explanations 2020 consense Reasonin 2019 (How Local Explana 2020 2020	BISICKBOWNLP EMINLP EMINLP ACL Aridv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL EMNLP EMNLP EMNLP EMNLP	bitps://www.ini.edu/natural-language/milemnip1 https://www.ini.edu/natural-language/milemnip1 https://www.aclweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rejani.vig https://www.cs.utexas.edu/-mil/papers/rejani.vig https://www.aclweb.org/anthology/P18-1032.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concesponded Rationales for Text Matching 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explainations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning interpretable Relationships between Entities, Relations at 132 Learning to Explain: Answering Why-Questions via Rephrasing 133 Learning to Explain: Datasets and Models for Identifying Valid Rea 136 Learning Variational Word Masks to Improve the Interpretability of	2019 se of Explicit 2020 soning over 2019 2018 all Explanat 2020 sming from 2020 sming from 2020 sming from 2019 chine Rese 2019 chine Rese 2019 chine Rese 2020 2018 2020 sment level 2019 and Conces 2020 2015 2019 assoning Ch 2020 Neural Tex 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL ACL ACL ACL EMNLP EMNLP	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.research-co https://www.research-co https://www.aclweb.org/	https://arxiv.org/parkip.com/park	Inderstanding Neural Abstractive Summarization Models via Uncertainty 193 Understanding neural networks through representation erasure 194 Universal adversarial triggers for attacking and analyzing NLP 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categori. 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualisation and diagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 204 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding Neural Models in NLP 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vigue: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT is not: Lessons from a new suite of psycholinguistic diagnostica 202 209 What can All do for me? evaluating machine learning interpretations in coope 201	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL ICAI-W ICASSP JAIR ACL NAACL EMNLP-IJCNLP EMNLP EMNLP TAGL IUI	cackeeb, orgianthology/N1 cackeeb, orgianthology/P1 corgipdf/1809.02701.pdf corgipdf/1809.02701.pdf corgipdf/1809.02701.pdf corgipdf/1809.08079.pdf corgipdf/1809.08037.pdf https://arxiv.orgipdf/2010.07892.pdf https://arxiv.orgipdf/2010.07892.pdf https://arxiv.orgipdf/2010.07892.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://www.ackeeb.orgianthology/D1 https://www.ackeeb.orgianthology/D1 https://www.ackeeb.orgianthology/D1 https://www.ackeeb.orgianthology/D1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://arxiv.orgipdf/2010.06775.pdf https://arxiv.orgipdf/2010.06775.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf
sonance Between Human and Machine In Human Rationales Improve Machine In Human Rationales Improve Machine In Multi-hop Readers Dream of Reasoni In Neural Dialog Systems Use the Converse I was it care what you asked? Understanding Importance I was String-Based Neural MT Learn Source Syntax? I was	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 retence Simplificatic 2019 2017 2020 2020 es Help Users Predi 2020 erid documents and 2018 tations in transform 2020 euge Explanations 2020 euge Expl	BISICKBOWNLP EMINLP EMINLP ACL Arriv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL EMNLP EMNLP EMNLP EMNLP EMNLP	bitps://www.isi.edu/natural-language/mt/emnlp1 https://www.isi.edu/natural-language/mt/emnlp1 https://www.isi.edu/natural-language/mt/emnlp1 https://www.aclweb.org/anthology/D16-1159.pdf https://arxiv.org/pd/1812.01193.pdf https://arxiv.org/pd/1812.01193.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mt/papers/rajani.vig https://www.cs.utexas.edu/-mt/papers/rajani.vig https://www.cs.utexas.edu/-mt/papers/rajani.vig https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-1032.pdf	119 KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivit 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularizati 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning interpretable Relationships between Entities, Relations of 132 Learning to Explain: Answering Why-Questions via Rephrasing 133 Learning to Explain: Datasets and Models for Identifying Valid Rea 136 Learning Variational Word Masks to Improve the Interpretability of 137 Lightly-supervised representation learning with global interpretability of 137 Lightly-supervised representation learning with global interpretability of	2019 se of Explicit 2020 soning over 2019 2018 all Explanat 2020 sming from 2020 sming from 2020 sming from 2019 chine Reset 2019 chine Reset 2019 chine Reset 2020 2018 2020 sment level 2019 and Concep 2020 2015 2019 assoning Ch 2020 Neural Tex 2020 ity 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboaNLP ICLR NAACL ACL ACL ACL EMNLP EMNLP EMNLP EMNLP	https://www.aclweb.org/ https://www.aclweb.org/	https://arxiv.orgicality.com/per/arthology/019-1187.pdf 7,03756.pdf 0,04119.pdf (anthology/2020.acl-dem 2,03342.pdf (anthology/2020.acl-dem 8,05801.pdf (anthology/2020.amnip-r 8,04127.pdf (anthology/2020.amnip-r 8,04127.pdf (anthology/2020.acl-main (anthology/2020.acl-ma	Inderstanding Neural Abstractive Summarization Models via Uncertainty 193 Understanding neural networks through representation erasure 194 Universal adversarial triggers for attacking and analyzing NLP 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categori. 200 201 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 202 Visualisation and diagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 204 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding the Effectiveness of BERT 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vige-e: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT is not: Lessons from a new suite of psycholinguistic diagnostics 202 209 What can All do for me? evaluating machine learning interpretations in coope 201 210 What do Neural Machine Translation Models Learn about Morphology? 201	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP F EMNLP F NAACL IJCAI-W F ICASSP JAIR ACL B NAACL EMNLP-IJCNLP EMNLP D EMNLP D EMNLP D EMNLP D EMNLP D EMNLP D EACL D IUI F ACL	cackeeb, orgianthology/N1 cackeeb, orgianthology/P1 corgipdf/1909, 05795, pdf corgipdf/1909, 05791, pdf corgipdf/1909, 05791, pdf corgipdf/1909, 05079, pdf corgipdf/1909, 05077, pdf torgipdf/1909, 05037, pdf https://arxiv.orgipdf/2010, 07892, pdf https://arxiv.orgipdf/2010, 07892, pdf https://arxiv.orgipdf/2005, 01218, pdf https://arxiv.orgipdf/2005, 01218, pdf https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/D1 https://www.ackeeb.org/anthology/N0 https://www.ackeeb.org/anthology/N0 https://www.ackeeb.org/anthology/N1 https://www.ackeeb.org/anthology/N1 https://www.ackeeb.org/anthology/N1 https://www.ackeeb.org/anthology/N1 https://www.ackeeb.org/anthology/N1 https://www.ackeeb.org/anthology/N1 https://arxiv.org/pdf/1908,05620,pdf https://arxiv.org/pdf/1908,05620,pdf https://arxiv.org/pdf/1907,13528.pdf https://arxiv.org/pdf/1907,13528.pdf https://arxiv.org/pdf/1907,13528.pdf https://arxiv.org/pdf/1907,13528.pdf
sonance Between Human and Machine In Human Rationales Improve Machine In Multi-hop Readers Dream of Reasonin In Neural Dialog Systems Use the Converse es it care what you asked? Understanding Importance es String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? ICA: Decision Tree-based Co-Attention Networks for Entri: Natural language inference with natural language et INTS: An Neural Programmer-Interpreter Model for Se tuce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA asser: A benchmark to evaluate rationalized nip models aluating and Characterizing Human Rationales aluating Explainable Al: Which Algorithmic Explanation aluating reutral network explanation methods using hyt bert: A visual analysis tool to explore learned represent pBERT: Representation Engineering with Natural Lang plain Yourself! Leveraging Language Models for Comm plainable Active Learning (XAL): An Empirical Study of plainable Automated Fact-Checking for Public Health C plainable Clinical Decision Support from Text plaining Black Box Predictions and Univelling Data Artit	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 retence Simplificatic 2019 2017 2020 2020 es Help Users Predi 2020 erid documents and 2018 tations in transform 2020 euge Explanations 2020	BISICKBOWNLP EMINLP EMINLP ACL Aridv NIPS-W ACL EMNLP ACL	bitps://www.iai.edu/natural-language/mt/emnlp1 https://www.iai.edu/natural-language/mt/emnlp1 https://www.aclweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.aclweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aclweb.org/anthology/P18-1032.pdf https://www.aclweb.org/anthology/P18-2020.emnlp-r https://www.aclweb.org/anthology/P18-2020.emnlp-r https://www.aclweb.org/anthology/P18-2020.emnlp-r https://www.aclweb.org/anthology/P18-2020.acl-main	KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning interpretable Relationships between Entities, Relations of 132 Learning to Explain: Answering Why-Questions via Rephrasing 133 Learning to Explain: Datasets and Models for Identifying Valid Rea 136 Learning Variational Word Masks to Improve the Interpretability of 137 Lightly-supervised representation learning with global interpretability of 138 Linguistic Knowledge and Transferability of Contextual Representation 138 Linguistic Knowledge and Transferability of Contextual Representation	2019 see of Explicit 2020 soning over 2019 2018 all Explanatt 2020 aming from 2020 and bag-of-c 2019 2019 on 2019 on 2019 chine React 2019 chine React 2020 2018 2020 ament level 2019 and Concec 2020 2015 2019 assoning Ch 2020 (Neural Tex 2020 ity 2019 attors 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL ACL ACL ACL EMNLP EMNLP EMNLP NAACL NAACL NAACL NAACL	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.research-co https://www.aclweb.org/	https://arxiv.orgicality.com/per/arthology/019-1187.pdf 7,03756.pdf 0,04119.pdf (anthology/2020.acl-dem 2,03342.pdf (anthology/2020.acl-dem 8,05601.pdf 5,11666.pdf (anthology/2020.acl-main (anthol	Inderstanding Neural Abstractive Summarization Models via Uncertainty 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 205 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu. 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N. 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Oc. 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categori. 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy. 201 201 Using regional saliency for speech emotion recognition. 202 Visualizing and Understanding Neural Machine Translation. 203 Visualizing and Understanding Neural Machine Translation. 204 Visualizing and Understanding Neural Models in NLP. 205 Visualizing and Understanding the Effectiveness of BERT. 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vige-e: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT is not: Lessons from a new suite of psycholinguistic diagnostics. 202 209 What can All do for me? evaluating machine learning interpretations in coope 201 201 What do Neural Machine Translation Models Learn about Morphology? 201 202 What do Neural Machine Translation Models Learn about Morphology? 201 202 What do Neural Machine Translation Models Learn about Morphology? 201 202 What do Neural Machine Translation Models Learn about Morphology? 203 204 What do Neural Machine Translation Models Learn about Morphology? 205 207 208 What do Neural Machine Translation Models Learn about Morphology? 207 208 What do Neural Machine Translation Models Learn about Morphology?	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL ICAI-W ICASSP JAIR ACL NAACL EMNLP-IJCNLP EMNLP EMNLP TACL IUI TACL IUI TACL ICIR	cackeeb, orgianthology/N1 cackeeb, orgianthology/P1 corgipdf/1809.02701.pdf corgipdf/1809.02701.pdf corgipdf/1809.08079.pdf corgipdf/1809.08079.pdf corgipdf/1809.08037.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://www.ackeeb.orgianthology/P1 https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf
sonance Between Human and Machine In Human Rationales Improve Machine In Human Rationales Improve Machine In Multi-hop Readers Dream of Reasoni In Neural Dialog Systems Use the Converse I was it care what you asked? Understanding Importance I was String-Based Neural MT Learn Source Syntax? I was String-Based Neural I language of String-Based Neural I language of String-Based Neural I language of String-Based Neural I support I was seen that I was seen that I was seen to be supported to the String-Based Neural Learning (SAL): An Empirical Study of plainable Automated Fact-Checking for Public Health Coptainable Clinical Decision Support from Text plainable Prediction of Medical Codes from Clinical Texplaining Black Box Predictions and Univelling Data Artitiplaining Character-Aware Neural Networks for Word-Learning Character-Aware Neural Neural Neural Networks for Word-Learning Character-Aware Neural Ne	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 retence Simplificatic 2019 2017 2020 2020 es Help Users Predi 2020 erid documents and 2018 lations in transform 2020 erid documents and 2018 lations in transform 2020 erid documents and 2018 lations in transform 2020 erid documents 2020 latins 2020 dd 2018 lacts through influer 2020 evel Prodiction: Do 2018	BISCHDONNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W ACL EMNLP ACL	bitps://www.iai.edu/naturai-language/milemnio1 https://www.iai.edu/naturai-language/milemnio1 https://www.aciweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1365.pdf	KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivit 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning Interpretable Relationships between Entities, Relations in 132 Learning to Explain: Answering Why-Questions via Rephrasing 133 Learning to Explain: Datasets and Models for Identifying Valid Rei 136 Learning Variational Word Masks to Improve the Interpretability of 137 Lightly-supervised representation learning with global interpretabil 138 Linguistic Knowledge and Transferability of Contextual Represent 139 LISA: Explaining Recurrent Neural Network Judgments via Layer-	2019 se of Explicit 2020 soning over 2019 2018 all Explanatt 2020 aming from 2020 anti bag-of-c 2019 2019 on 2019 chine React 2019 chine React 2020 2018 2020 ament level 2019 and Concec 2020 2015 2019 assoning Ch 2020 (Neural Tex 2020 ity 2019 attors 2019 wise Serns 2018	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboaNLP ICLR NAACL ACL ACL ACL EMNLP EMNLP EMNLP NAACL BlackboaNLP ICLR ICLR	https://www.aclweb.org/ https://www.aclweb.org/	https://arxiv.org/parkings/2020.emnip-r/anthology/D19-1187.pdl 7.03756.pdf 0.04119.pdf 0.04119.pdf (anthology/2020.acl-dem 2.03342.pdf adf/ad-ni/Qaa09tm 8.05601.pdf 5.11666.pdf (anthology/2020.emnip-r/anthology/2020.emni	Inderstanding Neural Abstractive Summarization Models via Uncertainty 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 205 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu. 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N. 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Oc. 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categori. 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy. 201 201 Using regional saliency for speech emotion recognition. 202 Visualizing and Understanding Neural Machine Translation. 203 Visualizing and Understanding Neural Machine Translation. 204 Visualizing and Understanding Neural Models in NLP. 205 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vige-e: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT is not: Lessons from a new suite of psycholinguistic diagnostics. 202 209 What can All do for me? evaluating machine learning Interpretations in coope 201 209 What do Neural Machine Translation Models Learn about Morphology? 201 202 What do Neural Machine Translation Models Learn about Morphology? 201 202 What Does BERT Learn about the Structure of Language? 203 204 What Does BERT Learn about the Structure of Language?	Arxiv EMNLP-IJCNLP ACL BlackboxNLP EMNLP NAACL ICAI-W ICASSP JAIR ACL NAACL EMNLP-IJCNLP EMNLP EMNLP TACL IUI ACL	c acleeb, orgianthology/N1 c acleeb, orgianthology/P1 corgipdf/1909.05697.pdf corgipdf/1909.02701.pdf corgipdf/1909.08079.pdf corgipdf/1909.08079.pdf corgipdf/1909.08077.pdf https://arxiv.orgipdf/2010.07892.pdf https://arxiv.orgipdf/2010.07892.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/1905.01218.pdf https://arxiv.orgipdf/1711.10203.pdf https://arxiv.orgipdf/1711.10203.pdf https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.056316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf
sonance Between Human and Machine between Human and Machine between Human and Machine between Human and Machine between Draining Importance between Draining Systems Use the Conver- bes it care what you asked? Understanding Importance bes String-Based Neural MT Learn Source Syntax? best String-Based Neural MT Learn Source Syntax? best String-Based Neural MT Learn Source Syntax? best String-Based Neural Programmer-Interpreter Model for Se bluce: Explaining model decisions through unsupervised insembling Visual Explanations for VQA best Can Neural Programmer-Interpreter Model for Se bluce: Explaining model decisions for VQA best Can Neural Explanations for VQA best Can Neural Explanations for VQA best Can Neural Explanation methods using hyd best A visual analysis tool to explore learned represent plaining Foreign Study of the Selection Support from Text best Can Neural Mean Support from Text best Can Neural Mean Support from Text best Can Neural Mean Support from Text best String-Based Sex Predictions and Univelling Data Artifi best String-Based Sex Predictions an	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 retence Simplificatic 2019 2017 2020 2020 es Help Users Predi 2020 erid documents and 2018 lations in transform 2020 erid documents and 2018 lations in transform 2020 erid documents and 2018 lations in transform 2020 erid documents 2020 erid documents 2020 erid documents 2020 erid production 2020 erid lation 2020 erid lation 2020 erid lation 2020 erid lation 2020 erid Production: Do 2018 erid Deep Architec 2018	BISICKDOWNLP EMINLP EMINLP ACL Aridy NIPS-W ACL EMINLP ACL	bitps://www.iai.edu/natural-language/milemnio1 https://www.iai.edu/natural-language/milemnio1 https://www.aciweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1332.pdf https://www.aciweb.org/anthology/P18-1335.pdf https://www.aciweb.org/anthology/P18-1365.pdf https://www.aciweb.org/anthology/P18-1365.pdf https://www.aciweb.org/anthology/P18-1365.pdf	120 Knowledge Aware Conversation Generation with Explainable Real 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning prodible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning interpretable Relationships between Entities, Relations in 132 Learning Interpretable Relationships between Entities, Relations in 133 Learning to Explain: Answering Why-Questions via Rephrasing 134 Learning to Explain: Datasets and Models for Identifying Valid Rei 135 Learning Variational Word Masks to Improve the Interpretability of 137 Lightly-supervised representation learning with global interpretabil 138 Linguistic Knowledge and Transferability of Contextual Represent 139 LISA: Explaining Recurrent Neural Network Judgments via Layer- 140 Localizing Moments in Video With Natural Language	2019 se of Explicit 2020 soning over 2019 2018 all Explanatt 2020 aming from 2020 ant bag-of-c 2019 2019 on 2019 chine React 2019 chine React 2020 2018 2020 ament level 2019 and Concec 2020 2015 2019 assoning Ch 2020 (Neural Tex 2020 ity 2019 attors 2019 wise Serns 2018 2017	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboaNLP ICLR NAACL ACL ACL ACL EMNLP EMNLP NAACL NACL BlackboaNLP ICCV	https://www.aclweb.org/ https://www.aclweb.org/	https://arxiv.org/parkips/sept/sept/sept/sept/sept/sept/sept/se	192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Ov 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categoric 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualizing and Understanding Neural Machine Translation 201 203 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding Neural Models in NLP 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vige-e: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT Is not: Lessons from a new suite of psycholinguistic diagnostics 202 209 What can All do for me? evaluating machine learning interpretations in coope 201 210 What do Neural Machine Translation Models Learn about Morphology? 201 211 What do you learn from context? Probing for sentence structure in contextual 201 212 What Does BERT Learn about the Structure of Language? 201 213 What does bert look at? an analysis of bert's attention 201	B Arxiv B EMNLP-IJCNLP D ACL B BlackboxNLP F EMNLP F NAACL CAI-W F ICASSP D JAIR D ACL B NAACL EMNLP-IJCNLP EMNLP EMNLP F ECCV TACL D IUI F ACL B ICLR D ACL B ICLR D ACL B BlackboxNLP	cackesb.org/anthology/N1/ cackesb.org/anthology/N1/ corg/pdf/1809.02791.pdf/ corg/pdf/1809.02791.pdf/ corg/pdf/1809.08079.pdf/ corg/pdf/1809.08079.pdf/ corg/pdf/1809.08037.pdf/ https://arxiv.org/pdf/1809.08037.pdf/ https://arxiv.org/pdf/1809.08037.pdf/ https://arxiv.org/pdf/2010.07882.pdf/ https://arxiv.org/pdf/2005.01218.pdf/ https://www.ackeeb.org/anthology/N1/ https://arxiv.org/pdf/1908.05620.pdf/ https://arxiv.org/pdf/1908.05620.pdf/ https://arxiv.org/pdf/1907.13528.pdf/ https://arxiv.org/pdf/1907.13528.pdf/ https://arxiv.org/pdf/1907.13528.pdf/ https://arxiv.org/pdf/1907.06316.pdf/ https://arxiv.org/pdf/1907.06316.pdf/ https://arxiv.org/pdf/1907.06316.pdf/ https://arxiv.org/pdf/1908.06341.pdf/ https://arxiv.org/pdf/1908.06341.pdf/ https://arxiv.org/pdf/1908.06341.pdf/ https://arxiv.org/pdf/1908.06341.pdf/
sonance Between Human and Machine b Human Rationales Improve Machine b Human Rationales Improve Machine b Multi-hop Readers Dream of Reasoni b Neural Dialog Systems Use the Conver- bes it care what you asked? Understanding Importance bes String-Based Neural MT Learn Source Syntax? bes String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for En- sni: Natural language inference with natural language e- stNTS: An Neural Programmer-Interpreter Model for Se- fuce: Explaining model decisions through unsupervised rembling Visual Explanations for VQA seer: A benchmark to evaluate retionalized nip models relusting Explainable Al: Which Algorithmic Explanation relusting Explainable Al: Which Algorithmic Explanation relusting Explainable Al: Which Algorithmic Explanation relusting Inspection Engineering with Natural Lang plain Yourself Leveraing (XAL): An Empirical Study of plainable Active Learning (XAL): An Empirical Study of plainable Automated Fact-Checking for Public Health Co plainable Clinical Decision Support from Text plaining Black Box Predictions and Unveiling Data Artit plaining Character-Aware Neural Networks for Word-Le plaining Simple Natural Language Inference	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019 2017 2020 2020 as Help Users Predi 2020 and documents and 2018 lations in transform 2020 auage Explanations 2020	BISCKBONNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL ACL ACL ACL CSCW EMNLP EMNLP NAACL ACL EMNLP BISCKBONNLP BISCKBONNLP ACL	bitps://www.isi.edu/natural-language/milemnip1 https://www.isi.edu/natural-language/milemnip1 https://www.aciweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-10365.pdf https://www.aciweb.org/anthology/P18-1365.pdf https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd	KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivit 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning Interpretable Relationships between Entities, Relations of 132 Learning to Explain: Answering Why-Questions via Rephrasing 133 Learning to Explain: Datasets and Models for Identifying Valid Rei 136 Learning Variational Word Masks to Improve the Interpretability of 137 Lightly-supervised representation learning with global interpretabil 138 Linguistic Knowledge and Transferability of Contextual Represent 139 LISA: Explaining Recurrent Neural Network Judgments via Layer- 140 Localizing Moments in Video With Natural Language 141 Listrivis: A tool for visual analysis of hidden state dynamics in recurrent 143 Listrivis: A tool for visual analysis of hidden state dynamics in recurrent	2019 se of Explicit 2020 soning over 2019 2018 all Explanatt 2020 aming from 2020 ant bag-of-c 2019 2019 on 2019 on 2019 chine Reatt 2019 chine Reatt 2020 2018 2020 ament level 2019 and Concer 2020 2015 2019 assoning Ch 2020 Neural Tes 2020 ity 2019 attors 2019 wise Serns 2018 2017 arrent neurs 2017	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL ACL ACL ACL ACL ACL EMNLP EMNLP NAACL NACL NA	https://www.aclweb.org/ https://www.aclweb.org/	Interpolacy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 emolp-r 8,05601 pdf (antholocy/2020 emolp-r 8,04127 pdf (antholocy/2020 emolp-r (anthol	Drg/pdf/1902.10186.pdf 192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-Ov 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using "Annotator Rationales" to Improve Machine Learning for Text Categoric 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualizing and Understanding Neural Machine Translation 201 203 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding Neural Models in NLP 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vige-e: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT Is not: Lessons from a new suite of psycholinguistic diagnostics 202 209 What can All do for me? evaluating machine learning interpretations in coope 201 210 What do Neural Machine Translation Models Learn about Morphology? 201 211 What do you learn from context? Probing for sentence structure in contextual 201 212 What does bert look at? an analysis of bert's attention 201 213 What does this word mean? explaining contextualized embeddings with natu 201	B Arxiv B EMNLP-IJCNLP D ACL B BlackboxNLP F EMNLP F NAACL CAI-W F ICASSP D JAIR D ACL B NAACL EMNLP-IJCNLP EMNLP EMNLP F ECCV TACL D IUI F ACL D ICLR D ACL D ICLR D BlackboxNLP EMNLP EMNLP	c acleeb, orgianthologa(N1) c acleeb, orgianthologa(P1) corgipdf/1809.02701.pdf corgipdf/1809.02701.pdf corgipdf/1809.08079.pdf corgipdf/1809.08079.pdf corgipdf/1809.08079.pdf corgipdf/1809.08037.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/1711.10203.pdf https://arxiv.orgipdf/1711.10203.pdf https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.13528.pdf https://arxiv.orgipdf/1907.056316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf
sonance Between Human and Machine Human Rationales Improve Machine Multi-hop Readers Dream of Reasons Neural Dialog Systems Use the Converse is it care what you asked? Understanding Importance as String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Enri: Natural language inference with natural language is INTS: An Neural Programmer-Interpreter Model for Seuce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA aser: A benchmark to evaluate rationalized nlp models saluating and Characterizing Human Rationales saluating Explainable Al: Which Algorithmic Explanation aluating neural network explanation methods using hybrart: A visual analysis tool to explore learned represent pBERT: Representation Engineering with Natural Language Inference plainable Active Learning (KAL): An Empirical Study of plainable Active Learning (KAL): An Empirical Study of plainable Active Learning (KAL): An Empirical Study of plainable Clinical Decision Support from Text plaining Black Box Predictions and Unveiling Data Artificialing Character-Aware Neural Networks for Word-Leplaining Simple Natural Language Inference plaining Simple Natural Language Inference plaining the Stars: Weighted Multiple-Instance Learning plaining the Stars: Weighted Multiple-Instance Learning	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explanations 2018 intence Simplificatic 2019 2017 2020 2020 as Help Users Predi 2020 and documents and 2018 lations in transform 2020 auage Explanations 2020	BISCKBONNLP EMNLP EMNLP ACL NeuIPS ACL Arxiv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL ACL CSCW EMNLP EMNLP EMNLP EMNLP NAACL ACL EMNLP BISCKBONNLP ACL EMNLP	bitps://www.iai.edu/naturai-language/milemnio1 https://www.iai.edu/naturai-language/milemnio1 https://www.iai.edu/naturai-language/milemnio1 https://www.aciweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-10365.pdf https://www.aciweb.org/anthology/P18-1365.pdf https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd	KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning Interpretable negation rules via weak supervision at documents of the Company of the Company of the Company 134 Learning to Explain Entity Relationships in Knowledge Graphs 135 Learning to Explain: Datasets and Models for Identifying Valid Rea 136 Learning to Explain: Datasets and Models for Identifying Valid Rea 137 Lightly-supervised representation learning with global interpretability of Localizing Moments in Video With Natural Language 140 Localizing Moments in Video With Natural Language 141 Listrivis: A tool for visual analysis of hidden state dynamics in records.	2019 se of Explicit 2020 soning over 2019 2018 all Explanatt 2020 aming from 2020 anti bag-of-c 2019 2019 on 2019 on 2019 ochine React 2019 chine React 2020 2018 2020 ament level 2019 and Concec 2020 2015 2019 assoning Ch 2020 (Neural Tes 2020 (ity 2019 attors 2019 wise Serns 2018 2017 arrent neurs 2017 obal Explan 2020	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL ACL ACL ACL ACL ACL EMNLP EMNLP NAACL NACL NA	https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.aclweb.org/ https://www.research-co https://www.research-co https://www.aclweb.org/ https:	Interpolacy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 emolp-r 8,05601 pdf (antholocy/2020 emolp-r 8,04127 pdf (antholocy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 ed-main (anthol	192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using 'Annotator Rationales' to Improve Machine Learning for Text Categori, 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional salidagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 204 Visualizing and Understanding Neural Mechine Translation 201 205 Visualizing and Understanding Neural Mechine Translation 101 206 Visualizing and Understanding Neural Mechine Translation NLP 201 207 Vige-e: Explaining, elaborating, and enhancing your answers for visual quest 201 208 What BERT 1 not: Leasons from a new suite of psycholinguistic diagnostics 202 209 What do Neural Machine Translation Models Learn about Morphology? 201 210 What do Neural Machine Translation Models Learn about Morphology? 201 211 What do you learn from context? Probing for sentence structure in contextual 201 212 What does this word mean? explaining contextualized embeddings with natu 201 213 What does this word mean? explaining contextualized embeddings with natu 201 215 What is one grain of sand in the desert? analyzing individual neurons in dee; 201	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP F EMNLP F NAACL F LICAI-W F ICASSP D JAIR D ACL B NAACL EMNLP-IJCNLP EMNLP D EMNLP D EMNLP D IUI F ACL D IUI F ACL D ICLR D ACL D BlackboxNLP EMNLP D EMNLP	c acleeb, orgianthology/N1 c acleeb, orgianthology/N1 corgipdf/1909, 02701 pdf corgipdf/1909, 02701 pdf corgipdf/1909, 02701 pdf corgipdf/1909, 08037 pdf corgipdf/1909, 08037 pdf corgipdf/1909, 08037 pdf corgipdf/1909, 08037 pdf https://arxiv.orgipdf/2010, 07892 pdf https://arxiv.orgipdf/2010, 07892 pdf https://arxiv.orgipdf/2005, 01218 pdf https://arxiv.orgipdf/1911, 10203 pdf https://arxiv.orgipdf/1711, 10203 pdf https://arxiv.orgipdf/1908, 05620 pdf https://arxiv.orgipdf/1908, 05620 pdf https://arxiv.orgipdf/1908, 05620 pdf https://arxiv.orgipdf/1907, 13528 pdf https://arxiv.orgipdf/1905, 06316 pdf https://arxiv.orgipdf/1905, 06316 pdf https://arxiv.orgipdf/1905, 04341 pdf
