

# INTERROLANG: Exploring NLP Models and Datasets through Dialogue-based Explanations

Nils Feldhus<sup>1</sup> Qianli Wang<sup>2,1</sup> Tatiana Anikina<sup>1,3</sup>  
Sahil Chopra<sup>3,1</sup> Cennet Oguz<sup>1,3</sup> Sebastian Möller<sup>2,1</sup>

<sup>1</sup>German Research Center for Artificial Intelligence (DFKI)

<sup>2</sup>Technische Universität Berlin, Germany

<sup>3</sup>Saarland Informatics Campus, Saarbrücken, Germany

{firstname.lastname}@dfki.de

## Abstract

While recently developed NLP explainability methods let us open the black box in various ways (Madsen et al., 2022), a missing ingredient in this endeavor is an interactive tool offering a conversational interface. Such a dialogue system can help users explore datasets and models with explanations in a contextualized manner, e.g. via clarification or follow-up questions, and through a natural language interface. We adapt the conversational explanation framework TALKTOMODEL (Slack et al., 2022) to the NLP domain, add new NLP-specific operations such as free-text rationalization, and illustrate its generalizability on three NLP tasks (dialogue act classification, question answering, hate speech detection). To recognize user queries for explanations, we evaluate fine-tuned and few-shot prompting models and implement a novel Adapter-based approach. We then conduct two user studies on (1) the perceived correctness and helpfulness of the dialogues, and (2) the simulatability, i.e. how objectively helpful dialogical explanations are for humans in figuring out the model’s predicted label when it’s not shown. We found rationalization and feature attribution were helpful in explaining the model behavior. Moreover, users could more reliably predict the model outcome based on an explanation dialogue rather than one-off explanations.

*Disclaimer: This paper contains material that is offensive or hateful.*

## 1 Introduction

Framing explanation processes as a dialogue between the human and the model has been motivated in many recent works from the areas of HCI and ML explainability (Miller, 2019; Lakkaraju et al., 2022; Weld and Bansal, 2019; Jacovi et al., 2023). With the growing popularity of large language models (LLMs), the research community has yet to present a dialogue-based interpretability

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 167

**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:

@ user i ' m as ##tou  
##nded that conservative ##s hold  
her up as some kind of  
intellectual . what a failure .

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

Figure 1: INTERROLANG dialogue with token-level attribution and adversarial example operations. Users are aware of IDs in the data, since we provide a dataset viewer (not shown).

framework in the NLP domain that is both capable of conveying faithful explanations<sup>1</sup> in human-understandable terms and is generalizable to different datasets, use cases and models.

One-off explanations can only tell a part of the

<sup>1</sup>While it might be tempting to use ChatGPT, we point out the black-box nature of proprietary software: Most interpretability methods require access to gradients, parameters or training data to make faithful explanations of their behavior. Lastly, it is not possible yet to connect other ML models to it for generating explanations.

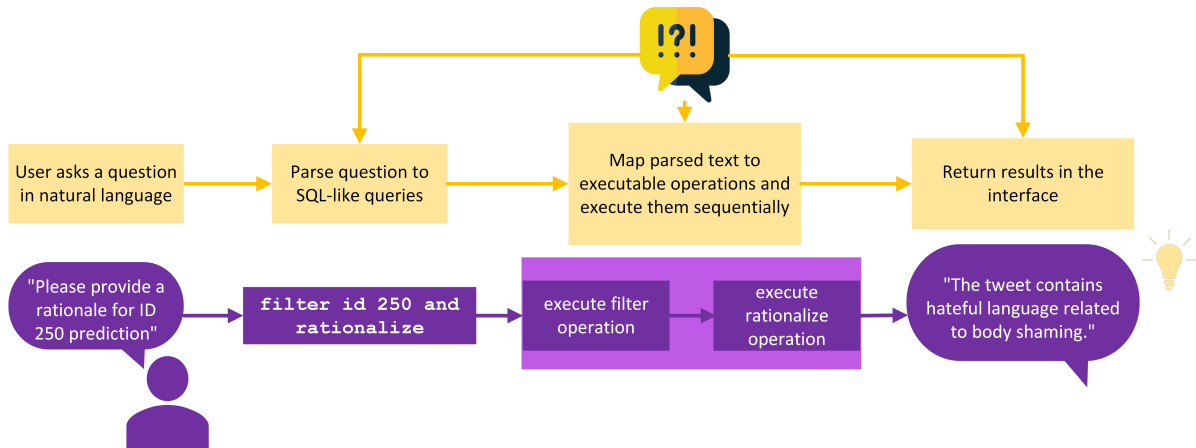


Figure 2: Illustration of how natural language queries from users are parsed into executable operations and their results are inserted in INTERROLANG responses presented through a dialogue interface.

overall narrative about why a model “behaves” a certain way. Saliency maps from feature attribution methods can explain the model reasoning in terms of what input features are important for making a prediction (Feldhus et al., 2023), while counterfactuals and adversarial examples show how an input needs to be modified to cause a change in the original prediction (Wu et al., 2021). Semantic similarity and label distributions can shed a light on the data which was used to train the model (Shen et al., 2023), while rationales provide a natural language justification for a predicted label (Wiegrefe et al., 2022). These methods do not allow follow-up questions to clarify ambiguous cases, e.g. a most important token being a punctuation (Figure 1) (cf. Schuff et al. 2022), or build a mental model of the explained models.

In this work, we build a user-centered, dialogue-based explanation and exploration framework, INTERROLANG, for interpretability and analyses of NLP models. We investigate how the TALKToMODEL (TTM, Slack et al. 2022) framework can be implemented in the NLP domain: Concretely, we define NLP-specific operations based on the aforementioned explanation types. Our system, INTERROLANG, allows users to interpret and analyze the behavior of language models interactively. We demonstrate the generalizability of INTERROLANG on three case studies – dialogue act classification, question answering, hate speech detection – for which we evaluate the intent recognition (parsing of natural language queries) capabilities of both fine-tuned (FLAN-T5, BERT with Adapter) and few-shot LLM (GPT-Neo). We find that an efficient Adapter setup outperforms few-shot LLMs, but that this

task of detecting a user’s intent is far from being solved. In a subsequent human evaluation (§5.3), we first collect subjective quality assessments on each response about the explanation types regarding four dimensions (correctness, helpfulness, satisfaction, fluency). We find a preference for mistakes summaries, performance metrics and free-text rationales. Secondly, we ask the participants about their impressions of the overall explanation dialogues. All of them were deemed helpful, although some (e.g., counterfactuals) have some potential for improvement. Finally, a second user study on simulatability (human forward prediction) provides first evidence for how various NLP explanation types can be meaningfully combined in dialogical settings. Attribution and rationales resulted in very high simulation accuracies and required the least number of turns on average, revealing a need for a longer conversation than single-turn explanations. We open-source our tool<sup>2</sup> that can be extended to other models and NLP tasks alongside a dataset collected during the user studies including various operations and manual annotations for the user inputs (parsed texts): Free-text rationales and template-based responses for the decisions of NLP models include explanations generated from interpretability methods, such as attributions, counterfactuals, and similar examples.

