Exploring NLP Models and Datasets through Dialogue-based Explanations

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, Sebastian Möller

German Research Center for Artificial Intelligence (DFKI)

Technische Universität Berlin

Saarland Informatics Campus

Why do we need dialogue-based explanation?

- Interactive conversational interface providing multi-turn dialogues and context
- One-off explanations not sufficient, potentially ambiguous
- Support various explanations in one single system
- Ease of use; More accessible to laypeople

Hello 😊, I'm a machine learning model trained to predict to answer offensive/non-offensive attitude based on tweets.

Let's get started. Ask me something!

Can you show me three most attributed tokens for id 16?

Text: @USER I'm astounded that conservatives hold her up as some kind of intellectual. What a failure.

Top 3 token(s): **. i conservative**

▾ The visualization:

Feedback

Please generate an adversarial sample for this instance.

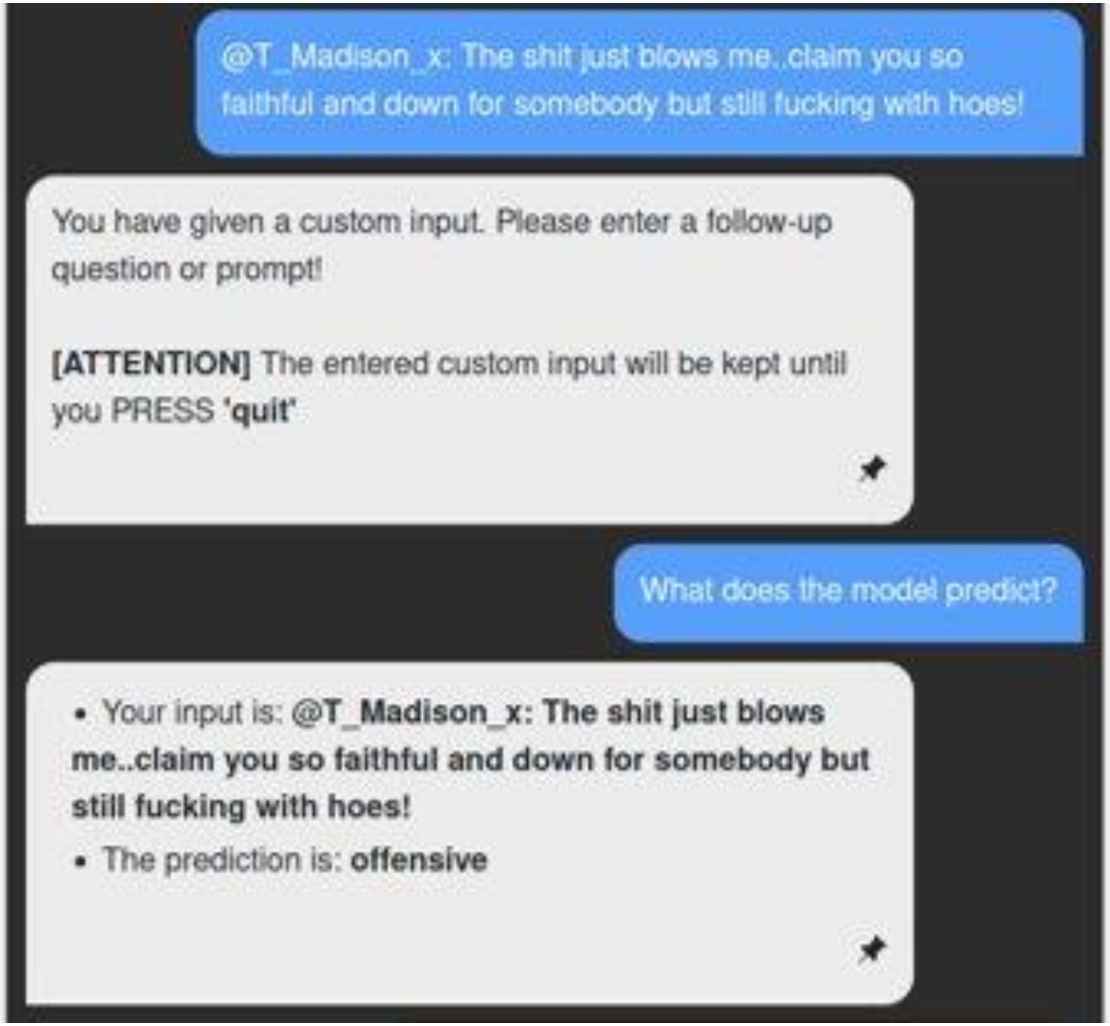
Label non-offensive (85.065%) --> offensive (57.769%)