sonance Between Human and Machine Human Rationales Improve Machine Multi-hop Readers Dream of Reasons Neural Dialog Systems Use the Converse is it care what you asked? Understanding Importance is String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Enri: Natural language inference with natural language is INTS: An Neural Programmer-Interpreter Model for Select Explaining model decisions through unsupervised sembling Visual Explanations for VQA isser. A benchmark to evaluate rationalized nlp models aluating and Characterizing Human Rationales aluating Explainable Al: Which Algorithmic Explanation sluating neural network explanation methods using hybert: A visual analysis tool to explore learned represent plainable Active Learning (KAL): An Empirical Study of plainable Clinical Decision Support from Text plaining Character-Aware Neural Networks for Word-Lealining Character-Aware Neural Networks for Word-Lealining Simple Natural Language Inference plaining Structure in Representation of Named Entities of the Stans: Weighted Multiple-Instance Learning sloiting Structure in Representation of Named Entities of the Structure in Representation of	Attention is of Werbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explainations 2018 intence Simplificatic 2019 2017 2020 2020 as Help Users Predi 2020 and documents and 2018 lations in transform 2020 and documents and 2018 lations in transform 2020 and Explanations 2020 and 2020 and 2020 and 2020 and 2018 lacts through influer 2020 aveil Prediction: Do 2018 ased Deep Architec 2018 2019 g for Aspect-Based 2014 using Active Learnis 2018	BISCREONNLP EMINLP EMINLP ACL NeuIPS ACL Arriv NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL ACL CSCW EMNLP EMNLP EMNLP EMNLP EMNLP EMNLP COLING	bitps://www.isi.edu/natural-language/milemnilp1 https://www.isi.edu/natural-language/milemnilp1 https://www.aciweb.org/arithology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.aciweb.org/arithology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-1032.pdf https://www.aciweb.org/arithology/P18-10365.pdf https://www.aciweb.org/arithology/P18-1365.pdf https://www.aciweb.org/arithology/W18-5403.pd https://www.aciweb.org/arithology/W18-5403.pd https://www.aciweb.org/arithology/W18-5403.pd https://www.aciweb.org/arithology/W18-5403.pd https://www.aciweb.org/arithology/W18-5403.pd https://www.aciweb.org/arithology/W18-5403.pd https://www.aciweb.org/arithology/W18-1062.pd	KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Matching 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning from Explanations with Neural Execution Tree 131 Learning Interpretable Relationships between Entities, Relations of 132 Learning to Explain: Answering Why-Questions via Rephrasing 133 Learning to Explain: Datasets and Models for Identifying Valid Rea 136 Learning to Explain: Datasets and Models for Identifying Valid Rea 137 Lightly-supervised representation learning with global interpretability of 138 Linguistic Knowledge and Transferability of Contextual Represent 139 LISA: Explaining Recurrent Neural Network Judgments via Layer 140 Localizing Moments in Video With Natural Language 141 Listrivis: A tool for visual analysis of hidden state dynamics in recurrent 142 Machine Guides, Human Supervises: Interactive Learning with Git 143 MathQA: Towards interpretable math word problem solving with o	2019 se of Explicit 2020 soning over 2019 2018 all Explanatt 2020 aming from 2020 anti bag-of-c 2019 2019 on 2019 on 2019 ochine Reest 2019 chine Reest 2020 2018 2020 ament level 2019 and Concest 2020 2015 2019 assoning Ch 2020 (Neural Tes 2020 (Ry 2019 attions 2019 witse Serns 2018 2017 arrent neurs 2017 obal Explar 2020 peration-bs 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL ACL ACL ACL ACL ACL EMNLP EMNLP NAACL NACL NACL NAACL	https://www.aclweb.org/ https://www.aclweb.org/	Interpolacy/2020 emolo-r (antholocy/2020 emolo-r (antholocy/2020 emolo-r (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 emolo-r 8,04127 pdf (antholocy/2020 emolo-r 8,04127 pdf (antholocy/2020 emolo-r (antholocy/2020 emolo-r (antholocy/2020 emolo-r (antholocy/2020 emolo-r (antholocy/2020 emolo-r (antholocy/2020 et-main (anthol	192 Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using 'Annotator Rationales' to Improve Machine Learning for Text Categori, 200 200 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional salidagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 204 Visualizing and Understanding Neural Machine Translation 201 205 Visualizing and Understanding Neural Mechine Translation 201 206 Visualizing and Understanding Neural Mechine Translation 201 207 Vige-e: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT is not: Lessons from a new suite of psycholinguistic diagnostics 202 209 What can Al do for me? evaluating machine learning interpretations in coope 201 210 What do Neural Machine Translation Models Learn about Morphology? 201 211 What do you learn from context? Probing for sentence structure in contextual 201 212 What does bert look at? an analysis of bert's attention 201 213 What does this word mean? explaining contextualized embeddings with natu 201 214 What does this word mean? explaining contextualized embeddings with natu 201 215 What is one grain of sand in the desert? analyzing individual neurons in deep 201 216 What you can cram into a single \$889 vector: Probing sentence embeddings 201	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP F EMNLP F NAACL F LICAI-W F ICASSP D JAIR D NAACL EMNLP-IJCNLP EMNLP EMNLP D EMNLP D EMNLP D IUI F ACL D ICLR D ICLR D ICLR D ICLR D ACL D BlackboxNLP EMNLP D EMNLP D AAAI	Lackweb, org/anthology/N1 Lackweb, org/anthology/P1 Long/pdf/1809.02701.pdf Long/pdf/1809.02701.pdf Long/pdf/1809.02701.pdf Long/pdf/1809.08079.pdf Long/pdf/1809.08037.pdf https://anxiv.org/pdf/2010.07882.pdf https://anxiv.org/pdf/2010.07882.pdf https://anxiv.org/pdf/2010.07882.pdf https://anxiv.org/pdf/2005.01218.pdf https://anxiv.org/pdf/2005.01218.pdf https://anxiv.org/pdf/2005.01218.pdf https://anxiv.org/pdf/2005.01218.pdf https://anxiv.org/pdf/2005.01208.pdf https://anxiv.org/pdf/2005.01208.pdf https://www.ackweb.org/anthology/P1 https://www.ackweb.org/anthology/P1 https://www.ackweb.org/anthology/P1 https://www.ackweb.org/anthology/P1 https://www.ackweb.org/anthology/P1 https://www.ackweb.org/anthology/P1 https://anxiv.org/pdf/2010.06775.pdf https://anxiv.org/pdf/1810.09648.pdf
sonance Between Human and Machine Human Rationales Improve Machine Delationales Dream of Reasonin Neural Dialog Systems Use the Converse String-Based Neural MT Learn Source Syntax? es String-Based Neural MT Learn Source Syntax? CA: Decision Tree-based Co-Attention Networks for Explicit Natural language inference with natural language et INTS: An Neural Programmer-Interpreter Model for Seluce: Explaining model decisions through unsupervised sembling Visual Explanations for VQA aser: A benchmark to evaluate rationalized nip models saluating and Characterizing Human Rationales aluating Explainable Al: Which Algorithmic Explanation aluating replainable Al: Which Algorithmic Explanation BERT: Representation Engineering with Natural Language Models for Committee Programmer Language Models for Committee Programmer Programmer Programmer Delainable Active Learning (KAL): An Empirical Study of plainable Automated Fact-Checking for Public Health Coplainable Clinical Decision Support from Text plaining Black Box Predictions and Unveiling Data Artificationing Character-Aware Neural Networks for Word-Leplaining Character-Aware Neural Networks for Word-Leplaining non-linear Classifier Decisions within Kernel-burning Classifier Decisions within Ke	Attention is of Verbs in Deep Li 2018 2016 2016 2016 2016 explainable Claim Ve 2020 explainations 2018 intence Simplificatic 2019 2017 2020 2020 as Help Users Predi 2020 and documents and 2018 defid documents and 2018 defid documents and 2018 defid documents and 2020 and Explanations 2020 and 2020 defid 2018 decis through Influer 2020 aveil Prediction: Do 2018 decis through Influer 2020	BIBICKBONNLP EMINLP EMINLP ACL NeuIPS ACL Arity NIPS-W ACL EMNLP ACL ACL ACL ACL ACL ACL ACL CSCW EMNLP EMNLP EMNLP EMNLP EMNLP EMNLP COLING ACL	bitps://www.iai.edu/naturai-language/milemnio1 https://www.iai.edu/naturai-language/milemnio1 https://www.iai.edu/naturai-language/milemnio1 https://www.aciweb.org/anthology/D16-1159.pdf https://arxiv.org/pdf/1812.01193.pdf https://arxiv.org/pdf/1812.01193.pdf https://www.aciweb.org/anthology/P19-1331.pdf https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.cs.utexas.edu/-mil/papers/rajani.vig https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-1032.pdf https://www.aciweb.org/anthology/P18-10365.pdf https://www.aciweb.org/anthology/P18-1365.pdf https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd https://www.aciweb.org/anthology/W18-5403.pd	KERMIT: Complementing Transformer Architectures with Encoder 120 Knowledge Aware Conversation Generation with Explainable Rea 121 Latent alignment and variational attention 122 Leakage-Adjusted Simulatability: Can Models Generate Non-Trivia 123 LEAN-LIFE: A Label-Efficient Annotation Framework Towards Lea 124 Learning concept embeddings for dataless classification via efficie 125 Learning Corresponded Rationales for Text Matching 126 Learning credible deep neural networks with rationale regularization 127 Learning Dynamics of Attention: Human Prior for Interpretable Ma 128 Learning Explainable Linguistic Expressions with Neural Inductive 129 Learning Explanations from Language Data 130 Learning Interpretable negation rules via weak supervision at documents of the Company of the Company of the Company 134 Learning to Explain Entity Relationships in Knowledge Graphs 135 Learning to Explain: Datasets and Models for Identifying Valid Rea 136 Learning to Explain: Datasets and Models for Identifying Valid Rea 137 Lightly-supervised representation learning with global interpretability of Localizing Moments in Video With Natural Language 140 Localizing Moments in Video With Natural Language 141 Listrivis: A tool for visual analysis of hidden state dynamics in records.	2019 se of Explicit 2020 soning over 2019 2018 all Explanat 2020 aming from 2020 anti bag-of-c 2019 2019 on 2019 on 2019 on 2019 on 2018 2020 2018 2020 ament level 2019 and Concep 2020 2015 2019 assoning Ch 2020 (Nounal Tes 2020 (Ry 2019 attions 2019 attions 2019 wise Serma 2018 2017 arrent neurs 2017 obal Explar 2020 penation-bas 2019 2019	NAACL EMNLP EMNLP NeuIPS EMNLP Findings ACL KIS ICLR ICDM NeuIPS EMNLP BlackboxNLP ICLR NAACL ACL ACL ACL ACL ACL EMNLP EMNLP NAACL NACL NA	https://www.aclweb.org/ https://www.aclweb.org/	Interpolacy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 emolp-r (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 ed-dem 2,03342 pdf (antholocy/2020 emolp-r 8,04127 pdf (antholocy/2020 emolp-r 8,04127 pdf (antholocy/2020 emolp-r (anthol	Understanding Neural Abstractive Summarization Models via Uncertainty 202 193 Understanding neural networks through representation erasure 201 194 Universal adversarial triggers for attacking and analyzing NLP 201 195 Unsupervised Alignment-based literative Evidence Retrieval for Mutti-hop Qu 202 196 Unsupervised Discrete Sentence Representation Learning for Interpretable N 201 197 Unsupervised Token-wise Alignment to Improve Interpretation of Encoder-De 201 198 Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguati 201 199 Using 'Annotator Rationales' to Improve Machine Learning for Text Categori, 201 201 Using Explanations to Improve Ensembling of Visual Question Answering Sy 201 201 Using regional saliency for speech emotion recognition. 201 202 Visualization and'diagnostic classifiers' reveal how recurrent and recursive ni 201 203 Visualizing and Understanding Neural Machine Translation 201 204 Visualizing and Understanding Neural Models in NLP 201 205 Visualizing and Understanding the Effectiveness of BERT 201 206 Vokenization: Improving Language Understanding with Contextualized, Visual 202 207 Vige-s: Explaining, elaborating, and enhancing your answers for visual questi 201 208 What BERT is not Lessons from a new suits of psycholinguistic diagnostics 202 209 What can All do for me? evaluating machine learning interpretations in coope 201 210 What do Neural Machine Translation Models Learn about Morphology? 201 211 What do you learn from context? Probing for sentence structure in contextual 201 212 What does bet look at? an analysis of bert's attention 201 213 What does this word mean? explaining contextualized embeddings with natu 201 214 What does this word mean? explaining contextualized embeddings with natu 201 215 What sone grain of sand in the desert? analyzing individual neurons in dee; 201 216 What you can cram into a single \$810 vector: Probing sentence embeddings 201 217 Why Attention is Not Explanation: Surgical Intervention and Causal Reasonir 202	B Arxiv B EMNLP-IJCNLP D ACL B ACL B BlackboxNLP F EMNLP F NAACL F LICAI-W F ICASSP D JAIR D NAACL EMNLP-IJCNLP EMNLP EMNLP D EMNLP D EMNLP D IUI F ACL D ICLR D ICLR D ICLR D ICLR D ACL D BlackboxNLP EMNLP D EMNLP D AAAI	Lackweb, orgianthology/N1 Lackweb, orgianthology/N1 Lackweb, orgianthology/P1 Largipdf/1808.02701.pdf Largipdf/1808.02701.pdf Largipdf/1808.08079.pdf Largipdf/1808.08079.pdf Largipdf/1808.08079.pdf Largipdf/1809.08037.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2010.07882.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://arxiv.orgipdf/2005.01218.pdf https://www.ackweb.orgianthology/P1 https://www.ackweb.orgianthology/N1 https://www.ackweb.orgianthology/P1 https://www.ackweb.orgianthology/P1 https://www.ackweb.orgianthology/P1 https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1908.05620.pdf https://arxiv.orgipdf/1810.09648.pdf https://arxiv.orgipdf/1810.09648.pdf https://arxiv.orgipdf/1810.09648.pdf https://arxiv.orgipdf/1810.09648.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.06316.pdf https://arxiv.orgipdf/1905.04341.pdf https://arxiv.orgipdf/1905.04341.pdf https://arxiv.orgipdf/1812.09355.pdf