## 2 Methodology

TALKToMODEL (Slack et al., 2022) is designed as a system for open-ended natural language dialogues for comprehending the behavior of ML models

<sup>2</sup><https://github.com/DFKI-NLP/InterroLang>

OLID example instance:	<i>ibelieveblaseyford is liar she is fat ugly libreal snowflake she sold her herself to get some cash !! From dems and Iran ! Why she spoke after JohnKerryIranMeeting ?</i>
<b>Operation</b>	<b>Description; Question + Explanation example</b>
<b>Attribution</b>	<code>nlppattribute</code> (instance, granularity)* <b>Desc:</b> Feature importances on instance at (token   sentence)-level <b>Q:</b> Which tokens are most important? <b>E:</b> <i>fat, ugly</i> and <i>liar</i> are most important for the hate speech label.
	<code>globaltopk</code> (dataset, k, classes) <b>Desc:</b> Top k most attributed tokens across the entire dataset <b>Q:</b> What are the three most important keywords for the hate speech label in the data? <b>E:</b> <i>dumb, fucking,</i> and <i>ugly</i> are the most attributed for the hate speech label.
<b>Perturbation</b>	<code>nlpcfe</code> (instance, number) <b>Desc:</b> Gets number natural language counterfactual explanations for a single instance <b>Q:</b> How do you flip the prediction? <b>E:</b> By replacing <i>liar, fat, ugly</i> with neutral nouns and adjectives.
	<code>adversarial</code> (instance) <b>Desc:</b> Gets number adversarial examples for a single instance <b>Q:</b> What is the minimal change needed to cause a wrong prediction? <b>E:</b> <i>I question the timing of Dr. Ford's statement following the #JohnKerryIranMeeting [...]</i>
	<code>augment</code> (instance) <b>Desc:</b> Generate similar instance <b>Q:</b> Can you generate one more example like this? <b>E:</b> <i>I'm skeptical of her integrity and perceive her as a figure manipulated by political agendas.</i>
<b>Rat.</b>	<code>rationalize</code> (instance) <b>Desc:</b> Explain an instance (prediction) in natural language (rationale generation) <b>Q:</b> In natural language, why is this text hateful? <b>E:</b> The text includes multiple instances of insults related to body shaming.
<b>NLU</b>	<code>keywords</code> (dataset, number) <b>Desc:</b> Show most frequent keywords in the dataset <b>Q:</b> What are the most frequent keywords in the dataset? <b>E:</b> <i>USA, president, democrats</i>
	<code>similar</code> (instance, number)* <b>Desc:</b> Gets number of training data instances most similar to the current one <b>Q:</b> What is an instance in the data very similar to this one? <b>E:</b> <i>@USER How is she hiding her ugly personality. She is the worst.</i>

Table 1: Set of INTERROLANG operations. Descriptions and exemplary question-explanation pairs are added for the hate speech detection use case (OLID). Operations marked with (\*) provide support for custom input instances received from users. This applies to single instance prediction as well (Table 8).

for tabular datasets (including only numeric and categorical features). Our system INTERROLANG retains most of its functionalities: Users can ask questions about many different aspects and slices of the data alongside predictions and explanations. INTERROLANG has three main components (depicted in Figure 2): A *dialogue engine* parses user inputs into an SQL-like programming language using either Adapters for intent classification or LLM that treats this task as a seq2seq problem, where user inputs are the source and the parses are the targets. An *execution engine* runs the operations in each parse and generates the natural language response. A *text interface* (Figure 4) lets users engage in open-ended dialogues and offers pre-defined questions that can be edited. This reduces the users’ workload to deciding on what to ask, essentially.

## 2.1 Operations

We extend the set of operations in TTM (App. B), e.g. feature attribution and counterfactuals, towards linguistic questions, s.t. they can be used in NLP settings and on Transformers. In Table 1, we cat-

egorize all INTERROLANG operations into Attribution, Perturbation, Rationalization, and Data.

**Attribution** Feature attribution methods can quantify the importance of input tokens (Madsen et al., 2022) by taking the final predictions and intermediate representations of the explained model into account. Next to simple token-level attributions, we can aggregate them on sentence-level or present global top  $k$  attributed tokens across the entire dataset (Rönnqvist et al., 2022).

**Perturbation** Perturbation methods come in many forms and have different purposes: We propose to include counterfactual generation, adversarial attacks and data augmentation as the main representatives for this category. While counterfactuals aim to edit an input text to cause a change in the model’s prediction (Wu et al., 2021), adversarial attacks are about fooling the model to not guess the correct label (Ebrahimi et al., 2018). Data augmentation replaces spans in the input, keeping the outcome the same (Ross et al., 2022).

**Rationalization** Generating free-text rationales in the form of parsing consecutive operations (Figure 2) and extend it with the ability to handle custom inputs and clarification questions (App. E).

Such natural language explanations are usually generated by either concatenating the input text with the prediction and then prompting a model to explain the prediction, or by jointly predicting and rationalizing. However, the task has not yet been explored within dialogue-based model interpretability tools.

**Similarity** Inspired by influence functions (Koh and Liang, 2017), this functionality returns a number of instances from the training data that are related to the (local) instance in question. Since influence functions are notoriously expensive to compute, as a proxy, we instead compute the semantic similarity to all other instances in the training data, and retrieve the highest ranked instances.

## 2.2 Intent recognition

We follow TTM and write pairs of utterances and SQL-like parses that can be mapped to operations (Table 1) as well as templates that can be filled.

We propose a novel Adapter-based solution for intent recognition and train a model which can classify intents representing the TERRQLANG operations (e.g. adversarial, counterfactual, etc.). We also train a separate Adapter model for the slot tagging, s.t. for each intent we can label the relevant slots. The slot types that can be recognized by the model included, number, class\_names, data\_type, metric, include\_token and sentence\_level. The training details of the Adapter-based approach are listed in Table 9.

The training data for intents are generated from the same prompts that are used for baselines (GPT-Neo and FLAN-T5-base) with the slot values randomly replaced by the actual values from the datasets (e.g., IDs, class names etc.). Some of the prompts are paraphrased to obtain more diverse training data. Adapter models for intents and slots are fine-tuned on top of the same bert-base-uncased model. The performance of this approach is compared to the prompt-based solution in Table 2. We add dialogue management

<sup>3</sup>Some of the slots are crucial for the intent interpretation and cannot be omitted (e.g., for the show operation) while other slots are optional and if not specified by the user the default value is chosen. We also implement additional checks for the case when the user input includes deictic expressions (e.g., “this” in “show me a counterfactual for this sample”) in which case the ID of the previous instance is selected.

## NLP Models

We selected three use cases in NLP with BERT-type Transformer models trained on standard datasets, all of which we offer users to explore.

### 3.1 Dialogue Act classification

DailyDialog (Li et al., 2017) is a multi-turn dialogue dataset that covers different topics related to our daily life (e.g., shopping, discussing vacation trips etc.). All conversations are human-written and there are 13,118 dialogues in total with 8 turns per dialogue on average. We limit the training set to the first 1,000 dialogues, the development set to 100 and the test set to 300 dialogues.