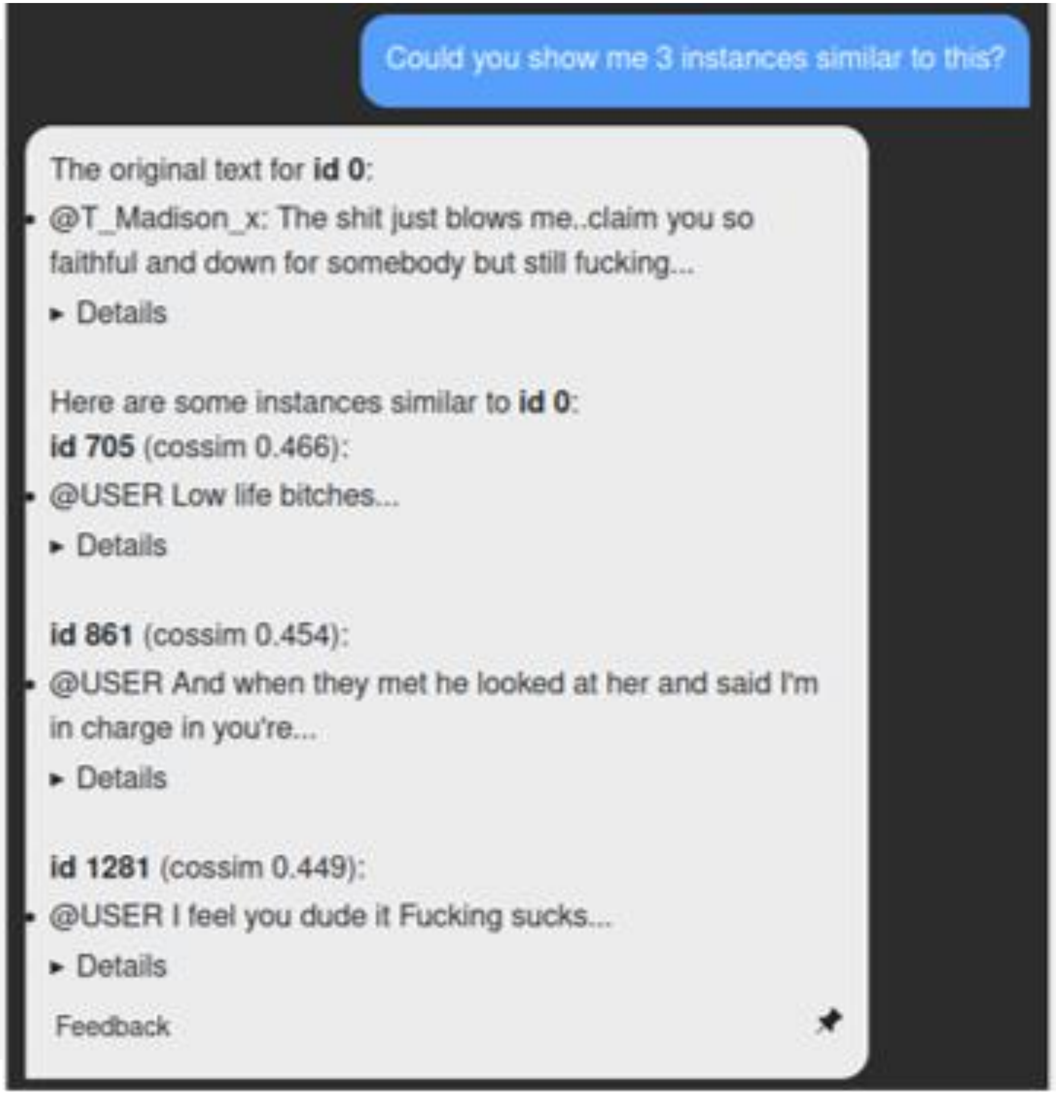
@ **USER** I ' m astounded that conservatives hold her up as some kind of intellectual . What a **failure** .

@ **user** i ' m astounded that conservatives hold her up as some kind of intellectual . what a **loser** .