Matching XAI Papers with XAI Question Bank?

43 User Questions

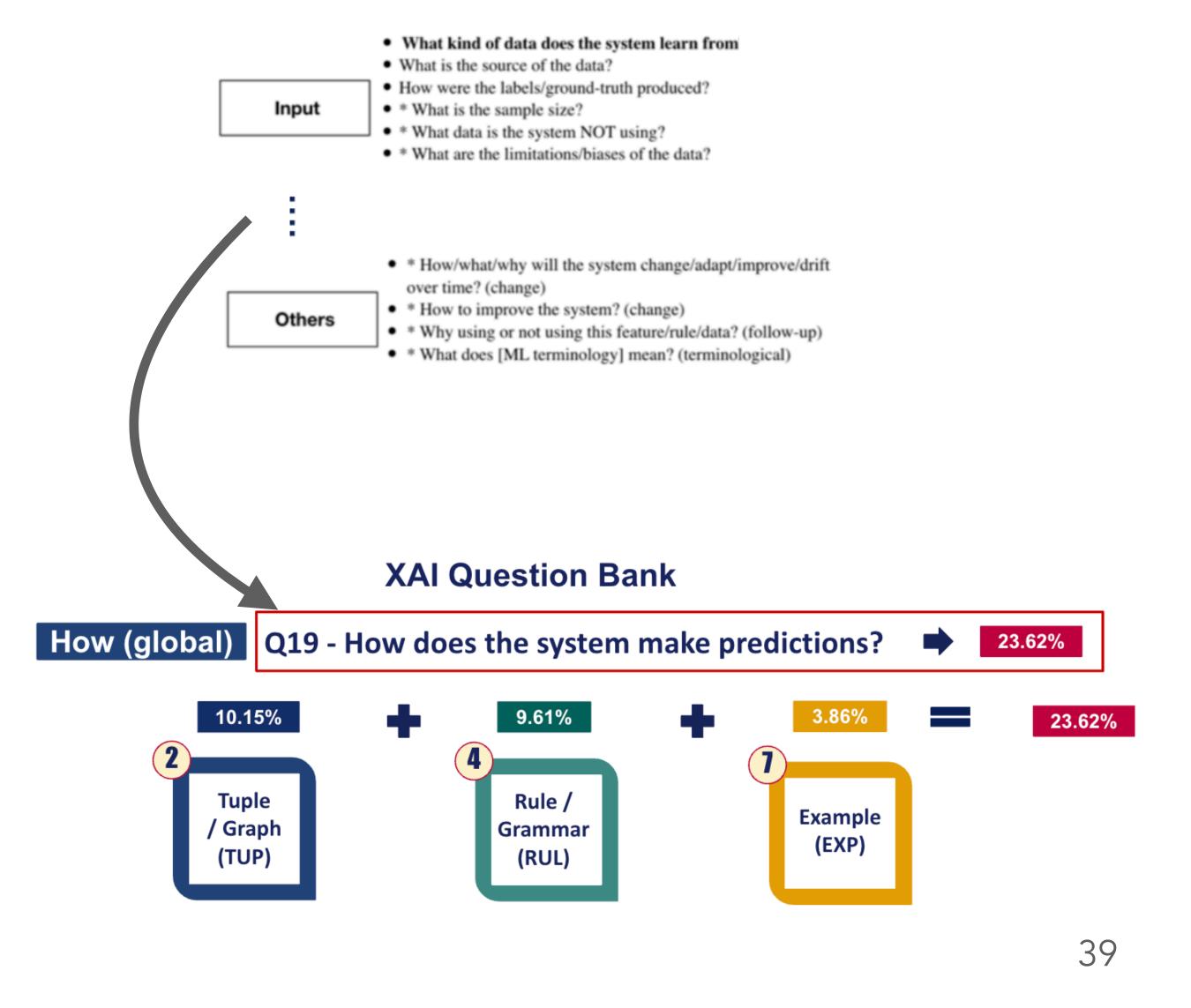
218 XAI Papers

	What kind of data does the system learn from?	1	ID	Title	Year	Venue	Paper URL
	What is the source of the data?		1	"Why should I trust you?" Explaining the predictions of any classifier	2016	KDD	https://arxiv.org/pdf/1602.04938.
Input	How were the labels/ground-truth produced? * What is the sample size?	•	2	Visualizing and Understanding Neural Models in NLP	2016	NAACL	https://www.aclweb.org/antholog
IIIput	 * What is the sample size? * What data is the system NOT using? 		3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taole
	* What are the limitations/biases of the data?	\ •	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.
	• * How/what/why will the system shapes/adent/improve/drift						
	 * How/what/why will the system change/adapt/improve/drift over time? (change) 		214	How much should you ask? On the question structure in QA systems	2018	BlackboxNLP	https://arxiv.org/pdf/1809.03734.
Others	* How to improve the system? (change)		215	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	https://arxiv.org/pdf/2011.06854.j
Others	 * Why using or not using this feature/rule/data? (follow-up) 		216	A Survey of the State of Explainable AI for Natural Language Processing	2020	AACL-IJCNLP	https://arxiv.org/pdf/2010.00711.j
	 * What does [ML terminology] mean? (terminological) 		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 ...

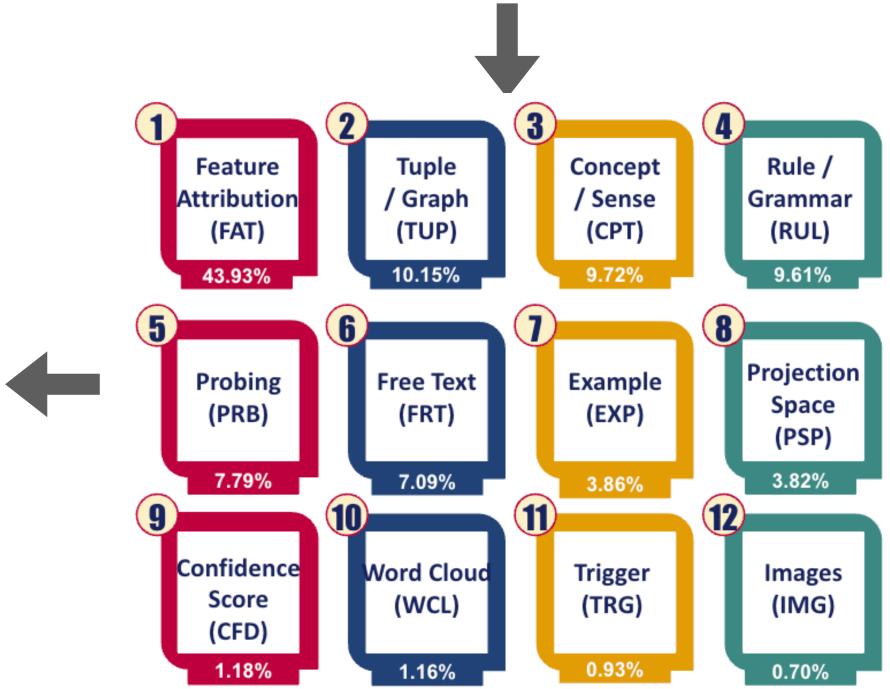
Matching Each User Question with XAI Forms in NLP

43 User Questions



2001 VAL Damaka

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/antholo
3	Rationalizing Neural Predictions	2016	EMNLP	https://people.csail.mit.edu/taol
4	BERT Rediscovers the Classical NLP Pipeline	2019	ACL	https://www.aclweb.org/antholo
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
	How much should you ask? On the question structure in QA systems Interpretable Multi-dataset Evaluation for Named Entity Recognition	2018 2020	BlackboxNLP	https://arxiv.org/pdf/1809.03734 https://arxiv.org/pdf/2011.06854
215	· · · · · · · · · · · · · · · · · · ·	2020		https://arxiv.org/pdf/2011.06854
214 215 216 217	Interpretable Multi-dataset Evaluation for Named Entity Recognition	2020	EMNLP	



Results: Heatmap for XAI Question Bank

		0.0%					1
	4.14/14.1	EVB	0.000/			VEAT/EDT/EVD	74.700/
	1-What kind of data does the system learn from?	EXP	3.86%			P/FAT/FRT/EXP	74.70%
	2-What is the source of the data?		*	W/by//	24-What instance feature leads to the system's prediction?	FAT	43.99%
Input/Data	3-How were the labels/ground-truth produced?		*	Why / Why not		P/FAT/FRT/EXP	74.70%
(0.55%)	4-What is the sample size?		*	(45.14%)	26-Why/how is this instance NOT predicted?	TRG	0.93%
	5-What data is the system NOT using?		•		27-Why is the instance predicted P instead of Q?	TRG	0.93%
	6-What are the limitations/biases of the data?		•		28-Why are [instance A and B] given different predictions? TRG/RUL/T	UP/FAT/FRT/EXP	75.62%
	7-How much data [like this] is the system trained on?		*		29-What would the system predict if this instance changes to?	CFD/EXP/TRG	5.97%
	8-What kind of output does the system give?	EXP	3.86%		30-What would system predict if this instance feature changes to?	CFD/FAT/TRG	46.10%
Output	9-What does the system output mean?		*		31-What would the system predict for [a different instance]?	CFD/TRG	2.11%
	10-How can I best utilize the output of the system?		•	What if /	32-How should this instance change to get a different prediction?	TRG	0.93%
(0.77%)	11-What is the scope of the system's capability?		•		33-How should instance feature change to get different prediction?	TRG	0.93%
	12-How's the output used for other systems modules?		•	(15.54%)	0.4 MHz + Lind of instance and a different and instance	TRG/EXP	4.79%
	13-How accurate/precise/reliable are the predictions?	CFD	1.18%		35-What's the scope of change permitted to get the same prediction?	TRG	0.93%
	14-How often does the system make mistakes?		*		36-What's the highest feature can have to get the same prediction?	TRG/FAT	44.91%
erformance	15-In what situations is the system to be incorrect?	CFD/EXP/TRG	5.97%		37-What is necessary feature present to guarantee this prediction?	TRG/FAT	44.91%
(2.03%)	16-What are the limitations of the system?		•		38-What kind of instance gets this prediction?	EXP	3.86%
	17-What kind of mistake is the system likely to make?	EXP	5.05%		39-How/what/why will the system change/improve/drift over time?		•
	18-Is the system's performance good enough for?		•	Others	40-How to improve the system?		•
	19-How does the system make predictions?	TUP/RUL/EXP	23 63%	(11.49%)	41-Why using or not using this feature/rule/data?	FAT/RUL/EXP	57.46%
How	20-What features does the system consider?	FAT	43.99%		42-What does [ML terminology] mean?		*
(Global)			53.60%		43-What are the results of other people using the system?		•
(30.31%)	21-What is the system's overall logic?	KULIFAT	33.00 /6	1			
	22-What kind of algorithm is used?		*				



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

Results: Heatmap for XAI Question Bank

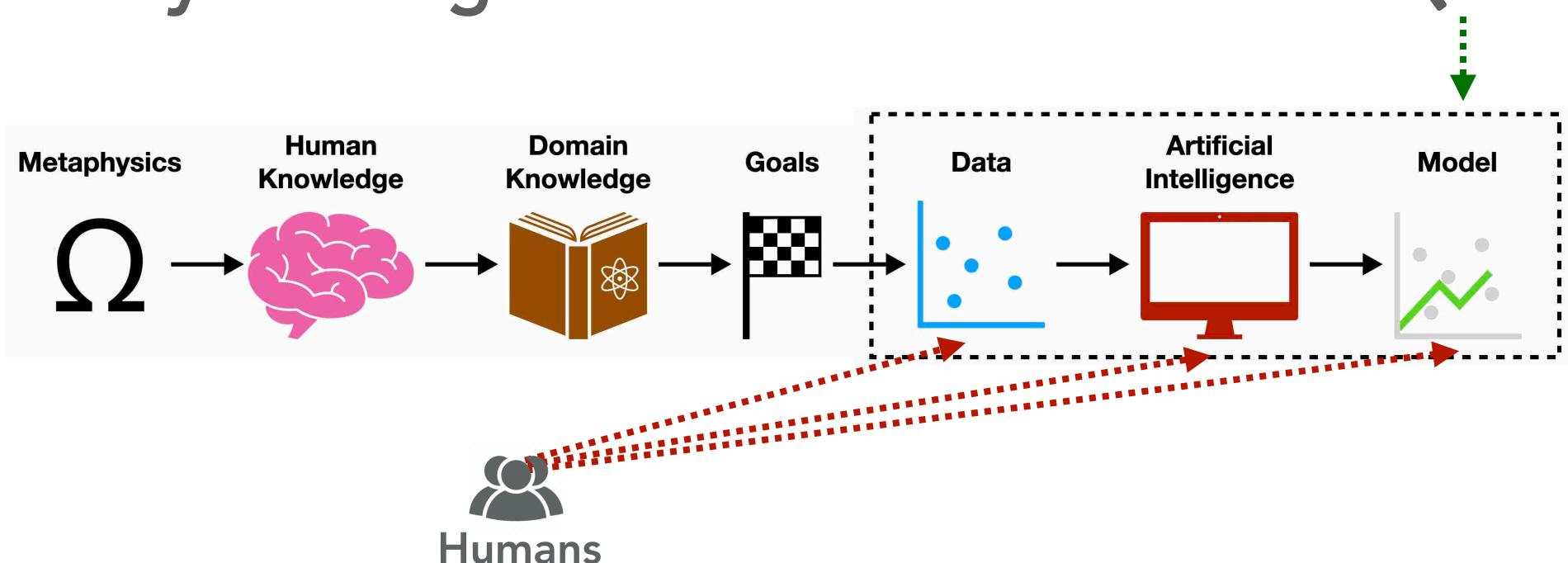
•				7			
	1-What kind of data does the system learn from?	EXP	3.86%		23-Why/how is this instance given this prediction?	TUP/FAT/FRT/EXP	74.70%
	2-What is the source of the data?	*			24-What instance feature leads to the system's prediction?	FAT	43.99%
	3-How were the labels/ground-truth produced?		*	Why / Why not	25-Why are [instance A and B] given the same prediction?	TUP/FAT/FRT/EXP	74.70%
Input/Data (0.55%)	4-What is the sample size?		*		26-Why/how is this instance NOT predicted?	TRG	0.93%
(0.0070)	5-What data is the system NOT using?		•		27-Why is the instance predicted P instead of Q?	TRG	0.93%
	6-What are the limitations/biases of the data?	•			28-Why are [instance A and B] given different predictions? TRG/RI	JL/TUP/FAT/FRT/EXP	75.62%
	7-How much data [like this] is the system trained on?		*		29-What would the system predict if this instance changes to?	CFD/EXP/TRG	5.97%
	8-What kind of output does the system give?	EXP			30-What would system predict if this instance feature changes to?	CFD/FAT/TRG	46.10%
	9-What does the system output mean?		*		31-What would the system predict for [a different instance]	? CFD/TRG	2.11%
Output	10-How can I best utilize the output of the system?		•		32-How should this instance change to get a different prediction?	TRG	0.93%
(0.77%)	11-What is the scope of the system's capability?		•	What if /	33-How should instance feature change to get different prediction?	TRG	0.93%
	12-How's the output used for other systems modules?		•	(15.54%)	34-What kind of instance gets a different prediction?	TRG/EXP	4.79%
				, 1010 1,0,	35-What's the scope of change permitted to get the same prediction?	TRG	0.93%
	13-How accurate/precise/reliable are the predictions?	CFD			36-What's the highest feature can have to get the same prediction?	TRG/FAT	44.91%
	14-How often does the system make mistakes?	AFD/FVD TDA	*	-	37-What is necessary feature present to guarantee this prediction?	TRG/FAT	44.91%
Performance (2.03%)	15-In what situations is the system to be incorrect?	CFD/EXP/TRG	5.97%		38-What kind of instance gets this prediction?	EXP	3.86%
(2.0570)	16-What are the limitations of the system?		•		39-How/what/why will the system change/improve/drift over time?		•
	17-What kind of mistake is the system likely to make?	EXP	5.05%		40-How to improve the system?		•
	18-Is the system's performance good enough for?		•	Others	41-Why using or not using this feature/rule/data?	FAT/RUL/EXP	57.46%
	19-How does the system make predictions?	TUP/RUL/EXP	23.63%	(11.49%)	42-What does [ML terminology] mean?		*
How (Global)	20-What features does the system consider?	FAT	43.99%		43-What are the results of other people using the system?		•
(30.31%)	21-What is the system's overall logic?	RUL/FAT	53.60%		To Trinat are the recalls of other people using the system?		
	22-What kind of algorithm is used?		*				

100.0%



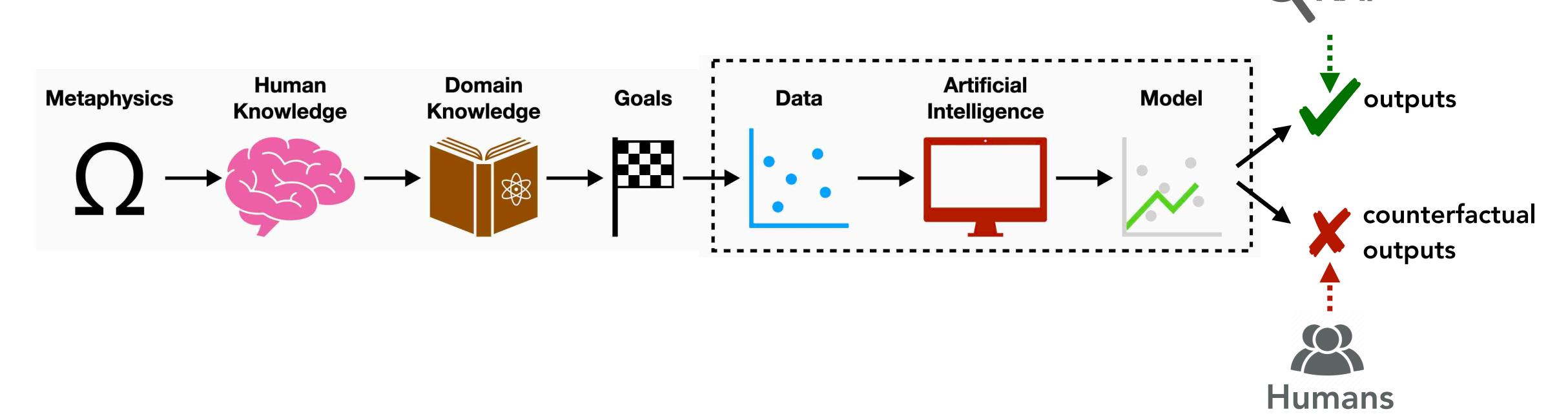
XAI largely ignored: diverse information across AI development process (data, model, deployment, etc.)