The dialogue act labels annotated in the dataset are as follows: Inform, Question, Directive and Commissive (see Figure 3a for the distribution of labels). Inform is about providing information in the form of statements or questions. Question is used when the speaker wants to know something and actively asks for information. Directives are about requests, instructions, suggestions and acceptance or rejection of offers. Commissives are labeled when the speaker accepts or rejects requests or suggestions (Li et al., 2017). The Transformer model trained on DailyDialog achieves F1 score 68.7% on the test set after 5 epochs of training with 5e-6 learning rate.

### 3.2 Question answering

We choose BoolQ (Clark et al., 2019) as the representative dataset which has been analyzed in the explainability context in many works (DeYoung et al., 2020; Atanasova et al., 2020; Pezeshkpour et al., 2022, i.a.). Each of the 16k examples consists of a question, a paragraph from a Wikipedia article, the title of that article, and a “yes”/“no” answer.

We let its validation set (3.2k instances) be predicted by a fine-tuned DistilBERT (Sanh et al., 2019) model<sup>5</sup> with an accuracy of 72.11%. We choose a smaller model, because it is more easily deployable and more error-prone which increases the need for explanations.

<sup>4</sup>The ground truth labels for the test set are not available.

<sup>5</sup><https://huggingface.co/andig11/distilbert-base-uncased-qa-boolq>

Dataset Parsing model	Size	BoolQ			OLID			DailyDialog		
		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-annotated development data, dev-gpt data augmented via GPT-3.5 test set created from questions asked by participants of the user study) GPT-Neosk = 20 shots in the prompt.

### 3.3 Hate speech detection

Hate speech detection is a challenging task to determine user entries on social media if offensive.

While better models for hate speech detection are continuously being developed, there is little research on the acceptability aspects of hate speech models. There have been a few studies on this task in the explainability literature, mostly using attributions or binary highlights (Mathew et al., 2021; Balkir et al., 2022; Attanasio et al., 2022).

OLID (Zampieri et al., 2019) is one of the common benchmark datasets and includes 14,100 tweets to be identified whether they are offensive. Each row in OLID consists of text and label and the label indicates if the twitter text is “offensive” or “non-offensive”. A fine-tuned bert-olid-en model is used to predict the validation set (2648 instances) and it can achieve an accuracy of 81.42%.

## 4 Interpretability and Analysis Components

For our implementation and experimental setup, we use the following tools and methods to realize the operations in Table 1:

**Attribution** Slack et al. (2022) automatically select “the most faithful feature importance method for users, unless a user specifically requests a certain technique”. We constrain feature importance to Integrated Gradients (Sundararajan et al., 2017) saliency scores that we obtain from CAPTUM (Kokhlikyan et al., 2020), which allows easy replacement with other saliency methods. The attributions are based on subtoken-level as generated by the underlying model, e.g. BERT. In our experiments. We also provide caching functionality to pre-compute and store the scores, thus reducing the

inference time and mitigating expensive reruns on static inputs.

**Perturbation** For counterfactual generation, we use the official Hugging Face implementation of POLYJUICE (Wu et al., 2021). Adversarial examples are generated via OPENATTACK (Zeng et al., 2021)<sup>8</sup>, where we choose EWWS (Ren et al., 2019) as the attacker for our models on a single instance. For data augmentation we use the NLP AUG library<sup>9</sup> and replace some tokens in the text based on their embedding similarity computed with the bert-based-casenn model. The percentage of words that are augmented for each text is set to 0.3. We display the replaced words in bold, so that the user can easily distinguish between the original instance and the augmented one.

**Rationalization** As a baseline, we use the parsing model GPT-Neo on a zero-shot setup to produce free-text explanations based on a concatenation of the input, the classification by the explained BERT-type model (Marasovic et al., 2022) and an instruction asking for an explanation. For an improved version, we produce plausible rationales from ChatGPT<sup>10</sup> and then prompt Dolly-v2-3B<sup>11</sup> for few-shot rationales. The rationales are pre-computed for all datasets.

**Natural language understanding** For computing the semantic similarity, we embed the data point using Sentence Transformers (Reimers and Gurevych, 2019) and compute the cosine similarity to other points (excluding the instance in question) in the respective dataset. In order to retrieve frequent keywords from the whole dataset, we apply

<sup>6</sup><https://huggingface.co/sinhala-nlp/mbert-olid-en>

<sup>7</sup><https://huggingface.co/uw-hai/polyjuice>

<sup>8</sup><https://github.com/thunlp/OpenAttack>

<sup>9</sup><https://github.com/makcedward/nlpaug>

<sup>10</sup><https://platform.openai.com/docs/api-reference/chat>, March 23 version

<sup>11</sup><https://huggingface.co/databricks/dolly-v2-3b>

	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
Perfb.	Counterfactual	40.00	27.03	21.62
	Adversarial example	61.90	40.00	37.50
	Augmentation	62.50	52.17	60.00

Table 3: Task A1 of the user study: Subjective ratings (% positive) on correctness, helpfulness and satisfaction for single turns (responses in isolation), macro-averaged (each user has the same weight, regardless of how many ratings they gave). Custom input operations are averaged with their “regular” counterparts.

the stopwords set defined in NLTK (Bird, 2006) and get a word frequency set. The operation can then return the most frequent keywords, with being defined through the user query.

## 5 Evaluation

We conduct our evaluation based on parsing accuracy and two user studies. After introducing the partitions we used to obtain the parsing (intent recognition) results (§5.2), we describe the setup of our human evaluation related to user experience and simulatability (§5.3).

### 5.1 Datasets

FLAN-T5-based and Adapter-based models are trained on the train set, which contains 505 pairs of user questions and prompts. We automatically extended the set for Adapter by filling in all possible slots with the values from the datasets (Fig. 9). The train set is a combination of manual creation (1) by us and subsequent augmentation using ChatGPT. For evaluation, we created three more partitions (dev, dev-gpt test) to evaluate the parsing accuracy, as presented in Table 2. The dev set has been manually created by us which consists of 102 pairs of user questions and parsed texts. To construct

Datasets	Corr.	Help.	Sat.	Flue.
BoolQ	3.6	3.3	2.5	3.1
OLID	2.9	3.4	3.0	3.1
DailyDialog	3.2	3.5	3.1	2.9

Table 4: Task A2 of the user study: Subjective ratings (Likert scale 1-5 with 1 being worst/disagree and 5 being best/fully agree) on correctness, helpfulness, satisfaction and fluency for entire dialogues.

the dev-gpt set, we leverage ChatGPT to generate semantically similar examples extracted from the dev set. The test set is obtained by collecting questions of participants who participated in the user study (§5.3). Unlike TTM, our NLP datasets don’t have a tabular format. Therefore, we had to adjust the parsing approach to be able to handle text inputs relevant to our NLP tasks.

### 5.2 Automated evaluation: Intent recognition

To answer the question of how well are user questions mapped onto the correct explanations and responses, for all three use cases, we compare the GPT-Neo-2.7B parsing proposed in Slack et al. (2022) with our novel Adapter-based solution (§2.2) and also fine-tune a custom parsing model based on FLAN-T5-based (Chung et al., 2022).