Feedback

Support for custom inputs | Similar examples





Rationale Generation

TalkToModel

Interactive Dialogues with ML Models

HUGGING FACE

+

NLP Model Token Attributions

Natural Language Counterfactuals

Rationale Generation with LLMs

Semantic Similarity

=

Task: Dialogue Act Classification

…

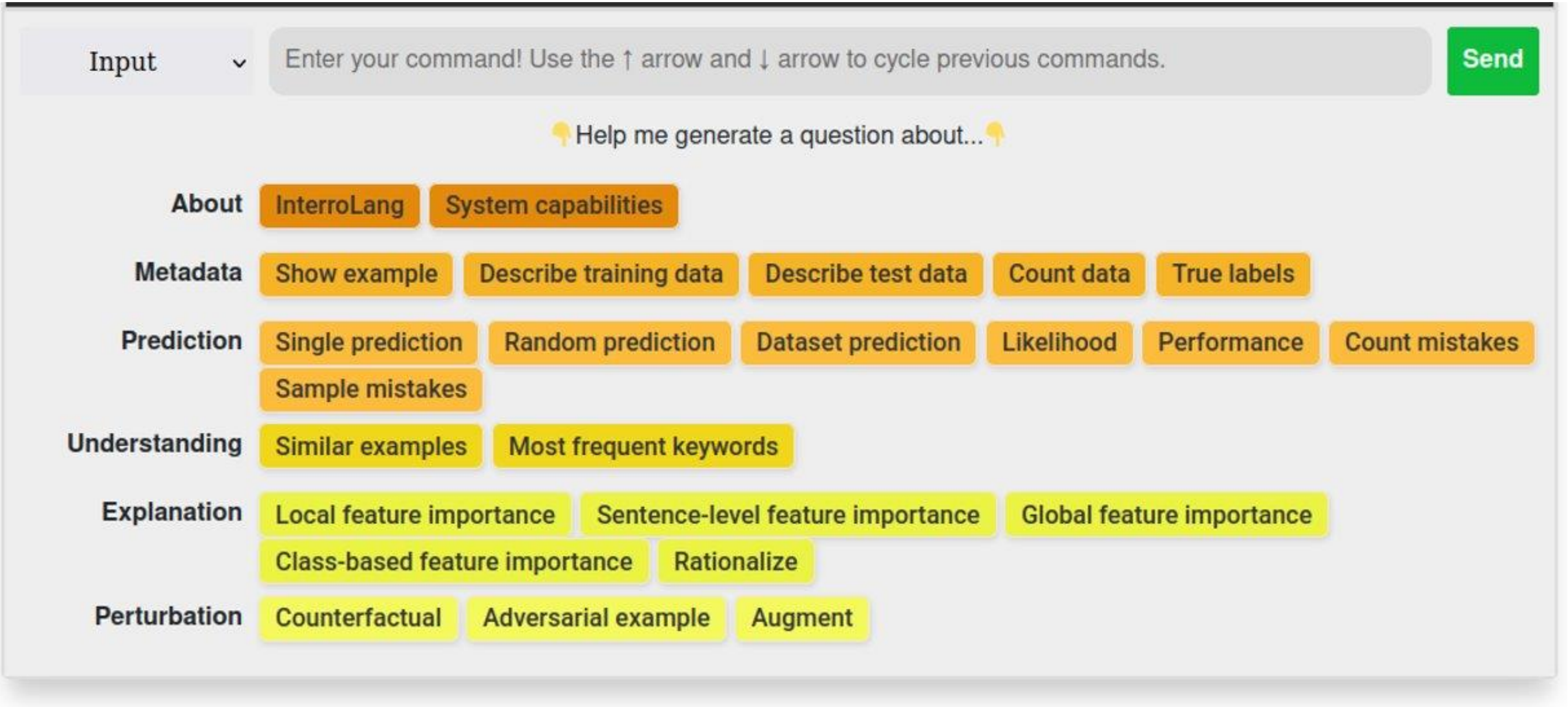
Task: Question Answering

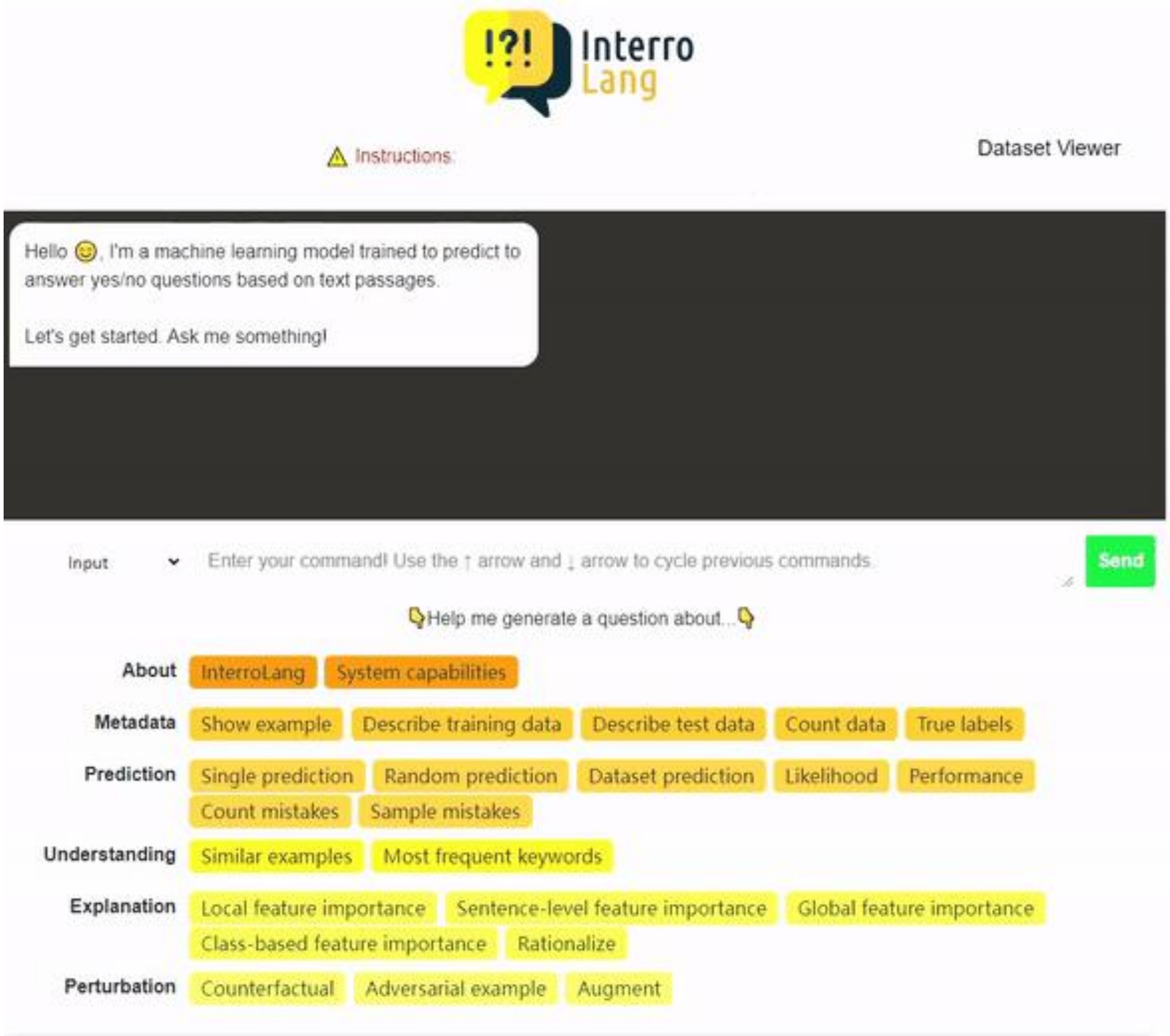
Task: Hate Speech Detection

Tweet: "*blasey ford is a fat ugly libral snowflake*"
Explain in natural language,
Why is this text hateful?

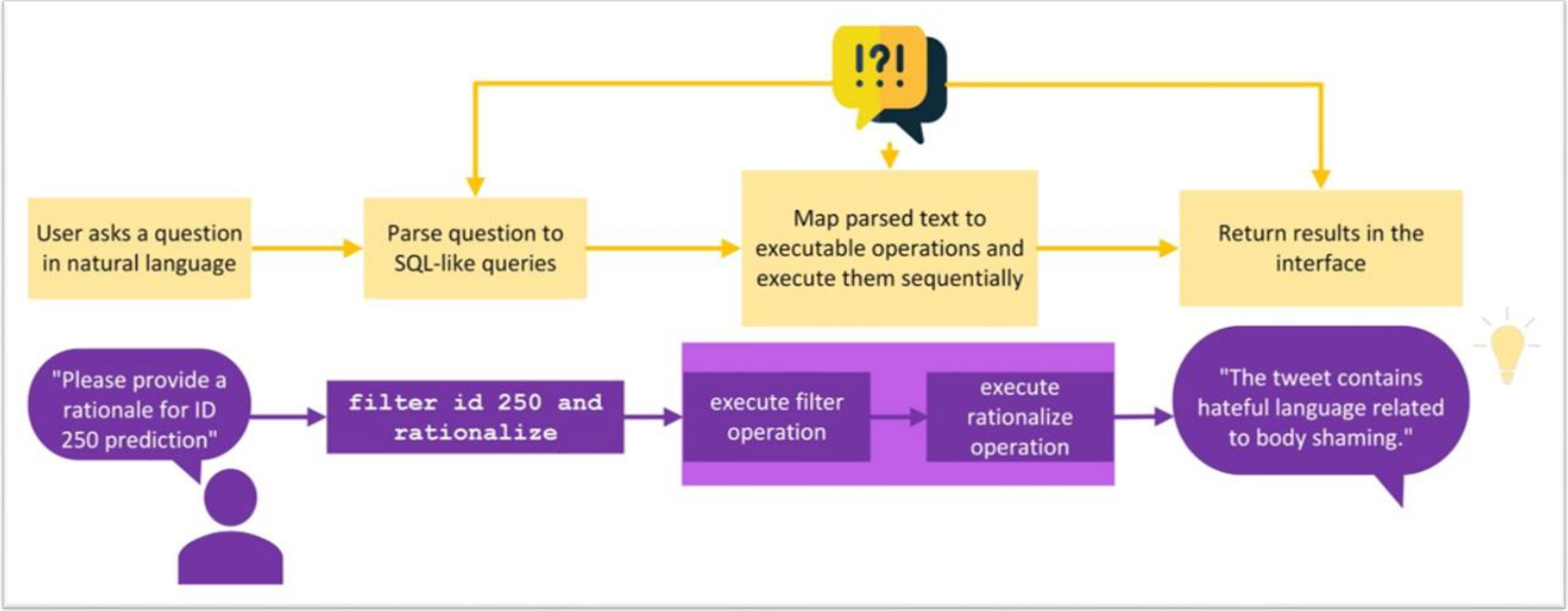
The tweet includes insults related to body shaming.

Operations


Input
Enter your command! Use the ↑ arrow and ↓ arrow to cycle previous commands.
Send
Help me generate a question about...
About
InterroLang
System capabilities
Metadata
Show example
Describe training data
Describe test data
Count data
True labels
Prediction
Single prediction
Random prediction
Dataset prediction
Likelihood
Performance
Count mistakes
Sample mistakes
Understanding
Similar examples
Most frequent keywords
Explanation
Local feature importance
Sentence-level feature importance
Global feature importance
Class-based feature importance
Rationalize
Perturbation
Counterfactual
Adversarial example
Augment

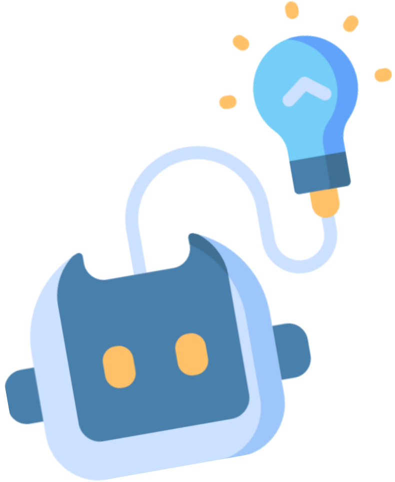

Interro Lang
Instructions
Dataset Viewer
Hello, I'm a machine learning model trained to predict to answer yes/no questions based on text passages.
Let's get started. Ask me something!
Input
Enter your command! Use the ↑ arrow and ↓ arrow to cycle previous commands.
Send
Help me generate a question about...
About
InterroLang
System capabilities
Metadata
Show example
Describe training data
Describe test data
Count data
True labels
Prediction
Single prediction
Random prediction
Dataset prediction
Likelihood
Performance
Count mistakes
Sample mistakes
Understanding
Similar examples
Most frequent keywords
Explanation
Local feature importance
Sentence-level feature importance
Global feature importance
Class-based feature importance
Rationalize
Perturbation
Counterfactual
Adversarial example
Augment

How to recognize user intent



Building blocks of **InterroLang**

Operation	**Tool / Model**
Intent recognition / Parsing	GPT-Neo (2.7B) FLAN-T5-base (250M) BERT + Adapter (110M)
Feature Attribution / Saliency Method	Captum Integrated Gradients
Counterfactuals	Polyjuice (GPT-2)
Adversarial Examples	OpenAttack
Data Augmentation	NLPAug
Rationalization	Dolly v2 (3B)
Similar Examples	SBERT



Experiments: Intent recognition

Dataset		BoolQ			OLID			DailyDialog		
Parsing model	Size	*dev*	*dev-gpt*	*test*	*dev*	*dev-gpt*	*test*	*dev*	*dev-gpt*	*test*
Nearest Neighbors	-	34.69	35	34.02	33.67	35	30.26	36.73	37	32.51
`GPT-Neo`	2.7B	**73**	70	72.54	71	72	67.11	70	66	70.44
`FLAN-T5-base`	250M	71	71	74.18	63	66	66.67	66	63	75.86
`BERT+Adapter`	110M	72.55	**76.86**	**79.33**	**72.55**	**76.86**	**84.25**	**72.55**	**77.69**	**83.94**

Table 2: Exact match parsing accuracy (in %) for the datasets and their three partitions (human-authored *dev* development data, *dev-gpt* data augmented via `GPT-3.5`, *test* set created from questions asked by participants of the user study). `GPT-Neo` uses $k = 20$ shots in the prompt.

Human evaluation: Subjective ratings

	Operations	Corr.	Help.	Sat.
Metadata	Show example	52.94	44.44	42.19
	Describe data	89.66	87.27	87.72
	Count data	56.41	44.44	45.83
	True labels	58.82	64.71	72.22
	Model cards	56.25	43.75	45.06
Prediction	Random prediction	57.59	60.71	65.52
	Single/Dataset prediction	53.42	53.52	54.17
	Likelihood	62.86	67.50	63.41
	Performance	72.50	65.79	76.19
	Mistakes	81.25	68.75	77.09
NLU	Similar examples	53.57	45.61	62.50
	Keywords	60.34	54.00	60.00
Expl.	Feature importance	55.88	42.25	50.00
	Global feature importance	50.00	50.00	31.32
	Free-text rationale	62.07	62.50	65.45
Pertb.	Counterfactual	40.00	27.03	21.62
	Adversarial example	61.90	40.00	37.50
	Augmentation	62.50	52.17	60.00

Subjective ratings (% positive) on **c**orrectness, **h**elpfulness and **s**atisfaction for single turns, macro-averaged.

Human evaluation: Simulatability

- User is exposed to: Input + Explanation
- User has to predict the expected model outcome
- Simulation accuracy: How often user prediction == Actual model outcome

Explanation types	Sim (all)	Sim ($t = 1$)	Help Ratio	#Turns Avg.
Local feature importance	91.43	93.10	**82.86**	3.85
Sent. feature importance	90.00	94.44	60.00	3.84
Free-text rationale	**94.74**	**100.00**	68.42	**3.70**
Counterfactual	85.00	80.00	25.00	4.14
Adversarial example	84.00	85.71	56.00	4.00
Similar examples	88.46	87.50	61.54	4.00

Table 5: Task B of the user study: Simulatability. Simulation accuracy (in %), simulation accuracy for explanations deemed helpful (in %), helpfulness ratio (in %), average number of turns needed to make a decision.

Takeaways for dialogue-based explanations

BERT + Adapter solution works best for intent recognition

- Smaller models outperform LLMs!

Human evaluators preferred **global explanations and analyses**

1. Metadata (Model cards / Datasheets)
2. Common mistakes made by the model
3. Performance metrics (Accuracy, F1, etc.)

Simulatability shows **multi-turn explanations** are necessary. Most useful explanation types:

1. Feature attribution
2. Free-text rationales

Outlook

- How well can InterroLang generalize to other tasks, modalities and larger LMs?
- How can we use the user feedback (ratings and responses) to improve the model?

InterroLang: Exploring NLP Models and Datasets through Dialogue-based Explanations

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, Sebastian Möller

Now on arXiv!

Accepted to *EMNLP 2023 Findings*.
To be presented at *BlackboxNLP* in Singapore on December 7!