Take-away Message



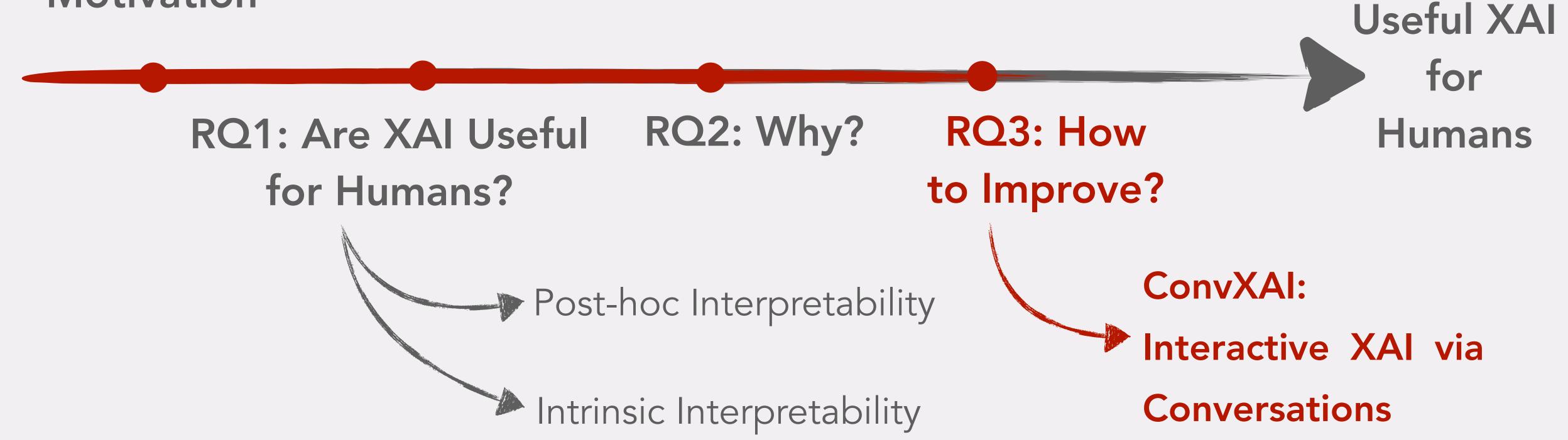
Users demand diverse XAI questions to understand AI models across AI development lifecycle, whereas existing XAI methods commonly answer only ONE XAI question.

Take-away Message



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

Background & Motivation





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



Hua Shen



Chieh-Yang Huang

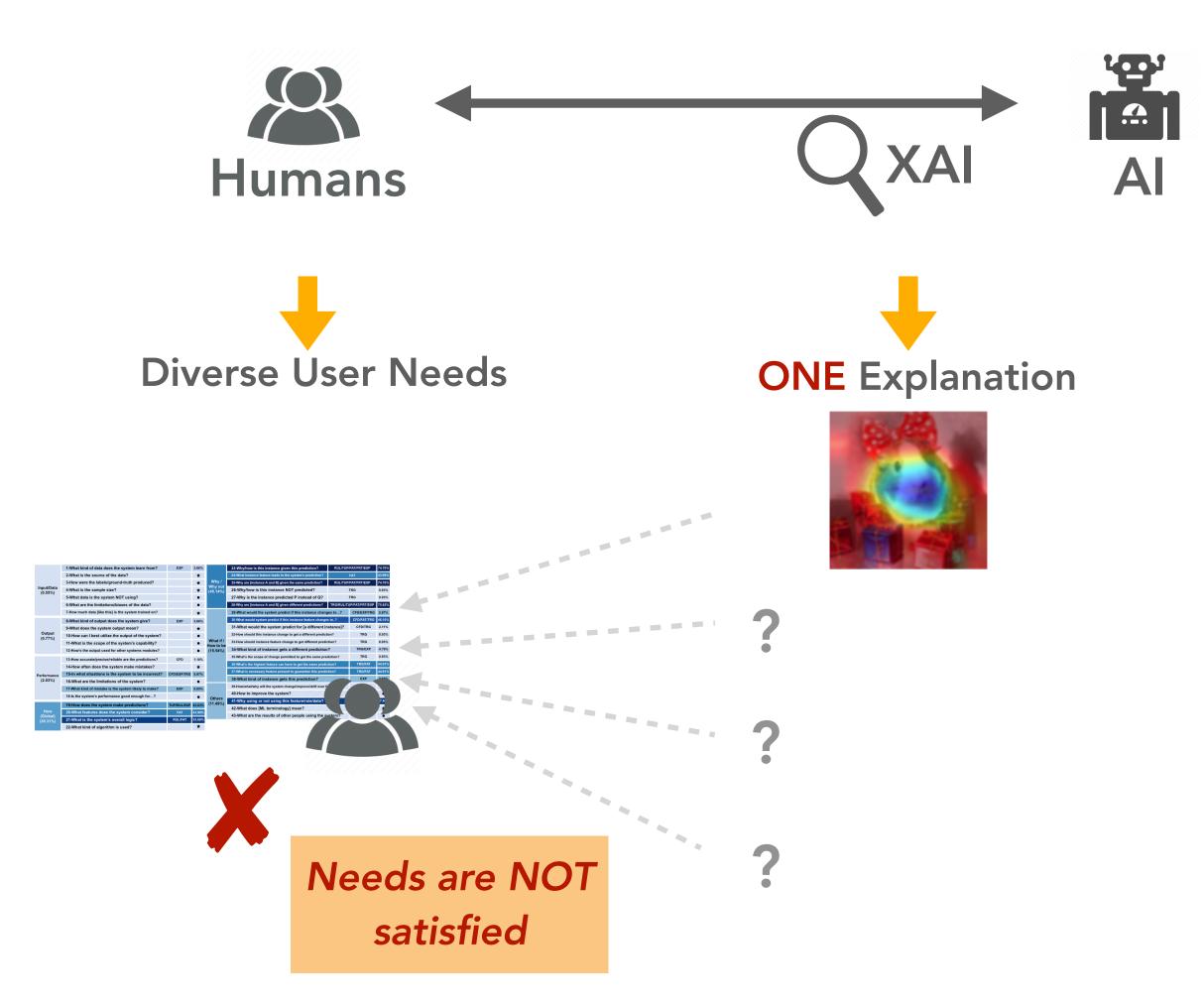


Sherry Wu



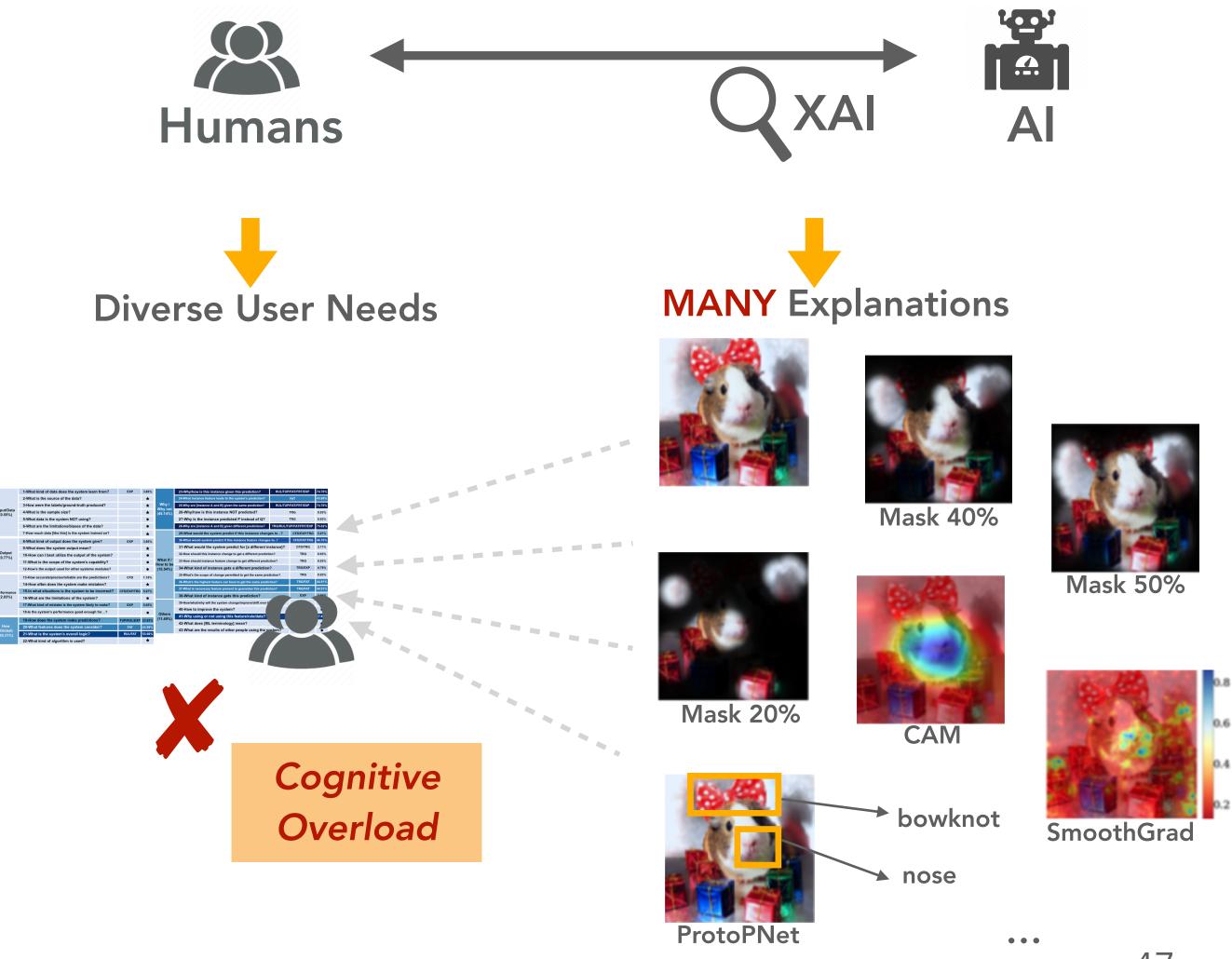
Kenneth Huang

Challenges of Existing XAI



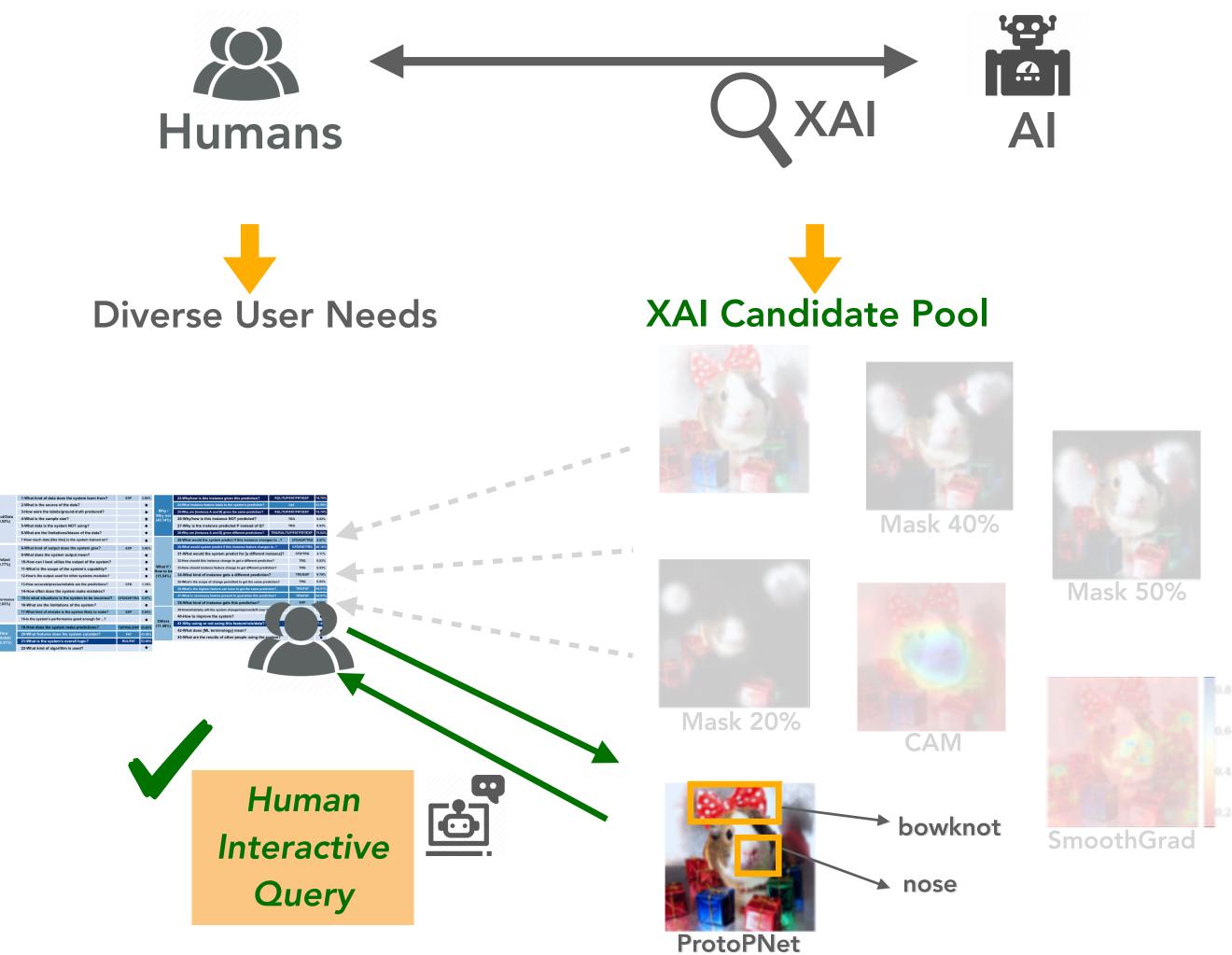
 Showing ONE specific explanation might NOT meet diverse XAI user needs.