### 5.3 Human evaluation

Dialogue evaluation research has raised awareness of measuring usability and understanding among many other criteria. There exist automated metrics based on NLP models for assessing the quality of dialogues, but their correlation with human judgments needs to be improved on (Mehri et al., 2022; Siro et al., 2022). While TTM is focused on usability metrics (ease, confidence, speed, likelihood to use), we target dialogue and explanation quality metrics.

#### 5.3.1 Subjective ratings

A more precise way are user questionnaires (Kelly et al., 2009). We propose to focus on two types of questionnaires: Evaluating a user’s experience (1) with one type of explanation (e.g. attribution), and (2) explanations in the context of the dialogue, with one type of downstream task (e.g., QA). An average of the second dimension will also provide a quality estimate for the overall system. Concretely, we let 10 students with computational linguistics and computer science back-

ground explore the tool and test out the available operations and then rate the following by giving a positive or negative review (Task A):

1. Correctness (C), helpfulness (H) and satisfaction (S) on the single-turn-level
2. CHS and Fluency (F) on the dataset-level (when nishing the dialogue)

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

### 5.3.2 Simulatability

We also conduct a simulatability evaluation (Task B), i.e. based on seeing only an explanation plus the original model input for a previously unseen instance. If the participant can correctly guess what the model predicted for that particular instance (which can also be a wrong classification) (Kim et al., 2016), the explanation that they saw would be deemed more helpful. We can then express an objective quality estimate of each type of explanation in terms of simulation accuracy, both in isolation and in combination with other explanations.

Each participant (four authors of this paper + two students from Task A) received nine randomly chosen IDs (three from each dataset). The list of operations (Table 5) is randomized for each ID serving as the itinerary. After each response, the participant can decide to either perform the simulation or continue with the next in the list. After deciding on a simulated label, they are tasked to assign one helpfulness rating to each operation: 1 = helpful; -1 = not helpful; 0 = unused. Let  $\mathcal{R}$  be the set of all ratings,  $\mathbb{I} \in \{0, 1\}$  our indicator function. We then calculate our Helpfulness Ratio as follows:

$$\text{Helpfulness Ratio} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \frac{1_{\mathbb{I}}(r)}{|R_j|}.$$

Let  $y_i$  be the model prediction at index  $i$  and  $y_i$  the user's guess on the model prediction, then the simulation accuracy is

$$\text{Sim(all)} = \frac{1}{|\mathcal{R}_j|} \sum_{i=1}^P \frac{1_{y_i}(y_i)}{|R_j|}.$$

Filtering for all cases where the operation was deemed helpful:

$$\text{Sim}(t = 1) = \frac{1}{|\mathcal{R}_j|} \sum_{i=1}^P \frac{1_{y_i}(y_i) \cdot 1_{\mathbb{I}}(r_i)}{1_{\mathbb{I}}(r_i)}.$$

## 6 Results and discussion

**Parsing accuracy** Table 2 shows that our Adapter-based approach (slot tagging and intent recognition) is able to outperform both the GPT-Neobaseline and the fine-tuned FLAN-T5 models, using much less parameters and trained on the automatically augmented prompts with replaced slot values.

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.

**Human preferences** Table 3 reveals that most operations were positively received, but there are large differences between the subjective ratings of operations across all three aspects (CHS). We find that data description, performance and mis-takes operations consistently perform highly, indicating that they're essential to model understanding. Among the repertoire of explanation operations, free-text rationale scores highest on average, followed by augmentation and adversarial examples, while counterfactuals are at the bottom of the list. The POLYJUICE GPT was often not able to come up with a perturbation (flipping the label) at all and we see the largest potential of improvement in the choice for a counterfactual generator. The dialogue evaluation in Table 4 also solidifies the overall positive impressions. While BoolQ scored highest on Correctness, DailyDialog was the most favored in Helpfulness and Satisfaction. Fluency showed no differences, mostly because the generated texts are task-agnostic. Satisfaction was lowest across the three use cases. Although the operations were found to be helpful and correct, the satisfaction still leaves some room for improvements, likely due to high affordances (too much information at once) or low comprehensiveness.

**Simulatability** Based on Table 5, we can observe that the results align with the conclusions drawn from Table 3. Specifically, free-text rationales provide the most assistance to users, while feature importance was a more useful operation for multi-turn simulation, compared to single-turn helpfulness ratings. On the other hand, counterfactual and adversarial examples are found to be least helpful, supporting the findings of Task A. Thus, their results may not consistently satisfy users' expectations. We detected very few cases where one operation was sufficient. Combinations of expla-

nations are essential: While attribution and rationales are needed to let users form their hypotheses about the model's behavior, counterfactuals and adversarial examples can be sanity checks that support analysis and accompanying visualization tools in port or counter them (Hohman et al., 2019). With practice. There are overlaps with the methods they integrate, but none of them offer explanations induced false trust and led the users to predict a different model output.

### 6.1 Dataset with our results

We compile a dataset from (1) our templates, (2) automatically generated explanations, and (3) human feedback on the rationales presented through the interface. The research community can use these to perform further analyses and train more robust and human-aligned models. We collected 1449 dialogue turns from feedback and 188 turns from the simulatability study (Task B).

## 7 Related Work

Dialogue systems for interpretability in ML interfaces and conversational agents for explanations. Most notably CONVXAI (Shen et al., 2023) very recently presented the first dialogue-based interpretability tool in the NLP domain. Their focus, however, is on the single task of LLMs as writing assistants. They also don't offer dataset exploration methods, their system is constrained to a single dataset (CODA-19) and they have not considered free-text rationalization, which we find is one of the most preferred types of operations. Dalvi Mishra et al. (2022) proposed an interactive system to provide faithful explanations using previous interactions as a feedback. Despite being interactive, it does not provide feasibility of generating rationales on multiple queries subsequently. Bertrand et al. (2023) wrote a survey on prior studies on "dialogic XAI", while Fig. 6 of Jacovi et al. (2023) highlights that interactive interrogation is needed to construct complete explanation narratives: Feature attribution and counterfactuals complement each other, the users can build a generalizable mental model

Visual interfaces for interpretability in NLP LIT (Tenney et al., 2020) AZIMUTH (Gauthier-Melançon et al., 2022) FAN (Mosca et al., 2023) and WEBSHAP (Wang and Chau, 2023) offer a broad range of explanations and interactive analyses on both local and global levels. ROBUSTNESS GYM (Goel et al., 2021) SEAL (Rajani et al., 2022),

User studies on NLP interpretability Most influential to our study design are simulatability evaluations (Hase and Bansal, 2020; Nguyen, 2018; González et al., 2021; Arora et al., 2022; Das et al., 2022; Feldhus et al., 2023). In terms of preference ratings, Strout et al. (2019) evaluated how these to perform further analyses and train more extractive rationales (discretized attributions) from different models are rated by human annotators. Helpfulness and satisfaction ratings were used in Schuff et al. (2020) and Ray et al. (2019).

## 8 Conclusion

We introduce our system INTERLANG, which is a user-centered dialogue-based system for exploring the NLP datasets and model behavior. This system enables users to engage in multi-turn dialogues. Based on the findings from our conducted user study, we have determined that one-off explanations alone are usually not sufficient or beneficial. In many cases, users may require multiple explanations to obtain accurate predictions and gain a better understanding of the system's output.

Future work includes making the bot more proactive, so that it can suggest new operations related to the user queries. We also want to investigate the feasibility of using a singular LLM for all tasks (parsing, prediction, explanation generation<sup>12</sup>, response generation) over the modular setup that we currently employ; Redesigning operations as API endpoints and training LLMs to call them autonomously take care of the entire dialogue management at once. Lastly, refining language models (increasing faithfulness or robustness, aligning with user expectations) through dialogues has gained traction (Lee et al., 2023; Madaan et al., 2023). While we are already collecting valuable data, our framework misses an automated feedback loop to iteratively improve the models.

<sup>12</sup>Operations have to be adapted in some cases, e.g., generating matrices for feature attribution (Sarti et al., 2023) and counterfactuals without an external library (Chen et al., 2023).



## Limitations

INTERRLANG does not exhaust all interpretability methods, because understanding and integrating them requires a lot of resources. We see feature interactions and component analysis, i.e. neuron- or layer-based interpretations, as the most promising future work.

INTERRLANG does not allow direct model comparison. The models are constrained to their datasets and the use cases are intended to be explored separately.

Users can enter custom inputs to get predicted and explained, but they can not modify the dataset on-the-fly, e.g., adding generated adversarial examples or augmentations directly to the current dataset and saving the updated version.

We do not offer a solution to mitigate biases or potential harmful effects of language models but INTERRLANG with its range of explanations is intended to point users into directions where the training data or model behavior is counter-intuitive.

## Ethics Statement

We incorporate OLID as one of our datasets, which may contain hateful or offensive words. However, it is important to note that we do not generate any new content that is hateful or offensive. Our usage of the OLID dataset is solely for the purpose of assessing the integration of the hate speech detection task to our system and generating plausible and useful explanations.

## References

- Siddhant Arora, Danish Pruthi, Norman Sadeh, William W. Cohen, Zachary C. Lipton, and Graham Neubig. 2022. [Explain, edit, and understand: Rethinking user study design for evaluating model explanations](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5277–5285.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. [A diagnostic study of explainability techniques for text classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.
- Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022. [Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection](#). In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.
- Esma Balkir, Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. [Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, pages 2672–2686, Seattle, United States. Association for Computational Linguistics.
- Astrid Bertrand, Tiphaine Viard, Raik Belloum, James R. Eagan, and Winston Maxwell. 2023. [On selective, mutable and dialogic XAI: A review of what users say about different types of interactive explanations](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.
- Steven Bird. 2006. [NLTK: The Natural Language Toolkit](#). In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. [e-SNLI: Natural language inference with natural language explanations](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Zeming Chen, Qiyue Gao, Kyle Richardson, Antoine Bosselut, and Ashish Sabharwal. 2023. [DISCO: Distilling phrasal counterfactuals with large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-tuned language models](#). arXiv, abs/2210.11416.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

- Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. [Interactive model cards: A human-centered approach to model documentation](#). In 2022 ACM Conference on Fairness, Accountability, and Transparency FAccT '22, page 427–439, New York, NY, USA. Association for Computing Machinery.
- Bhavana Dalvi Mishra, Oyvind Tafjord, and Peter Clark. 2022. [Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement](#). In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing pages 9465–9480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anubrata Das, Chitrang Gupta, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. 2022. [ProtoTEX: Explaining model decisions with prototype tensors](#). In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) pages 2986–2997, Dublin, Ireland. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics pages 4443–4458, Online. Association for Computational Linguistics.
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. [On the origin of hallucinations in conversational models: Is it the datasets or the models?](#) In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies pages 5271–5285, Seattle, United States. Association for Computational Linguistics.
- Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. [HotFlip: White-box adversarial examples for text classification](#). In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) pages 31–36, Melbourne, Australia. Association for Computational Linguistics.
- Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. 2023. [Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods](#). In Proceedings of the First Workshop on Natural Language Reasoning and Structured Explanations (NLRSE) Toronto, Canada. Association for Computational Linguistics.
- Gabrielle Gauthier-Melançon, Orlando Marquez Ayala, Lindsay Brin, Chris Tyler, Frédéric Branchaud-Charron, Joseph Marinier, Karine Grande, and Di Le. 2022. [Azimuth: Systematic error analysis for text classification](#). In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations pages 298–310, Abu Dhabi, UAE. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#) Commun. ACM 64(12):86–92.
- Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. [Robustness gym: Unifying the NLP evaluation landscape](#). In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations pages 42–55, Online. Association for Computational Linguistics.
- Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. [On the interaction of belief bias and explanations](#). In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 pages 2930–2942, Online. Association for Computational Linguistics.
- Peter Hase and Mohit Bansal. 2020. [Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?](#) In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics pages 5540–5552, Online. Association for Computational Linguistics.
- Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. [Gamut: A design probe to understand how data scientists understand machine learning models](#). In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.
- Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. 2023. [Diagnosing AI explanation methods with folk concepts of behavior](#). In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency FAccT '23, New York, NY, USA. Association for Computing Machinery.
- Diane Kelly, Paul B. Kantor, Emile L. Morse, Jean Scholtz, and Ying Sun. 2009. [Questionnaires for eliciting evaluation data from users of interactive question answering systems](#) Natural Language Engineering 15(1):119–141.
- Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. [Examples are not enough, learn to criticize! criticism for interpretability](#). In Advances in Neural Information Processing Systems volume 29. Curran Associates, Inc.
- Pang Wei Koh and Percy Liang. 2017. [Understanding black-box predictions via influence functions](#). In Proceedings of the 34th International Conference on Machine Learning volume 70 of Proceedings

- of Machine Learning Research, pages 1885–1894. PMLR.
- Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. [Captum: A unified and generic model interpretability library for PyTorch](#). arXiv, abs/2009.07896.
- Michał Kuźba and Przemysław Biecek. 2020. [What would you ask the machine learning model? identification of user needs for model explanations based on human-model conversations](#). In *ICML PKDD 2020 Workshops*, pages 447–459, Cham. Springer International Publishing.
- Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. [Rethinking explainability as a dialogue: A practitioner's perspective](#). HCAI @ NeurIPS 2022
- Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2023. [XMD: An end-to-end framework for interactive explanation-based debugging of NLP models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Toronto, Canada. Association for Computational Linguistics.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. [Chameleon: Plug-and-play compositional reasoning with large language models](#). arXiv, abs/2304.09842.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative refinement with self-feedback](#). arXiv, abs/2303.17651.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. [Post-hoc interpretability for neural NLP: A survey](#). *ACM Comput. Surv.*
- Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2022. [ConvXAI: a system for multimodal interaction with any black-box explainer](#). *Cognitive Computing* 15(2):613–644.
- Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. [Few-shot self-rationalization with natural language prompts](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. [Hatexplain: A benchmark dataset for explainable hate speech detection](#). *Proceedings of the AAAI Conference on Artificial Intelligence* 35(17):14867–14875.
- Shikib Mehri, Jinho Choi, L. F. D'Haro, Jan Deriu, Maxine Eskénazi, Milica Gasic, Kallirroi Georgila, Dilek Z. Hakkani-Tür, Zekang Li, Verena Rieser, Samira Shaikh, David R. Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. [Report from the NSF future directions workshop on automatic evaluation of dialog: Research directions and challenges](#). arXiv, abs/2203.10012.
- Tim Miller. 2019. [Explanation in artificial intelligence: Insights from the social sciences](#). *Artificial Intelligence* 267:1–38.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\* '19)*, page 220–229, New York, NY, USA. Association for Computing Machinery.
- Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, and Georg Groh. 2023. [IFAN: An explainability-focused interaction framework for humans and NLP models](#). arXiv, abs/2303.03124.
- Dong Nguyen. 2018. [Comparing automatic and human evaluation of local explanations for text classification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.
- Van Bach Nguyen, Jörg Schlötterer, and Christin Seifert. 2022. [Explaining machine learning models in natural conversations: Towards a conversational XAI agent](#). arXiv, abs/2209.02552.
- Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron Wallace. 2022. [Combining feature and instance attribution to detect artifacts](#). *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1934–1946, Dublin, Ireland. Association for Computational Linguistics.
- Nazneen Rajani, Weixin Liang, Lingjiao Chen, Margaret Mitchell, and James Zou. 2022. [SEAL: Interactive tool for systematic error analysis and labeling](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 359–370, Abu Dhabi, UAE. Association for Computational Linguistics.