Challenges of Existing XAI



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

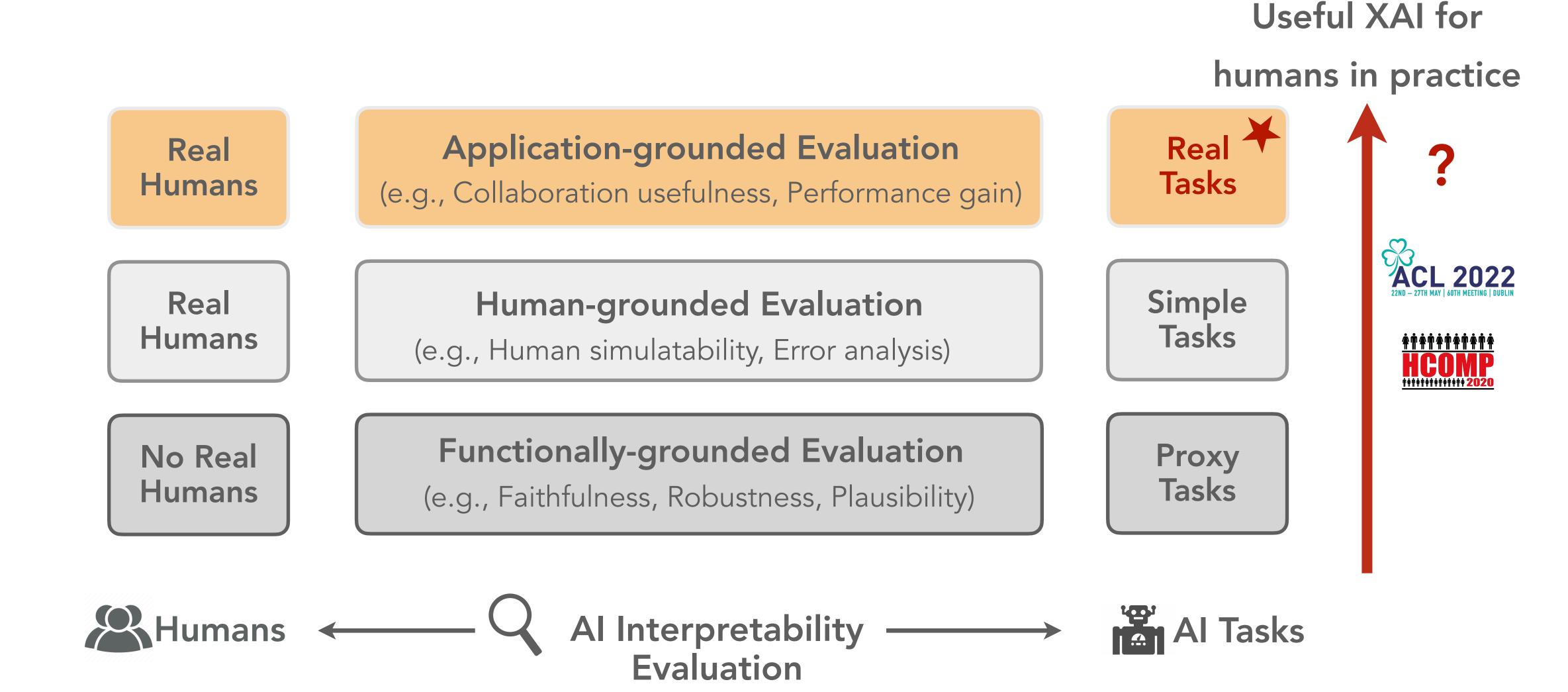
Solution: Conversational XAI



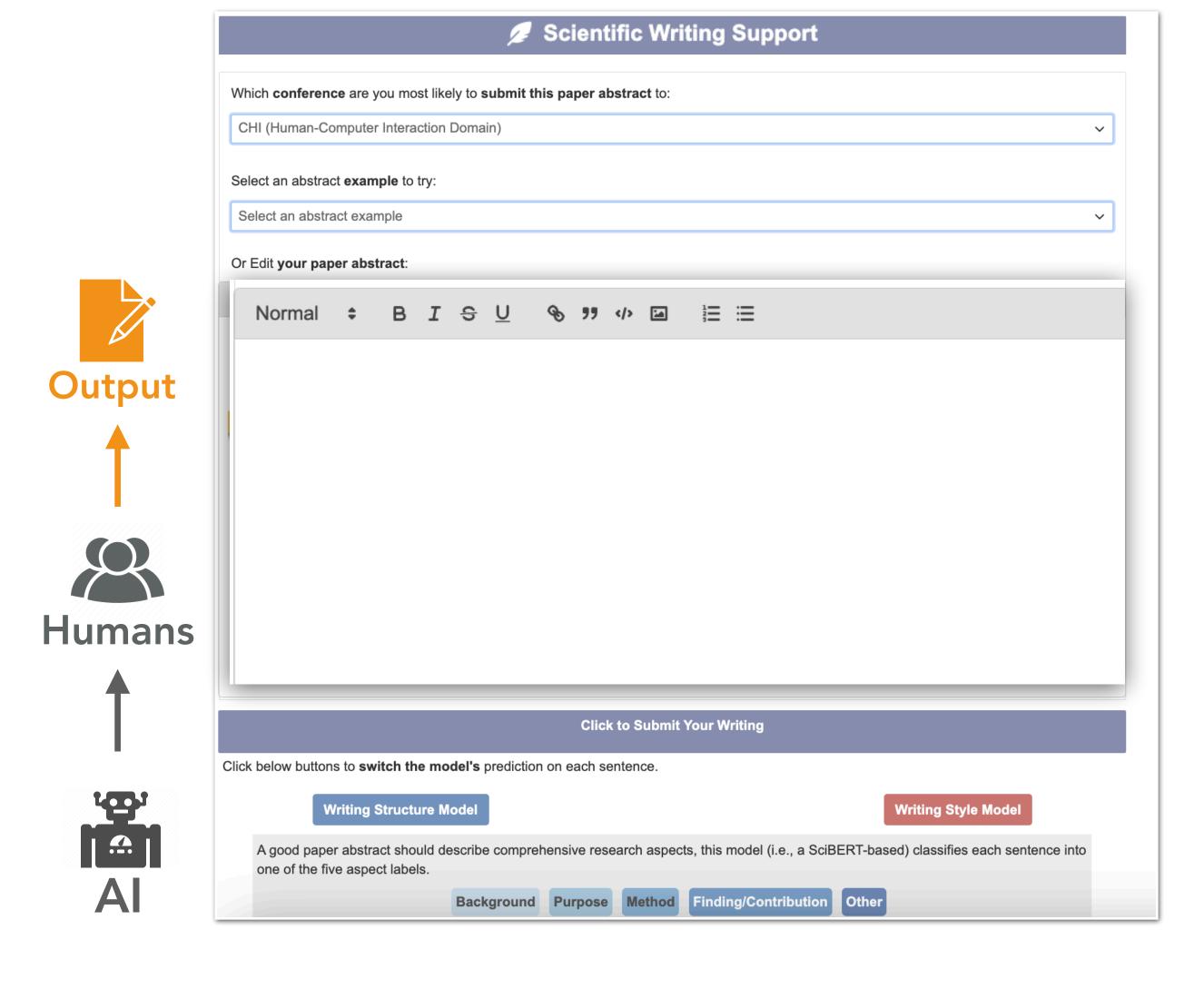
- Showing ONE specific explanation might NOT meet diverse XAI user needs.
- Showing MANY explanations at one time may lead to cognitive overload for humans

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

What Task should we apply Conversational XAI?



Real Task: Al-assisted scientific writing by humans

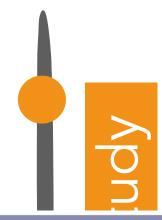


Reasons:

- 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;

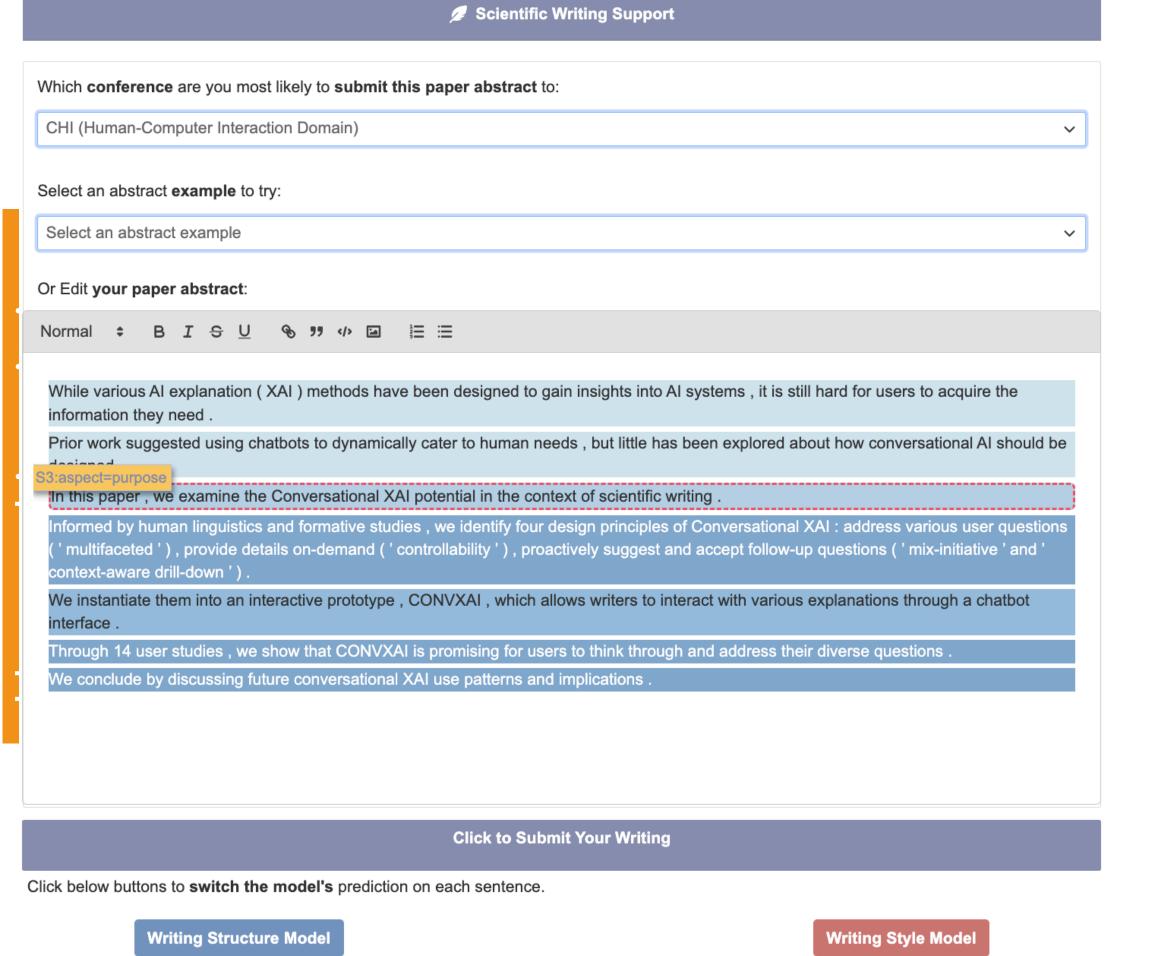
How to design the system?

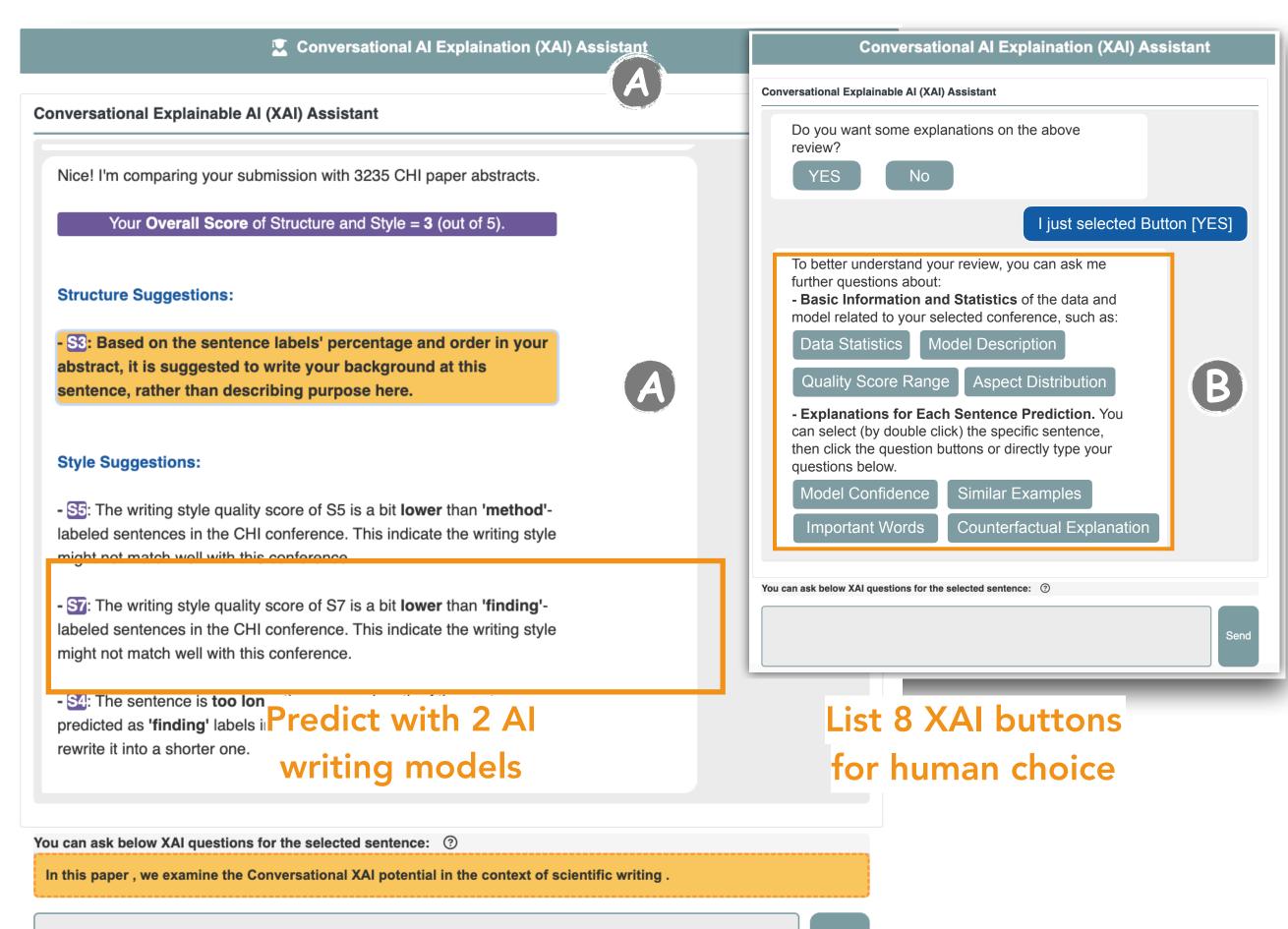
Goal: useful conversational XAI for AI-assisted scientific writing