- Arijit Ray, Yi Yao, Rakesh Kumar, Ajay Divakaran, Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Giedrius Burachas. 2019. [Can you explain that? lucid explanations help human-ai collaborative image retrieval](#). In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 7, pages 153–161.
- Nils Reimers and Iryna Gurevych. 2019. [SentenceBERT: Sentence embeddings using Siamese BERT-networks](#). In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Chen. 2019. [Generating natural language adversarial examples through probability weighted word saliency](#). In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.
- Samuel Rönnqvist, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter, and Veronika Laippala. 2022. [Explaining classes through stable word attributions](#). In Findings of the Association for Computational Linguistics: ACL 2022, pages 1063–1074, Dublin, Ireland. Association for Computational Linguistics.
- Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. [Tailor: Generating and perturbing text with semantic controls](#). In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019.
- Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. [Inseq: An interpretability toolkit for sequence generation models](#). In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Toronto, Canada. Association for Computational Linguistics.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). arXiv, abs/2302.04761.
- Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. [F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering](#). In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7076–7095, Online. Association for Computational Linguistics.
- Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao 'Kenneth' Huang. 2023. [ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing](#). In Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing CSCW '23 Demo, New York, NY, USA. Association for Computing Machinery.
- Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. [Understanding user satisfaction with task-oriented dialogue systems](#). In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '22, page 2018–2023, New York, NY, USA. Association for Computing Machinery.
- Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. [TalkToModel: Explaining machine learning models with interactive natural language conversations](#). Trustworthy and Socially Responsible Machine Learning Workshop (TSRML) @ NeurIPS 2022.
- Julia Strout, Ye Zhang, and Raymond Mooney. 2019. [Do human rationales improve machine explanations?](#) In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 56–62, Florence, Italy. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 3319–3328. PMLR.
- Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. [The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models](#). In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 107–118, Online. Association for Computational Linguistics.
- Sttorio Torri. 2021. [Textual eXplanations for intuitive machine learning](#). Master's thesis, Politecnico di Milano, dec.
- Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Sasko, Albert Villanova, Quentin Lhoest, Julien Chaumond, Margaret Mitchell, Alexander M. Rush, Thomas

- Wolf, and Douwe Kiela. 2022. [Evaluate & evaluation on the hub: Better best practices for data and model measurement](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128–136, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zijie J. Wang and Duen Horng Chau. 2023. [Webshap: Towards explaining any machine learning models anywhere](#). In *WWW '23 Companion*, page 262–266, New York, NY, USA. Association for Computing Machinery.
- Daniel S. Weld and Gagan Bansal. 2019. [The challenge of crafting intelligible intelligence](#). *Commun. ACM* 62(6):70–79.
- Christian Werner. 2020. [Explainable ai through rule-based interactive conversation](#). In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference*.
- Sarah Wiegrefe, Jack Hessel, Swabha Swayambdipta, Mark Riedl, and Yejin Choi. 2022. [Reframing human-AI collaboration for generating free-text explanations](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. [Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.
- Yang Xiao, Jinlan Fu, Weizhe Yuan, Vijay Viswanathan, Zhoumianze Liu, Yixin Liu, Graham Neubig, and Pengfei Liu. 2022. [DataLab: A platform for data analysis and intervention](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 182–195, Dublin, Ireland. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [Predicting the type and target of offensive posts in social media](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.
- Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. [OpenAttack: An open-source textual adversarial attack toolkit](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371.
- A Explanatory dialogue systems**
- Table 6 and Table 7 show the range of existing natural language interfaces and conversational agents for explanations.
- B TALK TOMODEL operations**
- Most TTM operations belonging to their ML, Conversation and Description categories can be trivially adapted. Here, we document the changes:
- Due to Transformers being explained instead of the much smaller sklearn models, we applied small changes such as pre-computing predictions (similar to the tricks we used for attributions and rationales).
- Metadata** For metadata, we provide an operation following the basic idea of model cards (Mitchell et al., 2019) which supplies information related to model details, intended use of the model, etc., and analogously, datasheets (Gebu et al., 2021) for training/test data documentation. User questions can target specific aspects of this structured information and the system replies in natural language and/or tabular formats.
- Table 8 shows the rest of the TERRLANG operations not depicted by Table 1.
- C Label distributions of NLP use cases**
- Figure 3 shows the label distributions of DailyDialog, OLID and BoolQ.
- D Adapter training details**
- Table 9 shows the hyperparameters and training time for the Adapter models for dialogue act classification and slot tagging.
- E Dialogue management**
- TTM, after translating user utterances into a grammar of production rules, composes its results in a template-filling manner while ensuring semantic coherence between multiple operations. They further argue that such a response generation approach prevents hallucinations commonly found in neural networks and conversational models (Dziri et al., 2022). However, it makes the dialogue less natural.

Implementations	Task data			Model
	Num	CV	NLP	
DR_ANT (Kuřba and Biecek, 2020)	■			RF
ERIC (Werner, 2020)	■			DT
Torri (2021)	■			RF
TALKTOMODEL (Slack et al., 2022)	■			RF
XAGENT (Nguyen et al., 2022)	■	■		RF, CNN
CONVXAI (Malandri et al., 2022)	■			DT, RF
CONVXAI (Shen et al., 2023)			CODA-19	Tf
<b>INTERROLANG (ours)</b>			BoolQ DailyDialog OLID	Tf

Table 6: Explananda (Task and model) comparison of existing implementations of natural language interfaces and conversational agents for XAI. We can see that applications to NLP tasks have started to surface only recently. **Task data** Num = Numeric/Tabular. CV = Computer vision. **Explained model** AOG = And-Or graph. DT = Decision Tree. RF = Random Forest. CNN = Convolutional neural network. Tf = Transformer.

Implementations	Explanation types					Intent recognition / Parsing of user questions				Resp	DST	Evaluation	
	FA	CF	Mt	Sim	RG	Comm	Embeds	Fine-Tuned	Few-Shot			Auto	Hum
Kuřba and Biecek (2020)	■	■				DiF				DiF	DiF		
Werner (2020)	■	■					fastText			Rule			
Torri (2021)	■	■						GPT-2		Rule			Like
Slack et al. (2022)	■	■	■				MPNet	T5	GPT-Neo/-J	Rule	Rule	ExM	Like
Nguyen et al. (2022)	■	■	■				SimCSE			Rule		ExM, F1	
Malandri et al. (2022)	■	■	■			RASA				Rule	Rule		Like
Shen et al. (2023)	■	■	■	■			SciBERT			Rule	Rule		
<b>INTERROLANG (ours)</b>	■	■	■	■	■		MPNet	BERT+Adap, FLAN-T5	GPT-Neo	Rule	Rule, Adap	ExM	Like

Table 7: Explanans (XAI modules) comparison of existing implementations of natural language interfaces and conversational agents for XAI. **Explanation types** FA = Feature Attribution. CF = Counterfactual Generation. Mt = Meta information about the model. Sim = Similar examples. RG = Rationale generation. **Intent recognition** Comm = Commercial product (RASA = RASA NLU; DiF = Google DialogFlow). Embeds = Nearest neighbor based on sentence embedding. **Response generation / Dialogue state tracking** Rule = Rule- and template-based response. **Evaluation** : **Automated**: ExM = Exact match accuracy. **Human**: Like = Likert-scale rating.

<b>Filters</b>	<code>filter(id)</code> <code>includes(token)</code>	Access single instance by its ID Filter instances by token occurrence
<b>Prediction</b>	<code>predict(instance)*</code> <code>predict(dataset)</code> <code>likelihood(instance)</code> <code>mistakes(dataset)</code> <code>score(dataset, metric)</code>	Get the prediction of the given instance Get the prediction distribution across the dataset Obtain the given instance’s probability for each class Count number of wrongly predicted instances Determine the relation between predictions and labels
<b>Data</b>	<code>show(list)</code> <code>countdata(list)</code> <code>label(dataset)</code>	Showcase a list of instance Count number of instances within the given list Describe the label distribution across the dataset
<b>Meta</b>	<code>data(dataset)</code> <code>model()</code>	Information related to training/test data Metadata of the model
<b>About</b>	<code>function()</code> <code>self()</code>	Inform the functionality of the system Self-introduction
<b>Logic</b>	<code>and(op1, op2)</code> <code>or(op1, op2)</code>	Concatenation of multiple operations Selection of multiple filters

Table 8: TTM operations used in INTERROLANG. \*Prediction operation provides support for custom input instances received from users.

That is why we also add a range of pre-defined responses for fallback that are chosen at random

when applicable. Moreover, the GPT-based rationales are also the first example of a fully model-generated response. Our system also recognizes when the user just wants to acknowledge the bot’s response or intends to finish the conversation and it generates the appropriate responses (see App. G for an example).

When designing dialogue systems, the task of keeping track of the dialogue history is essential to better inform the selection of the next action or response. Thus, we store the previous operations and ids and can resolve deictic expressions like “this sample” or “it” to the ID of the previously mentioned instance. We also check the prediction scores of the intent recognition module to see if there is some problem interpreting the user input, e.g., if several intents get very high scores

Parameters	Dialogue Act Classification	Slot Tagging
Base Model	<i>bert-base-uncased</i>	<i>bert-base-uncased</i>
Learning Rate	1e-4	1e-3
Number of Epochs	10	8
Batch Size	32	32
Optimizer	AdamW	AdamW
Number of Labels	23	15
Avg. Training Time	53 min	32 min
Avg. Model Size	3.6MB	3.6MB
Training Set	39,635	3,810
Development Set	11,010	635

Table 9: Training parameters for the Adapter-based parsing models. The best performing model was selected based on the loss on the development set. All samples are based on the original prompts automatically augmented through the slot value replacements.

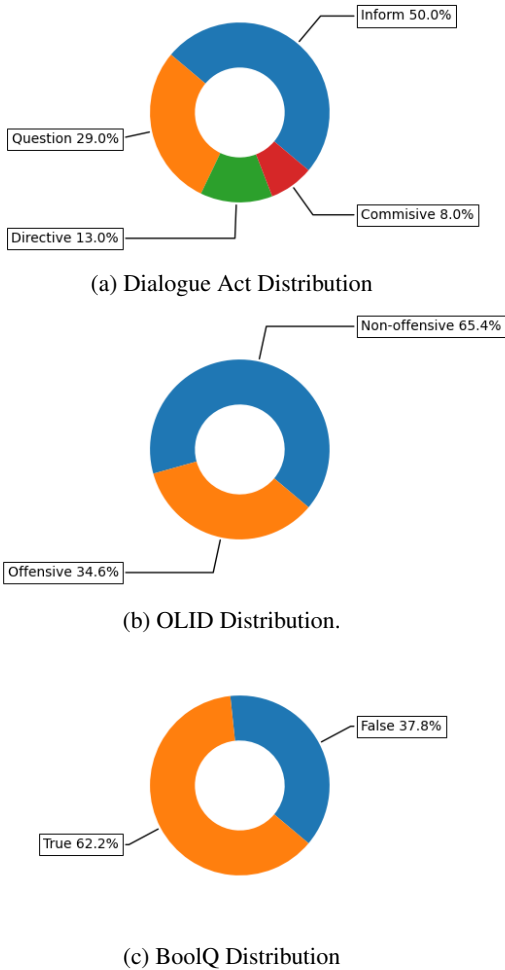


Figure 3: Label distribution of all three datasets.

INTERROLANG asks a clarification question to disambiguate between operations. Also, if we have an intent but some of its non-default slots are missing (not recognized) we can generate a clarification question to resolve it, e.g., “Could you please specify for which instance I should provide a counterfactual?”. This gives us more flexibility and makes the dialogue flow more natural.

## F Interface

We extend the TTM interface (Slack et al., 2022) in the following ways:

- **Custom inputs:** Compared to TTM, which only allows user to use instances from three pre-defined datasets, we provide a selection box that allows individual inputs from the user to be considered.
- **Text search:** A search engine that allows the user to filter the dataset according to strings. If a query is present, subsequent operations will consider the subset where this filter is applicable.
- **Dataset viewer:** This shows the first ten instances of the dataset (their IDs and the contents of the text fields) at the start, but in order to make the navigation through the data easier for the user, it will update according to both string filters and operations like label filters.