A Preliminary System

Predicting with 2 Al writing models





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

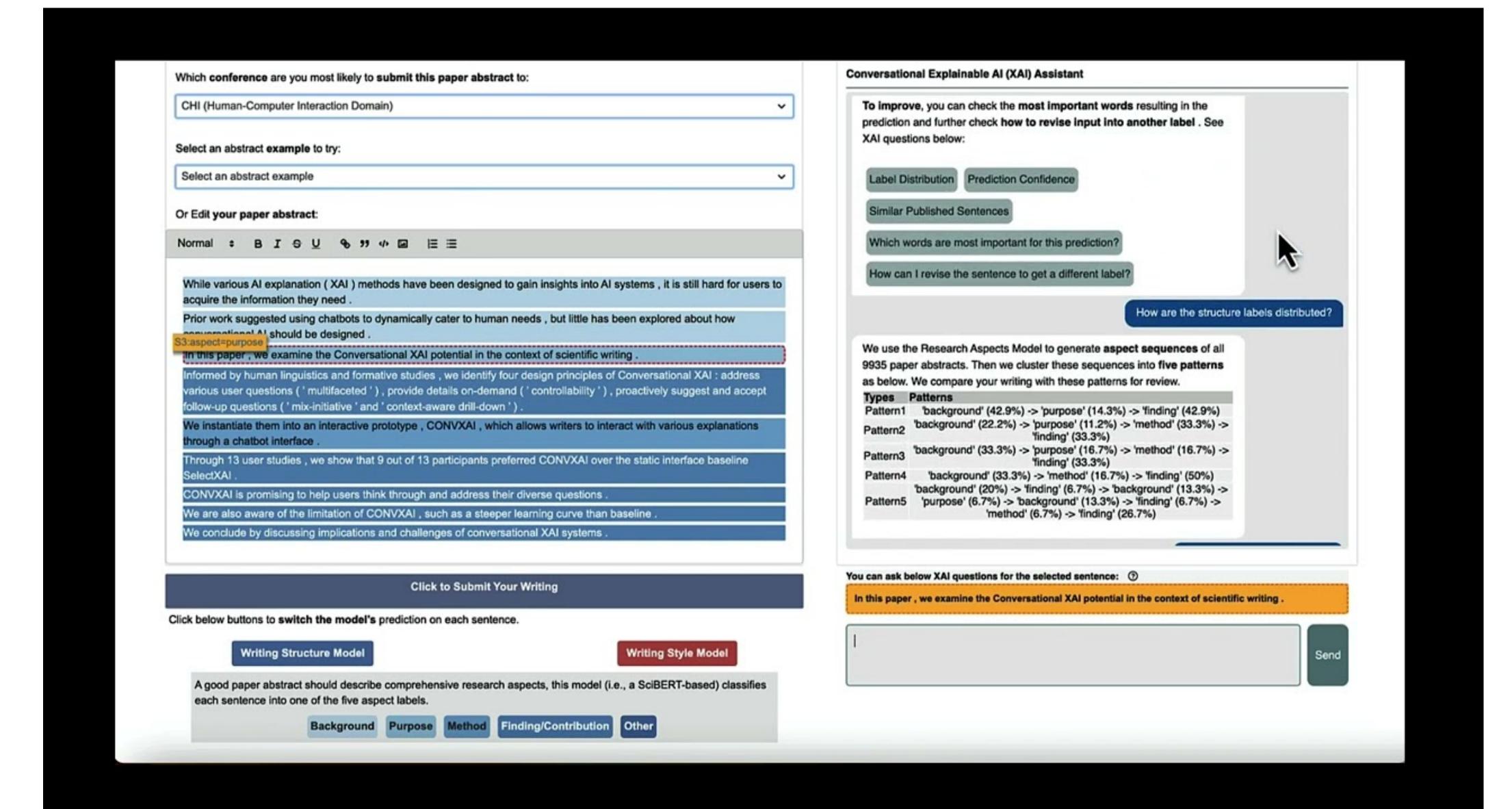
Maintain the conversation history to generate responses with user needs

P4

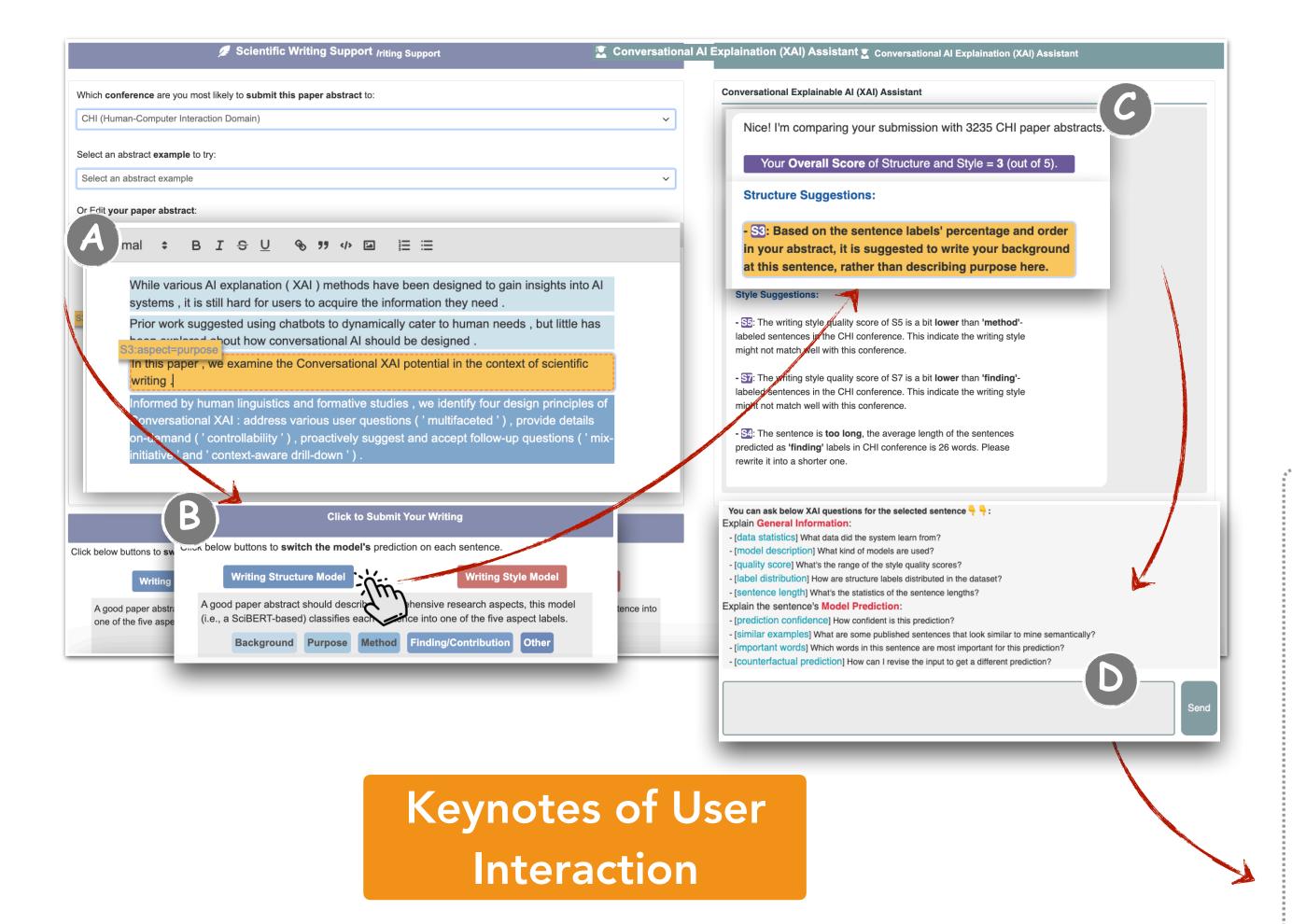
Controllability

Enable humans to customize XAI with personalized needs

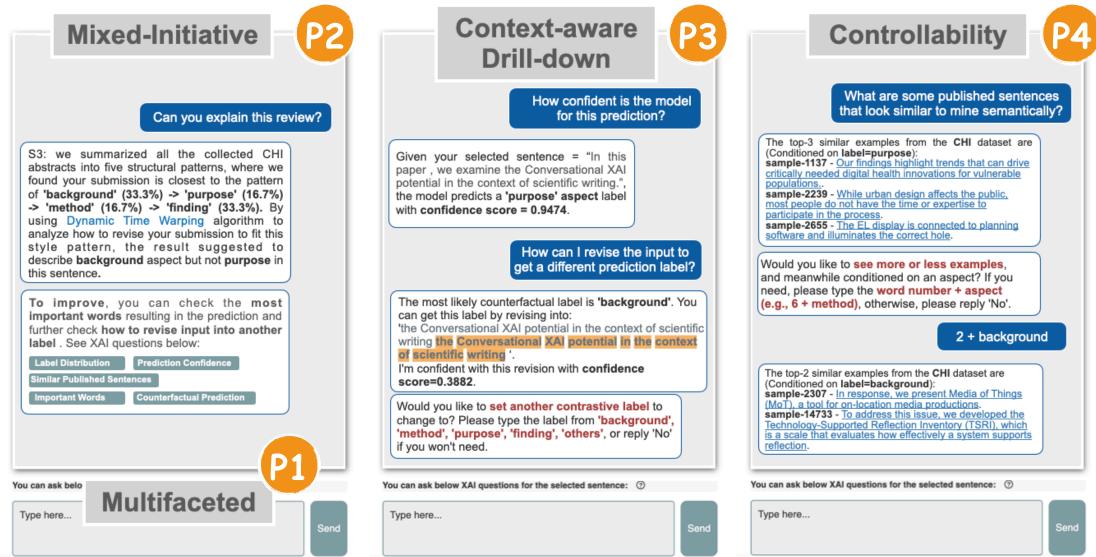
ConvXAI Demo:







Implementing Four Design Principles



54

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

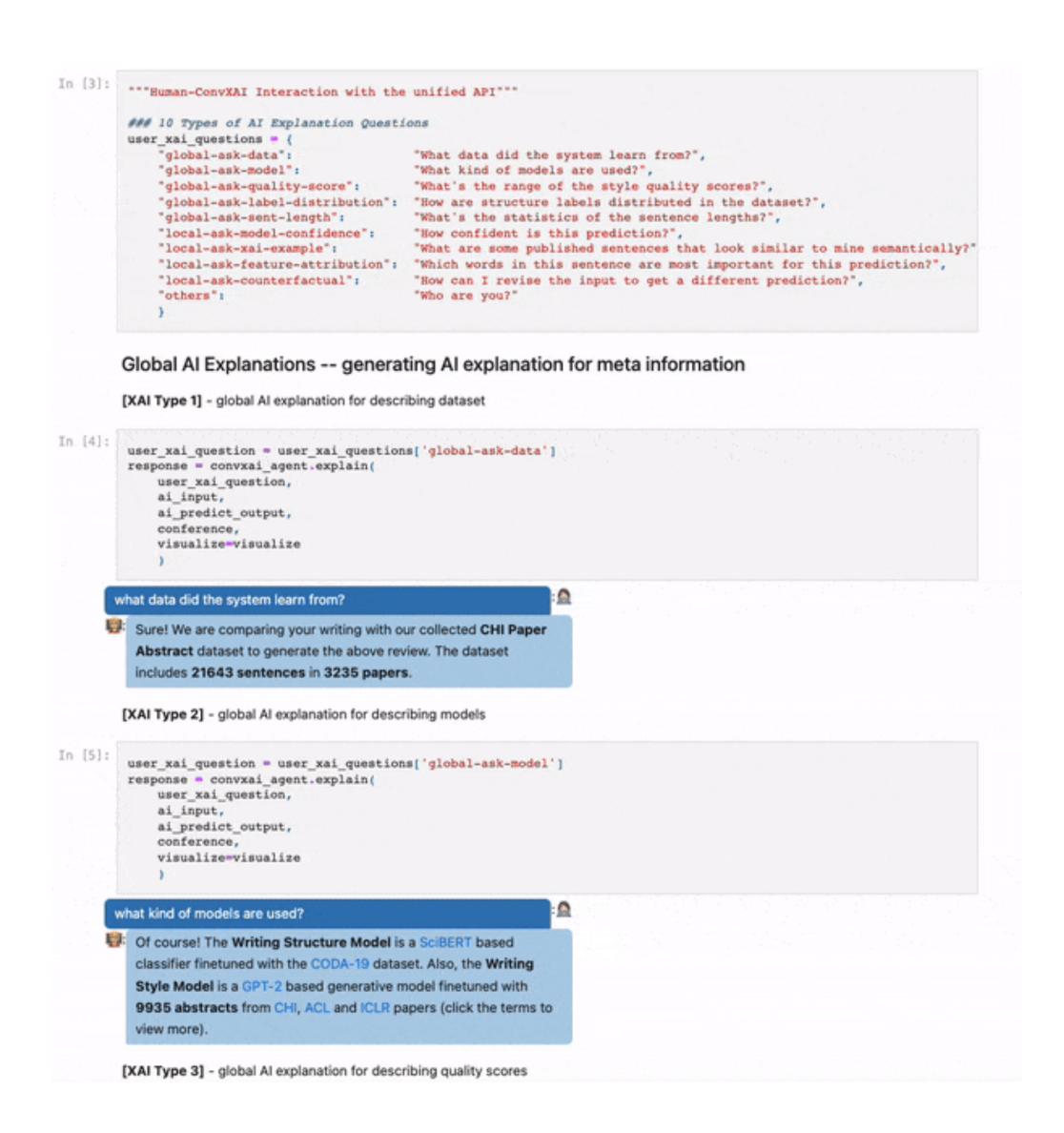
Technical Challenges & Contributions

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.

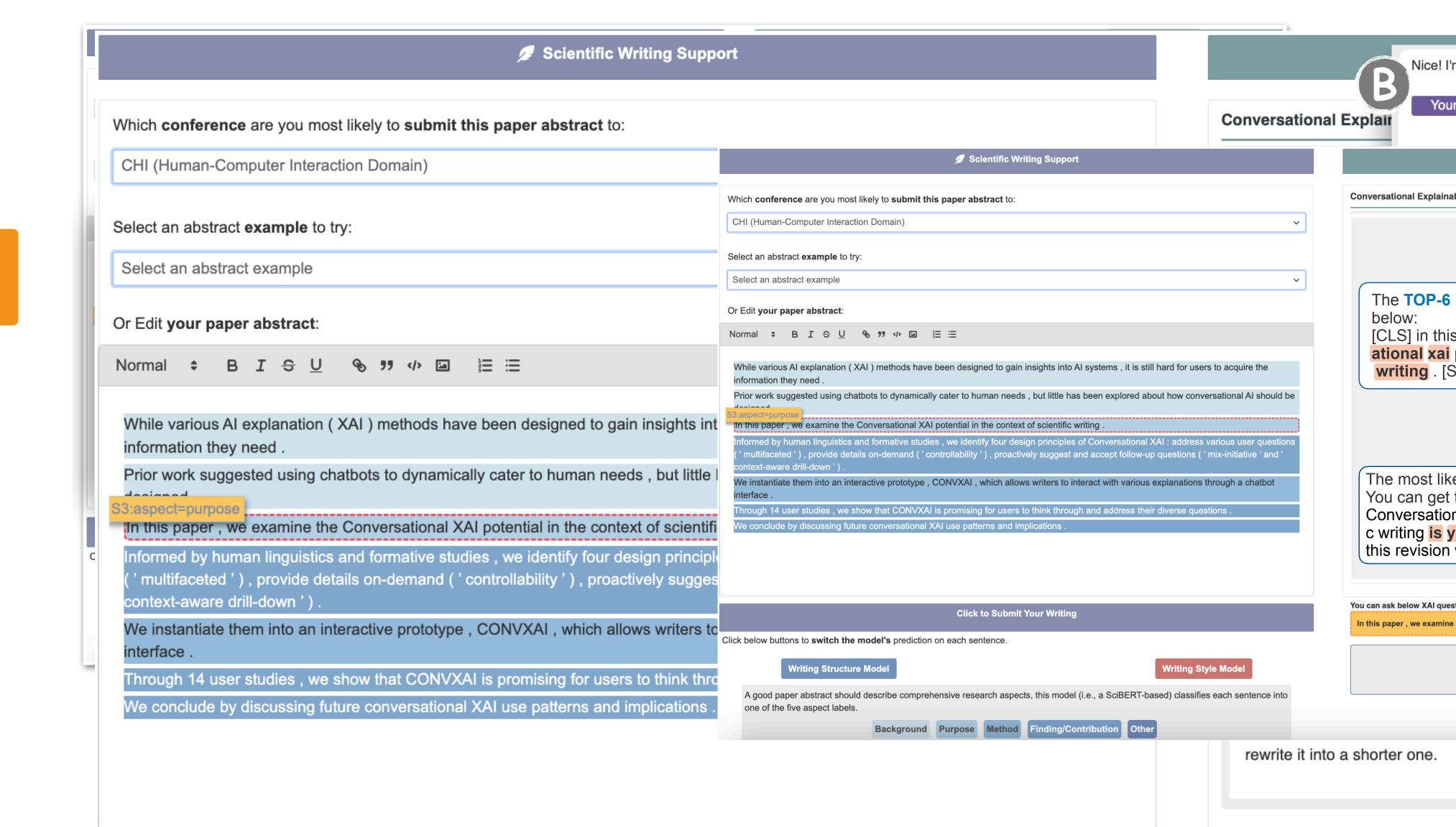


Evaluate ConvXAI with real human studies



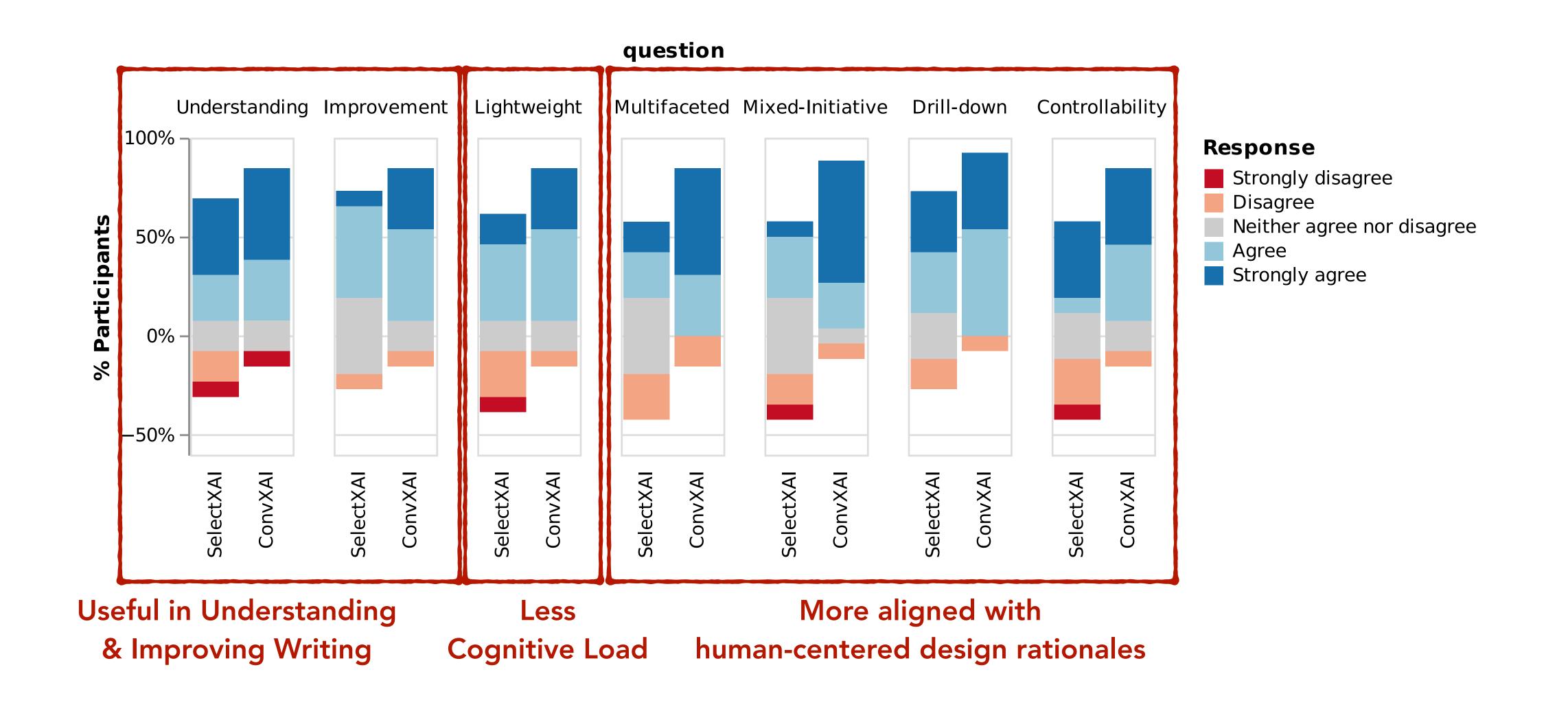
Baseline System (SelectXAI)

Within-Subjects
Study Design



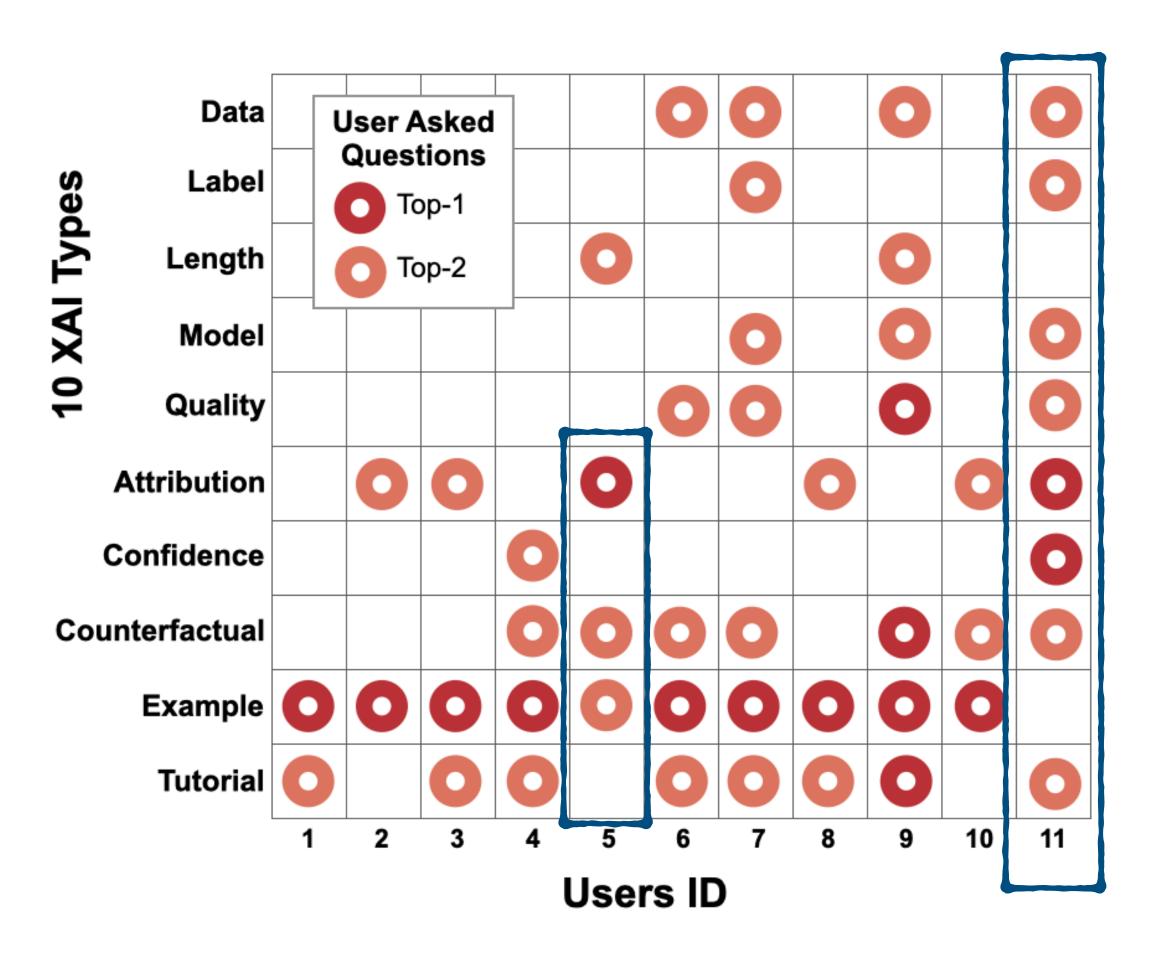
Survey results of human study in Task1

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



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

Table 1. Survey results of human-perceived usefulness rating.

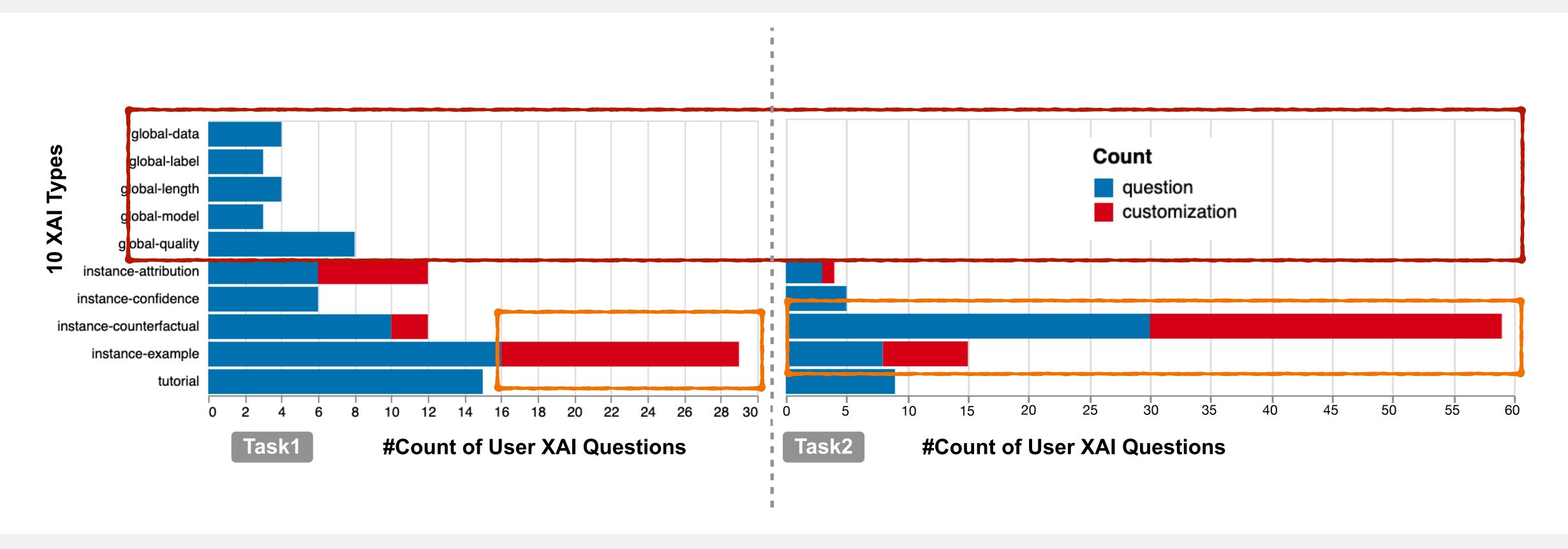
Condition↑	Grammar	ly (1-100)	Model Quality (1-5)		Model Structure (1-5)		Human Qu	ality (1-10)	Human Structure (1-10)	
Condition	Original	Improved	Original	Improved	Original	Improved	Original	Improved	Original	Improved
SelectXAI	04.0 (140.4)	85.1 (±5.52)	2.82 (0.75)	3.05 (0.64)	4 40 (0 27)	4.75 (0.38)	6.5 (1.60)	6.50 (1.30)	6.5 (1.07)	6.63 (1.19)
ConvXAI	84.8 (±10.4)	86.6 (±6.50)	2.82 (0.75)	3.18 (0.71)	4.19 (0.37)	4.31 (0.46)	6.5 (1.69)	6.38 (0.93)	6.5 (1.07)	6.63 (1.19)
Р	-	0.6264	-	0.6965	-	0.0560	-	0.8281	-	1.00

Table 2. 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



Real Humans **Application-grounded Evaluation**

(e.g., Collaboration usefulness, Performance gain)

Real Tasks CSCW 2023 Demo

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





Al Interpretability Evaluation





RQ1: Are XAI Useful for Humans?

RQ2: Why?

RQ3: How to Improve?

Limitation

• In real world human-Al tasks, "how to quantify human's subjective goal of XAI usefulness, and align it with objective Al predictions" is still challenging.

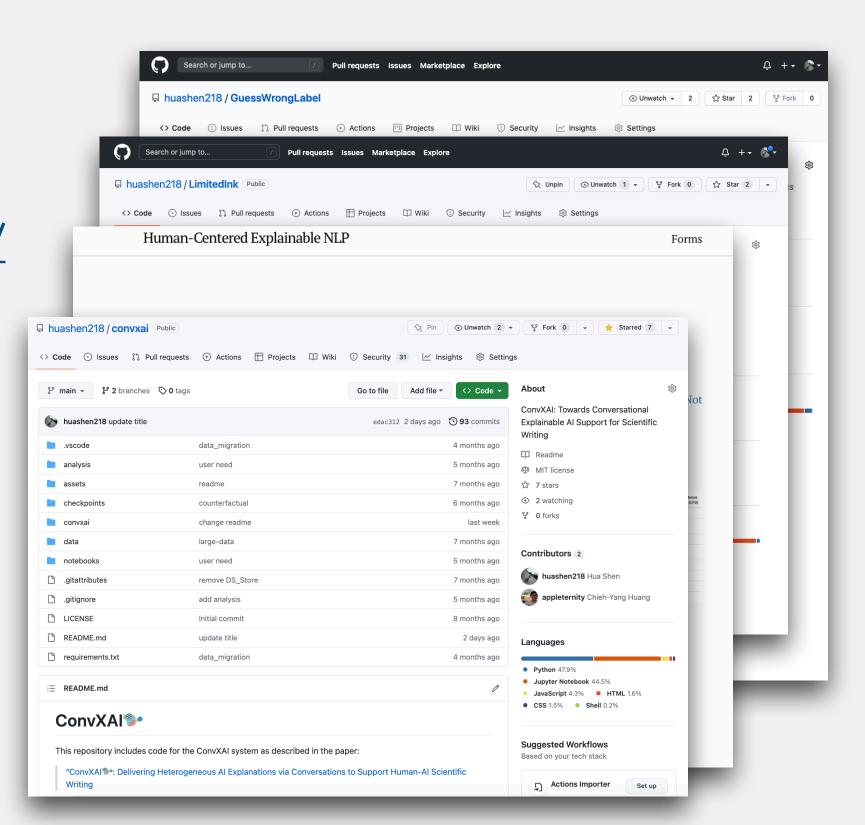
Future Work

- 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/ <u>GuessWrongLabel</u>
- 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-Al 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 Al 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 Al 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 Groupadapted Fusion Network. ICASSP, 2022. (Amazon Al 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.



















































