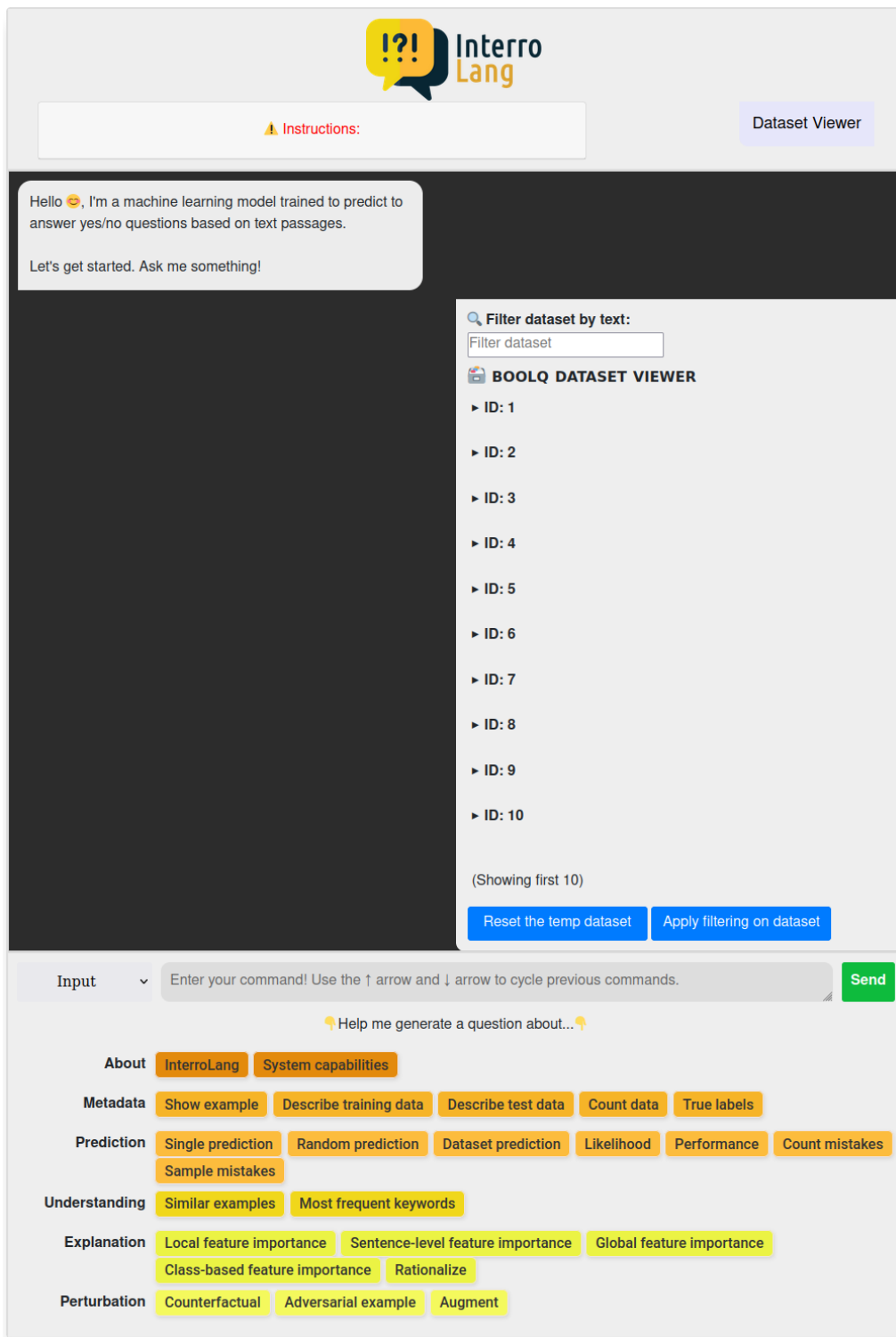


Figure 4: INTERROLANG interface with initial welcome message, opened dataset viewer (BOOLQ) and sample generator buttons.



## G Sample Dialogue (BoolQ with Adapter)

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!

User: Hi! Which kind of a model do you use?

### General Information:

Name	Content
model_name	Distilbert-base-uncased-qa-boolq
model_summary	This model is a distilled version of the BERT base model. It was introduced in this paper. The code for the distillation process can be found here. This model is uncased: it does not make a difference between english and English.
task	Question Answering

User: And what is the dataset?

### Training Data Details:

Name	Content
train_data_name	BoolQ Dataset: a question answering dataset for yes/no questions
train_data_source	plain text
train_data_language	English
train_data_number	9427

The data contains information related to **Boolean question answering (yes/no)**.

The exact **feature names** in the data are listed as follows:

- question
- passage

The dataset has following **labels**:

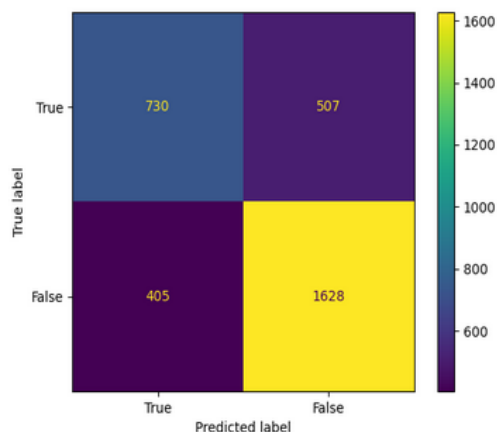
- True
- False

The model scores *72.11% accuracy* on the data.

User: Can you show me how often the model makes incorrect predictions?

For **all** the instances in the data, the model is incorrect 912 out of 3270 times (error rate 0.279).

Here is the confusion matrix:



User: Can you show me some examples of the mistakes?

For **all** the instances in the data, the model is incorrect 912 out of 3270 times (error rate 0.279).

▼ Here are the ids of instances the model predicts incorrectly:

0, 10, 13, 19, 24, 33, 36, 38, 42, 45, 52, 55, 59, 60, 61, 62, 65, 66, 68, 71, 77, 79, 82, 84, 90, 91, 92, 93, 97, 101, 103, 108, 112, 113, 119, 122, 123, 127, 129, 130, 134, 140, 146, 154, 155, 159, 161, 164, 165, 166, 169, 172, 174, 182, 184, 188, 191, 198, 202, 213, 226, 228, 229, 233, 237, 259, 264, 268, 270, 271, 272, 275, 276, 284, 288, 294, 297, 301, 305, 321, 324, 326, 334, 339, 342, 343, 350, 360, 370, 387, 388,

User: Ok, great! What about the id 42? Can you show me this sample?

For the data with **id equal to 42**, the features are

question: is there an interstate that goes coast to coast  
passage: U.S. Route 20 -- U.S. Highway 20 (US 20) is an east-west United States highway that stretches from the Pacific Northwest all the way to New England. The ``0" in its route number indicates that US 20 is a coast-to-coast route. Spanning 3,365 miles (5,415 km), it is the longest road in the United States, and particularly from Idaho to Massachusetts, the route roughly parallels that of Interstate 90 (I-90), which is in turn the longest Interstate Highway in the U.S. There is a discontinuity in the official designation of US 20 through Yellowstone National Park, with unnumbered roads used to traverse the park.

User: Please show me the token attributions for this id.

**Question:** is there an interstate that goes coast to coast  
**Passage:** U.S. Route 20 -- U.S. Highway 20 (US 20) is an east-west United States highway that stretches from the Pacific Northwest all the way to New England. The ``0" in its route number indicates that US 20 is a coast-to-coast route. Spanning 3,365 miles (5,415 km), it is the longest road in the United States, and particularly from Idaho to Massachusetts, the route roughly parallels that of Interstate 90 (I-90), which is in turn the longest Interstate Highway in the U.S. There is a discontinuity in the official designation of US 20 through Yellowstone National Park, with unnumbered roads used to traverse the park.

Top 5 token(s): **disco ##nu that ##nti -**

▼ The visualization:



