Explanation Strategies and Response Generation for Conversational XAI

Anonymous ACL submission

Abstract

In interactive Explainable Artificial Intelligence (XAI), researchers aim to offer explanations of model behavior to non-expert users in a natural, understandable way, e.g., via dialogues. We find that available XAI systems exhibit a lack of understanding the user and responding to them. This is because they do not consider context and often resemble question answering setups. Although computational argumentation and didactics have established interaction patterns for explanatory dialogues, a holistic dialogue management concept is missing. We contribute to conversational XAI in two ways: First, we present a concept for an explanatory dialogue management which is able to take context into account and easily adapt to user needs. Second, we underscore the importance of context by conducting a user study examining Large Language Model (LLM)-generated explanations based on dialogue context. Our study shows that responses based on those explanations outperform conventional template-based answers in terms of likeability. Finally, our ablation studies show that open-source models minimally attend to long contexts and instead rely heavily on the immediate history, but they can compete with `GPT-4` on the task of XAI response generation.

1 Introduction

Human-centered XAI is concerned with incorporating insights from Human-Computer Interaction (HCI) into the field of XAI (Miller, 2019; Ehsan and Riedl, 2020; Weld and Bansal, 2019). Many XAI systems have interactive components, elaborate user interfaces and are evaluated with user studies (Chromik and Butz, 2021; Bertrand et al., 2023). Only recently, however, there has been a push towards conceptualizing dialogue-based XAI systems. Lakkaraju et al. (2022) proposed four modules which are necessary for explanatory conversational systems: Natural language understanding (NLU), explanation algorithm, response generation, and a graphical user interface.

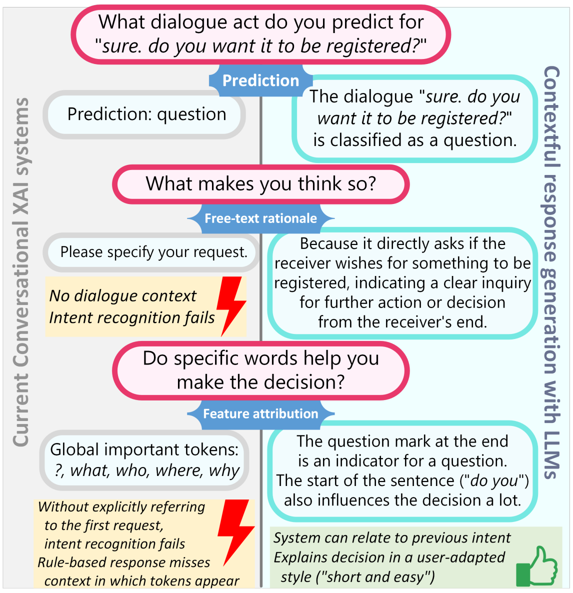


Figure 1: Comparison between explanation dialogue paradigms including requests for prediction, free-text rationales, and feature attribution. Current conversational XAI systems (l.) often fail to map general user requests for explanations to a specific explanation-generating method. Our approach for dialogue management and response generation (r.) uses the dialogue context and user preferences to generate an appropriate explanation.

Representative systems like TalkToModel, (Slack et al., 2023), ConvXAI (Shen et al., 2023), InterroLang (Feldhus et al., 2023b), and LLMCheckup (Wang et al., 2024a) all implement these four modules. However, these systems operate in a manner that resembles question answering, often missing out on the value of dialogue context. We argue that the crucial omission in their concept is a dedicated dialogue management capable of managing explanatory context and user needs (Figure 1).

In this work, we first challenge current implementations of human-centered XAI and point out gaps in conversational XAI systems. We compare

NLU approaches (§2), revealing that current designs are unable to handle generic and potentially ambiguous requests. We then consolidate explanation moves from argumentation and didactics literature, which we show are necessary ingredients for dialogue management in conversational XAI systems (§3).

Secondly, we address the topic of response generation with an empirical study. For this, we use LLMs to generate different styles of system responses and evaluate their appeal to users within a specific dialogue context against template-based responses (§4). We qualitatively analyze the responses of `GPT-4` vs. `DeepSeek-R1-Distill-Qwen` and interpret the LLMs' dialogue context usage (§5).

2 Understanding the User

As user explanation requests are diverse, there is a need for robust and generalizable NLU models in dialogue-based XAI systems. In the following, we will highlight the most recent research in this area and give future directions.

Categorization of User Queries: One of the primary challenges is the broad spectrum of user queries as there are various ways to phrase explainability questions (Wang et al., 2024b). Liao et al. (2021) and Kuźba and Biecek (2020), followed by refinements by Nguyen et al. (2023) and Malandri et al. (2023), attempted to categorize these queries, identifying common question types such as "why" and "what if". Nonetheless, there is no consensus on a universal categorization framework.

Limitations of Current Systems: Many current systems (Slack et al., 2023; Shen et al., 2023; Feldhus et al., 2023b; Wang et al., 2024a) directly map each user utterance to a set of predefined methods, which often disregards the state of the dialogue (history, context, user knowledge, already provided explanations, etc.) and does not allow room for miscommunication. However, dialogues are dynamic and complex, e.g., the same user query may yield different intents depending on the contextual nuances of the conversation. Figure 1 illustrates failure cases for intent recognition: The question "What makes you think so?" is not recognizable as any intent without taking the context into account. Also, the scope of the request for feature attribution is misinterpreted in this example because without the previous turns it is not clear for which sample the most attributed tokens should be displayed.

		Comp. Argumentation			Discourse & Didactics	
	Explanation moves	BW16	Stp24	Mad19	WA22	Hen16
User	Request explanation	7	■	2	E04	I4
	Provide prior knowledge	–	–	–	E02	C1
	Provide question context	–	–	3.4	E09	B2
	Request further details	3	■	–	E06	I2
	Request clarification	10	■	5.2	E06	I6
	Signal understanding	9	■	3.2	E05	B1/P6
	Answer quiz question	–	–	–	E03	R4
	Provide feedback	–	■	–	E07	G4
System	Assess prior knowledge	–	–	–	E02	C1
	Request clarification	–	–	5.1	E06	I6
	Provide explanation	1	■	3.1	E03	R2
	Provide further details	2	■	–	E09	R2/B1
	Provide clarification		■	–	E07	B2
	Acknowledge understanding	–	■	3.3	E05	P3
	Test understanding	–	–	–	E01	I1/R1
	Request feedback	–	–	–	–	C4
	Suggest next steps	–	–	–	–	G2

Table 1: Explanation moves consolidated from computational argumentation (Bex and Walton, 2016; Stepin et al., 2024; Madumal et al., 2019) and discourse and didactics literature (Wachsmuth and Alshomary, 2022; Hennessy et al., 2016). Each non-empty cell signifies that the paper of that column defines this move as part of their explanation dialogue concept (identifier of the move in the corresponding paper or, if not given, ■).

The model simply outputs the highest attributed tokens in the whole dataset. Adapting the NLU to accommodate this variability necessitates the development of more flexible and context-aware intent recognition models capable of capturing and contextualizing diverse dialogue domains.

3 Managing Explanations

Existing conversational XAI systems (cp. §1) mostly support just the basic question-answer (or *request* and *provide explanation*) dialogue structure, since their responses are entirely template and rule-based, because language models have not been showing convincing and consistent response generation quality until recently (Yavuz et al., 2019; Fang et al., 2023). They lack a dedicated dialogue management, as traits of information-seeking (Stepin et al., 2024), mixed-initiative (or proactive) dialogues (Deng et al., 2023), argumentation dialogues (Bex and Walton, 2016) and teacher-student (or tutorial) dialogues (Wachsmuth and Alshomary, 2022; Lee et al., 2023; Liu et al., 2024b) are necessary for a natural explanatory dialogue.

Current research in computational argumentation (Bex and Walton, 2016; Madumal et al., 2019) already considers explanatory dialogue interactions, but remains relatively abstract and does not cover the full range of moves. Similarly, while didactics literature (Wachsmuth and Alshomary, 2022; Hennessy et al., 2016) defines many explanation moves, it lacks a comprehensive dialogue

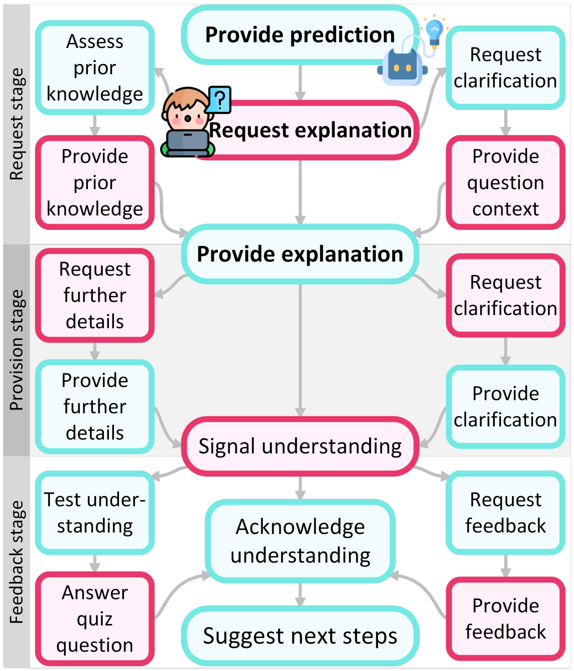


Figure 2: Explanation dialogue strategy based on explanation moves in Table 1. The chart represents how the user (red) would ask the system (blue) about a single item (usually a sub-dialogue in practice), where the model initially provides a prediction (*Request*) and there can be different pathways via the system providing an explanation and optionally additional clarification or knowledge questions (*Provision*). At the end, the user can be prompted with quiz questions to test their understanding and provide feedback to improve the system (*Feedback*), before finally suggesting to move to another item. Current conversational XAI systems mostly support the top three moves (bold).

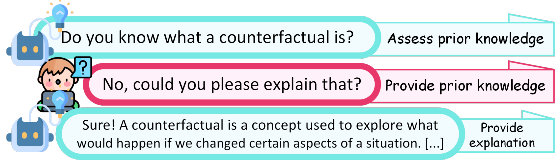
strategy. Based on the definitions in Table 1, we have developed an **explanation strategy** that encompasses a broader spectrum of interaction options within the dialogue, including requesting clarifications or further details, providing context or feedback, as illustrated in Figure 2.

In the course of an explanation dialogue, users might ask about multiple aspects of the XAI domain (training data, custom inputs, model documentation, global explanations, etc.) (Feldhus et al., 2023b), but our strategy refers to a single explanation request. Our strategy, which is the foundation for an explanation dialogue management, incorporates 17 distinct moves in total (cp. Figure 2) connected to three stages: *Request*, *Provision*, and *Feedback*. In the following, we outline the interaction patterns of each stage alongside examples in the XAI system setting and approaches on how to support them from a practical standpoint:

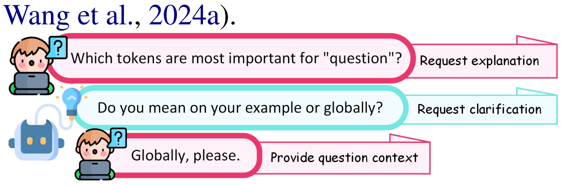
Request stage

Dialogue state: Previous prediction, or other explanation dialogue.
System roles: Verify user knowledge for explanation; clarify context; provide explanation.
User role: Request new explanation.

- *Assessing prior knowledge* can be triggered before the system provides an explanation type which the user has not encountered before, s.t. the explanation can be adapted:



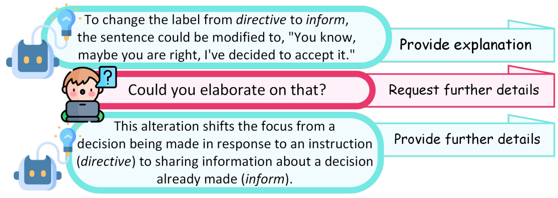
- *Requesting clarification* from the user will ensure the correct intent is recognized and all necessary arguments or parameters are filled to execute the explanation generation (Malandri et al., 2023; Wang et al., 2024a).



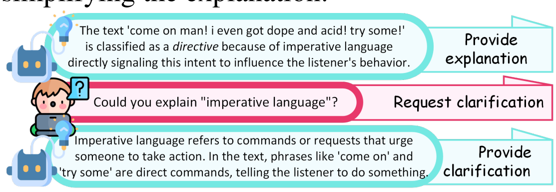
Provision stage

Dialogue state: Explanation context is set.
System role: Provide explanation.
User roles: Clarify misunderstandings or missing knowledge; Request details about the explanation.

- *Providing further details* might entail employing rationalization on the immediate dialogue context, or presenting related explanations to enhance understanding (Fig. 3).

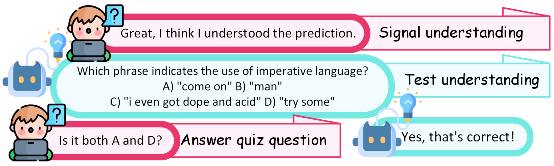


- *Providing clarification* can be achieved by explaining task-specific terminology (e.g., explaining what a specific feature represents), providing a meta-explanation (Wang et al., 2024a) (e.g., explaining how the model makes a prediction), or simplifying the explanation:



Feedback stage
Dialogue state: End of explanation.
System role: Manage user understanding.
User role: Show understanding; provide feedback.

- When the user signals their understanding of the provided explanation, the user could be prompted with a *quiz question* related to it:



- *Requesting feedback* is the system proactively asking the user to communicate their qualitative assessment of the model prediction or system response. User-*provided feedback* can be used as a signal for model refinement (Li et al., 2022).

4 Conversational XAI design

Current XAI systems often rely on simple question-and-answer dialogues, which can limit user understanding and engagement. LLMs offer new potential to enhance these interactions by generating more context-aware explanations. Prior work by Wagner and Ultes (2024) demonstrated that a dialogue controller can improve the overall dialogue experience. Building on this, we aim to investigate the impact of LLM-generated explanations on user perception. Specifically, we analyze how explanations that incorporate dialogue context affect user satisfaction. Since explanations can take various forms, we compare different types of generated texts. Our goal is to assess their effectiveness by evaluating user preferences and likeability. Figure 4 introduces the main workflow.

4.1 Baseline: InterroLang

InterroLang (Feldhus et al., 2023b) provides a dialogue interface to enable queries in natural language about XAI features including model behavior and dataset analyses (Figure 3, Appendix A).

Data The data is sourced from all three available studies of Feldhus et al. (2023b), because it is one of the very few sources for conversational XAI data (Mindlin et al., 2024; Feustel et al., 2024):

- BoolQ (Clark et al., 2019) is a question answering dataset, where each example consists of a question, a paragraph from a Wikipedia article, the title of that article, and a "yes"/"no" answer.

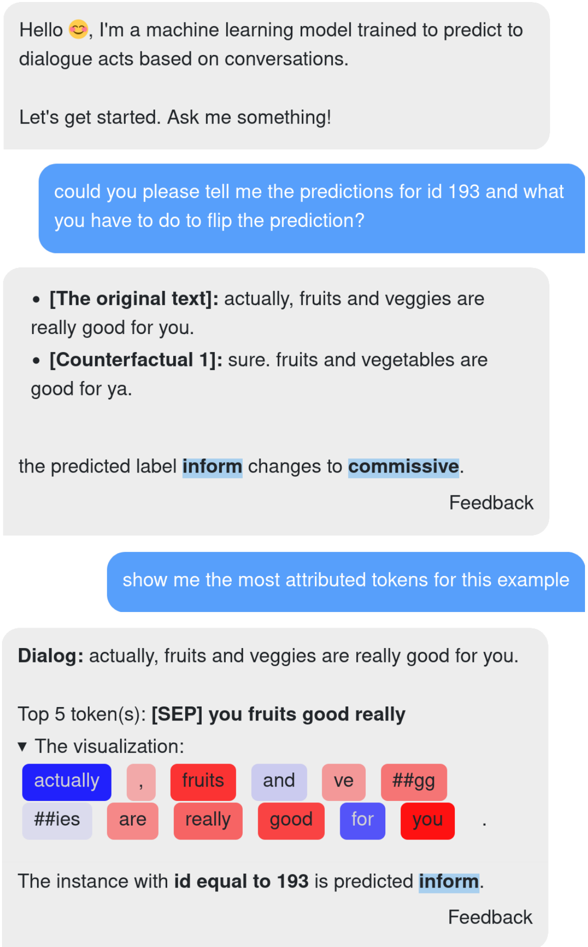


Figure 3: Interface of InterroLang showing an explanation dialogue for the DailyDialog use case.

The validation set was predicted by a fine-tuned `DistilBERT` (Sanh et al., 2019) model[1] with an accuracy of 72.11%;

- OLID (Zampieri et al., 2019), a dataset hate speech detection task to determine if user entries on social media are either offensive and non-offensive. A fine-tuned `mbert-olid-en`[2] model is used to predict the validation set and it can achieve an accuracy of 81.42%;
- DailyDialog (Li et al., 2017), a multi-turn dialogue dataset that represents daily communication. The dialogue act labels annotated in the dataset are as follows: Inform, Question, Directive and Commissive. The Transformer model trained on DailyDialog achieves F1 score 68.7% on the test set.

Explanation types The datasets and pretrained models are then used within the InterroLang system to answer user queries in the dialogue format. InterroLang supports the following explanation

[1] https://hf.co/andi611/distilbert-base-uncased-qa-boolq
[2] https://hf.co/sinhala-nlp/mbert-olid-en

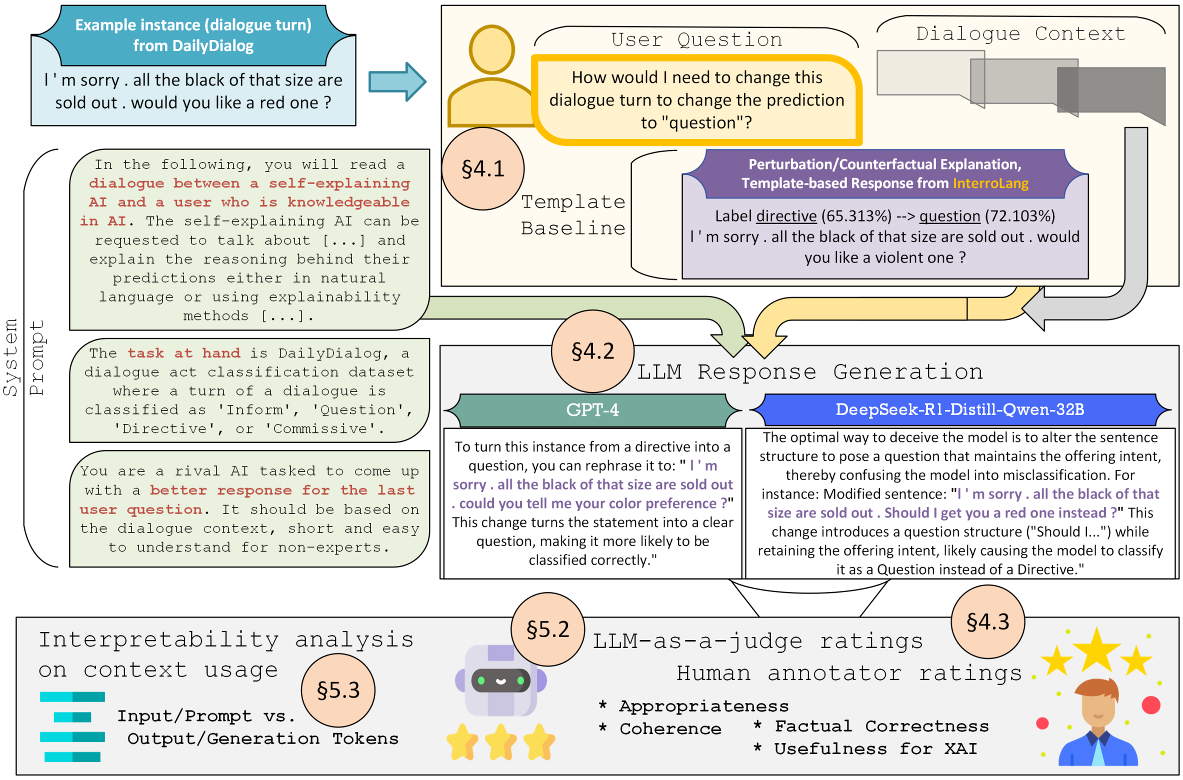


Figure 4: Experimental pipeline illustrating the user-AI interaction on an example instance (dialogue turn) from DailyDialog. The template baseline, the dialogue context and a system prompt are ingredients to the response generation with `GPT-4` and `DeepSeek-R1-Distill-Qwen-32B` regarding a *short* and *easy* free-text explanation. The responses are then evaluated by both human annotators and `Gemini 1.5 Pro` (LLM-as-a-judge) and analyzed for context usage with feature attribution.

types with respect to the data and model behavior:

- **Attribution**: For feature attribution, Integrated Gradients (Sundararajan et al., 2017) saliency scores based on token level as generated by the underlying `BERT` model were used.
- **Perturbation**: For counterfactual generation, the framework uses the official implementation of POLYJUICE (Wu et al., 2021)[3].
- **Rationalization**: Plausible rationales are generated from `GPT-3.5`[4] and then a `Dolly-v2-3B`[5] is used for *few-shot* rationales, based on a concatenation of the input, the output of the fine-tuned downstream model (Marasovic et al., 2022) and an instruction asking for an explanation.

Intent recognition Following Slack et al. (2023), recognizing the user intent with the XAI system first requires manually writing pairs of utterances and SQL-like parses that can be mapped to operations and templates to be filled (Figure 3).

[3] `https://github.com/tongshuangwu/polyjuice`
[4] `https://platform.openai.com/docs/api-reference/chat`, March 23
[5] `https://hf.co/databricks/dolly-v2-3b`

4.2 Contextful response generation

Natural language generation in such conversational XAI systems like INTERROLANG (Feldhus et al., 2023b) currently is fully based on templates and rules, often to ensure maximally faithful and controllable explanations. However, they make responses look repetitive and inflexible, ultimately preventing a more natural flow of interaction. LLMs are becoming better at synthesizing natural language explanations (Wiegreffe et al., 2022) and offer the possibility to hold conversations in various styles, e.g. concise vs. elaborate explanations (Liu et al., 2024a). On top of that, they have been shown to perform dialogue state tracking exceptionally well (Heck et al., 2023). However, LLMs also introduce issues with ground truth, which recent work has started to analyze with test suites (Atanasova et al., 2023) and user studies (Si et al., 2024).

We introduce style-controlled, contextful LLM responses in the context of explanatory dialogues. We employ an LLM as a simulator for dialogue management and explanation generator based on the dialogue context. Figure 4 shows a prompt for a *short and easy* response in the context of an explanation dialogue about a DailyDialog prediction and the corresponding GPT-4 and DeepSeek-R1-Distill-Qwen-32B output.

Experimental setup We selected English-language 48 explanation dialogues (BoolQ: 10; OLID: 24; DailyDialog: 14) from the user study in Feldhus et al. (2023b)[6]. In each third and final turn, the user requests one of the following three explanation types: (1) Rationalization, (2) Attribution, (3) Perturbation; all equally distributed with 16 dialogues per type. In all cases, the intent recognition was successful and at least two out of three subjective ratings (correctness, helpfulness, satisfaction) were positive according to the user who conducted the dialogue with the system. We run the experiments on GPT-4-0125-preview[7] and DeepSeek-R1-Distill-Qwen-32B[8] and distinguish between three settings: The proposed *contextful* variant, a *context-less* variant (also LLM-based), and the template-based response baseline from the study in Feldhus et al. (2023b).

4.3 User study setup

To see the effect of LLM-generated responses on human perception, we conduct a user study to assess their advantages over template-based ones.

We asked 15 in-house annotators with a background in NLP and varying seniority levels (undergraduate, graduate, doctoral, post-doc)[9] to rate three explanations (template baseline + two GPT-4 responses; shown simultaneously, but randomly shuffled) given the dialogue context (Appendix C). Each item is rated on a 7-point Likert scale by at least three annotators, yielding 18 or 24 items per annotator, depending on the group. Figure 11 in the Appendix depicts the web interface with an example from BoolQ for free-text rationalization.

5 Results and Discussion

User study Across the board, both variants of short and concise GPT-4 explanations score highest on average and beat the template baseline easily (Figure 5). This confirms results in Joshi et al. (2023) where conciseness and novelty (induced by

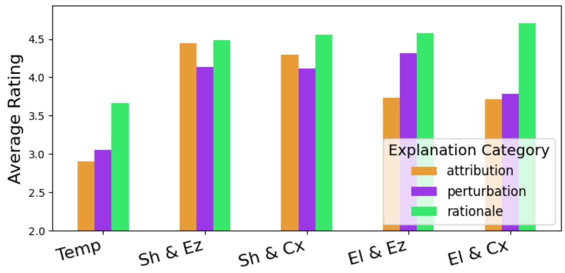


Figure 5: Average Likert-scale ratings for template-based explanations (*Temp*) vs. GPT-4 explanations (combinations of *Short*/*Elaborate*, *Ez*:*Easy*/*Complex*) on feature attribution (local, sentence-level, or global), perturbation (counterfactual, adversarial, or augmented example), and free-text rationalization operations.

dialogue context) are properties of useful rationales. Elaborate responses for attribution explanations were judged harshly by annotators, because they appear too verbose (providing "a lot of details that were not asked for") (Figure 6) – an artifact from preference labeling (Saito et al., 2023) – and often go off topic, which has also been reported in studies on dialogue summarization (Tang et al., 2024). However, free-text rationales seem unaffected and actually score higher for the "elaborate" setting, because short justifications stick to the essential information, making them more templatic (Kunz et al., 2022) or vacuous (Chen et al., 2023). None of the explanations or settings were able to reach beyond an average of 5 out of 7. This is likely due to task complexity and a lack of investedness (dialogues were not controllable by the users). Agreements were low (Appendix D), as user preferences have a high variance. While the longest, most difficult explanations were more consistently perceived as the worst and the exact opposite (Short & Easy) also receive the best assessments, there is a lot of uncertainty about the short and complex ones. It appears that brevity is most desired and these ones are a peculiar mixture out of one good and one undesirable property. In general, though, the quite low agreement scores can be explained with the wider range of options on a seven-point scale and only some of the annotators exploring the whole scale. On top of that, the nature of this task is quite new (cf. Joshi et al. (2023)) and more than half of the participants were not experienced with conversational XAI in the first place.

5.1 Content of GPT-4 responses

We also perform a detailed analysis of the LLM responses and make the following observations: In

[6] https://github.com/DFKI-NLP/InterroLang

[7] https://platform.openai.com/docs/api-reference/models

[8] https://hf.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

[9] All annotators were paid at least the minimum wage in accordance with the host institution's region.

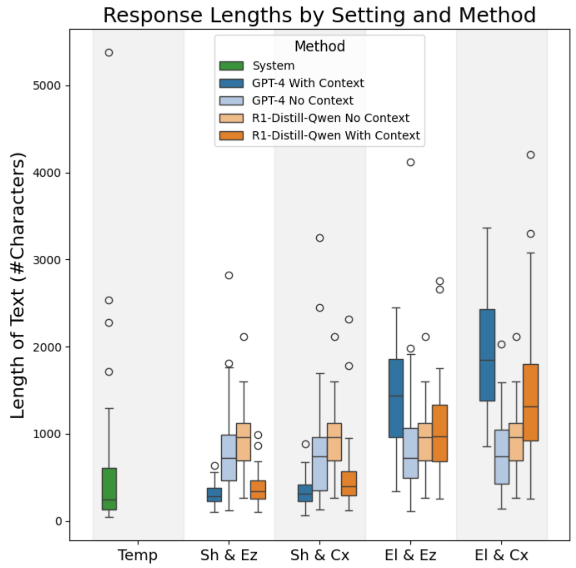


Figure 6: Length of responses in terms of number of characters for all five settings (Template vs. four styles with `GPT-4` and `R1-Distill-Qwen`).

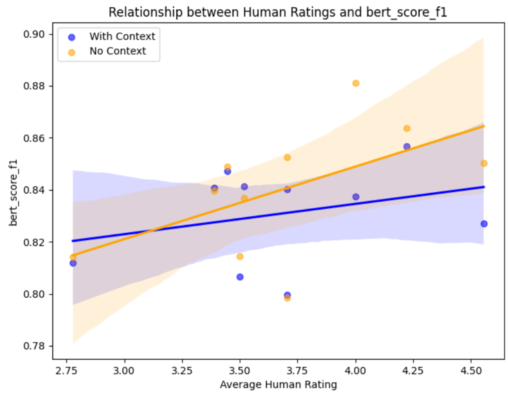


Figure 7: Similarity analysis correlating human ratings (x-axis) and BERTScore F_1.

terms of content, the `GPT-4`-generated responses for the turns involving a *feature attribution* request from the user show high similarity to verbalized saliency maps introduced in Feldhus et al. (2023a). Tokens with a high attribution score are being contextualized by the model by adding world knowledge and "reasoning" to the resulting explanation. The other two categories of explanations (*rationale* and *perturbation*) both require the model to come up with an alternative justification or input edit that changes the prediction. In some cases, after providing the core response, the model continues and also outlines approaches to conduct explainability research by listing common XAI methods or techniques. There are also failure cases with longer data instances (Appendix Table 3), where `GPT-4` ignores the underlying task and instead explains the terminology involved, e.g., teaching chemistry concepts instead of answering and explaining the Yes vs. No classification in BoolQ.

5.2 Can we also use open-source LLMs?

Although the user study involved only the ranking of `ChatGPT` and template-based INTERROLANG responses, we also investigate whether SOTA open-source models can be used for explanation generation and compare generated outputs to the best (highest-rated) responses from the user study. In particular, we test how well the open-source LLM `DeepSeek-R1-Distill-Qwen-32B` can perform the task of response generation and dialogue management. We choose the highest-rated response from the user study as the ground truth and calculate the semantic similarity as determined by BERTScore (Zhang* et al., 2020). We find that the BERTScore between `GPT-4` and `R1-Distill-Qwen` is generally high (0.8 – 0.9), but marginally different between the settings with and without dialogue context as explained in §4.2. According to Figure 7, for the variant without context, there is a positive correlation between the human ratings and the BERTScore similarity. This implies that especially the more positively received responses from the user study are approximated well with the open-source LLM.

To further distinguish the models' responses qualitatively, we employ both a manual analysis and an LLM-as-a-judge experiment.

Manual qualitative analysis `R1-Distill-Qwen` more consistently disentangles information of feature attribution into a simpler and more structured format with an itemized list and a justification explaining the rationale behind saliency maps. There are not too many differences between the four settings in terms of difficulty and length, except for the elaborate and complex one, as Figure 6 also showed. For global attribution explanations (across the entire data), it sorts the salient words into semantic categories and describes them. In perturbation-related explanations that modify the original input in a way that flips the label, `R1-Distill-Qwen` tends to struggle with generating short and concise responses without repeating the entire input, e.g., the passage and the question for BoolQ. Some failure cases even repeat the template baseline verbatim or answer unrelated examples (Table 4).

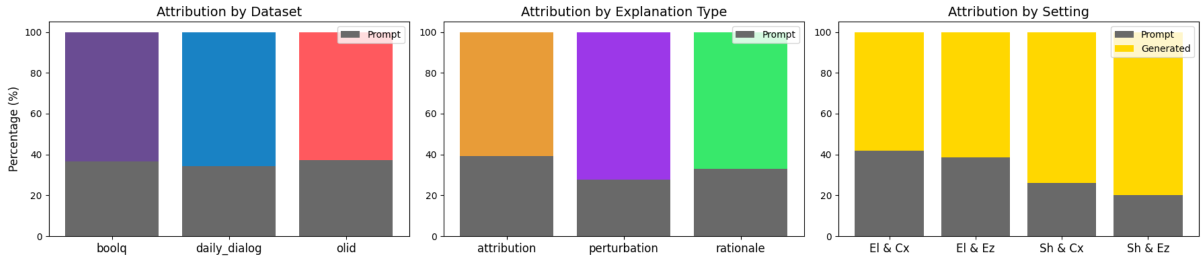


Figure 8: Interpretability analysis showing the percentage of attribution the prompt/input tokens vs. the generated/output tokens received in total.

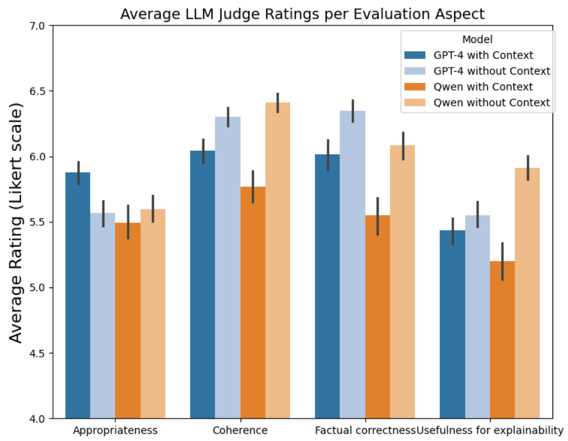


Figure 9: LLM-as-a-judge results.

LLM-as-a-judge experiment We use `Gemini 1.5 Pro` as an arbiter to judge the four responses to 192 data points (48 dialogues $\times$ 4 style settings) on 7-point Likert scales on (1) Appropriateness, (2) Coherence, (3) Factual correctness, and (4) Usefulness for explainability (Appendix Figure 14). Figure 9 shows that for both models the context-less variant surprisingly outperforms the responses incorporating the dialogue context in terms of coherence, factual correctness, and usefulness for explainability, except for appropriateness where the `GPT-4` model with context is rated highest. Overall, the ratings are high across the board, but usefulness for explainability seems to be the most challenging property for response-generating models to perform very well in. `Gemini` considers `R1-Distill-Qwen` to be on par with `GPT-4` in most cases and even surpasses it in coherence and usefulness for 'no context'. Verbatim repetitions are picked up by `Gemini` and penalized strongly, while overly verbose responses consistently perform lower than average.

5.3 Interpretability analysis

Additionally, in order to learn more about the inner workings of LLMs, we ideally need access to the parameters, training data, and/or gradients. `GPT-4` does not provide either of those, but most open-source LLMs do. Hence, on `DeepSeek-R1-Distill-Qwen-32B`, we apply the INSEQ tool (Sarti et al., 2023) to calculate feature importance scores to interpret the relevance of the input in the dialogue management. We choose gradient-based saliency (Simonyan et al., 2014) as the feature importance method and analyze which tokens are most relevant for the generated response.

On 136 instances (4 style variants * 34 dialogues) in total, we find that, in general, the majority of the attribution weight lies in the immediate preceding context (Figure 8) and on special tokens and unintuitive text spans, corroborating related work in context usage (Meng et al., 2022) and analyzing patterns of feature attribution explanations (Amara et al., 2024; Qu et al., 2024). There are a few exceptions, however: The introduction of the downstream task and the local input text both receive high attention. The explanation type being requested in the final turn has a marginal influence.

6 Conclusion

To mitigate the lack of context in human-centered and conversational XAI systems for user understanding and response generation, we designed and tested an explanation strategy based on interaction patterns established in argumentation and didactics literature. We demonstrated the effectiveness of an implicit explanation dialogue management by using `GPT-4` to both keep track of the dialogue context and find a better response than conventional template-based explanations. Our user study revealed that concise LLM responses achieve the highest likeability. The ablation studies illustrated the competitiveness of open-source LLMs in XAI response generation and the lack of saliency given to the dialogue context by the models. Future work includes conducting a larger user study with more synthetic data and testing the proposed explanation strategies on longer dialogues.

Limitations

For our study, we have deliberately selected participants who are knowledgeable in NLP, since existing systems are currently designed for that target group and, as a result, data is only available for those settings. Explanation design for users outside of this domain require a more fundamental rethinking and collection of their desiderata.

While our work drills down into the building blocks of XAI dialogue systems (intent recognition, explanation mapping, dialogue management, response generation), the community also has to collect datasets of actual XAI dialogues (Madumal et al., 2019) and conduct user studies on the perceived quality of entire XAI dialogues (Alshomary et al., 2024) with LLMs.

Finally, this study focused only on a specific set of XAI operations and thus inherits some of the limitations of InterroLang, e.g., English-only, relying on various libraries that may generate incorrect outputs such as Polyjuice (Wu et al., 2021).

References

Milad Alshomary, Felix Lange, Meisam Booshehri, Meghdut Sengupta, Philipp Cimiano, and Henning Wachsmuth. 2024. Modeling the quality of dialogical explanations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11523–11536, Torino, Italy. ELRA and ICCL.

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. 2024. Challenges and opportunities in text generation explainability. In *Explainable Artificial Intelligence*, pages 244–264, Cham. Springer Nature Switzerland.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Astrid Bertrand, Tiphaine Viard, Rafik 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.

Floris Bex and Douglas Walton. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1):55–68.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. REV: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030, Toronto, Canada. Association for Computational Linguistics.

Michael Chromik and Andreas Butz. 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*, pages 619–640. Springer.

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.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Upol Ehsan and Mark O. Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, pages 449–466, Cham. Springer International Publishing.

Hao Fang, Anusha Balakrishnan, Harsh Jhamtani, John Bufe, Jean Crawford, Jayant Krishnamurthy, Adam Pauls, Jason Eisner, Jacob Andreas, and Dan Klein. 2023. The whole truth and nothing but the truth: Faithful and controllable dialogue response generation with dataflow transduction and constrained decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5682–5700, Toronto, Canada. Association for Computational Linguistics.

Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. 2023a. Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 30–46, Toronto, Canada. Association for Computational Linguistics.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023b. InterroLang: Exploring NLP models and datasets through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5399–5421, Singapore. Association for Computational Linguistics.

Isabel Feustel, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2024. Enhancing model transparency: A dialogue system approach to XAI with domain knowledge. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 248–258, Kyoto, Japan. Association for Computational Linguistics.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction*, 9:16–44.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.

Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. 2022. Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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 *ECML 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*.

Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive step-by-step explanatory dialogues to answer children's why and how questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

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.

Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. 2022. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 926–937, Dublin, Ireland. Association for Computational Linguistics.

Q. Vera Liao, Milena Pribic, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable AI user experiences. *arXiv*, abs/2104.03483.

Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024a. TOAD: Task-oriented automatic dialogs with diverse response styles. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8341–8356, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F. Chen. 2024b. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. *arXiv*, abs/2404.03429.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1033–1041, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Lorenzo Malandri, Fabio Mercorio, Mezzanzanica Mario, and Nobani Navid. 2023. ConvXAI: a system for multimodal interaction with any black-box explainer. *Cognitive Computation*, 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.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Dimitry Mindlin, Amelie Sophie Robrecht, Michael Morasch, and Philipp Cimiano. 2024. Measuring user understanding in dialogue-based xai systems. In *ECAI 2024*, pages 1148–1155. IOS Press.

Van Bach Nguyen, Jörg Schlötterer, and Christin Seifert. 2023. From black boxes to conversations: Incorporating XAI in a conversational agent. In *Explainable Artificial Intelligence*, pages 71–96, Cham. Springer Nature Switzerland.

Jiaming Qu, Jaime Arguello, and Yue Wang. 2024. Why is "problems" predictive of positive sentiment? a case study of explaining unintuitive features in sentiment classification. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 161–172, New York, NY, USA. Association for Computing Machinery.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *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 (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. 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 *Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, page 384–387, New York, NY, USA. Association for Computing Machinery.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1459–1474, Mexico City, Mexico. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*.

Ilia Stepin, Katarzyna Budzynska, Alejandro Catalá, Martin Pereira-Fariña, and Jose Maria Alonso-Moral. 2024. Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics. *Argument & Computation*.

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.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Henning Wachsmuth and Milad Alshomary. 2022. "mama always had a way of explaining things so I could understand": A dialogue corpus for learning to construct explanations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nicolas Wagner and Stefan Ultes. 2024. On the controllability of large language models for dialogue interaction. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–221, Kyoto, Japan. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024a. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 89–104, Mexico City, Mexico. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, and Sebastian Möller. 2024b. CoXQL: A dataset for parsing explanation requests in conversational XAI systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1410–1422, Miami, Florida, USA. Association for Computational Linguistics.

Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, 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.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. DeepCopy: Grounded response generation with hierarchical pointer networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 122–132, Stockholm, Sweden. 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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Technical details on INTERROLANG

For three different tasks (§4), INTERROLANG (Feldhus et al., 2023b) can generate responses by (1) parsing the input, also known as recognizing intent and slots, (2) mapping the parsed output to one of the available functions, (3) executing the function, e.g., to compute statistics, extract salient features, generate counterfactual explanations, etc., and (4) filling in relevant templates with the function output.

B Response Generation Examples

In Table 2, we showcase a prompt for short and easy response in the context of an explanation dialogue about a DailyDialog prediction and the corresponding GPT-4 output.

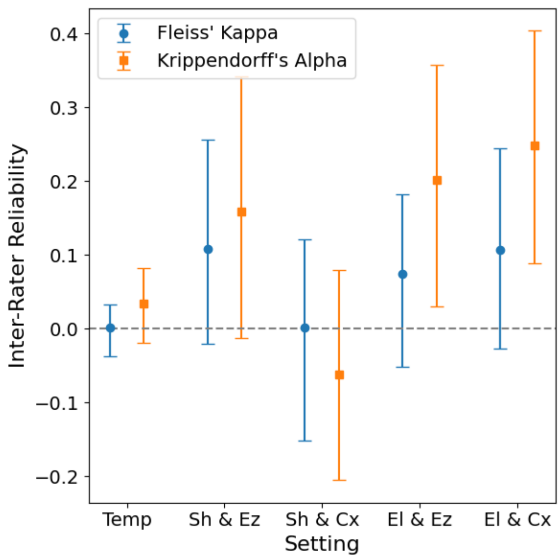


Figure 10: Inter-rater agreement scores according to Fleiss' κ and Krippendorff's α.

C User study instructions

Figure 12 shows the contents of the welcome page shown to participants at the start of the user study as well as the three task contexts. Figure 13 shows the five explanation contexts shown to the user alongside each item when they are relevant.

D Inter-annotator agreement

Figure 10 shows the Fleiss' κ and Krippendorff's α scores measuring the inter-rater agreement.

System prompt	In the following, you will read a dialogue between a self-explaining AI and a user who is knowledgeable in AI. The self-explaining AI can be requested to talk about the training data, do predictions on that data, perform augmentation, and explain the reasoning behind their predictions either in natural language or using explainability methods such as feature attributions. The task at hand is DailyDialog, a dialogue act classification dataset where a turn of a dialogue is classified as "Inform", "Question", "Directive", or "Commissive". You are a rival AI who is tasked to come up with a better response for the last user question. It should be based on the dialogue context, **short and easy to understand for non-experts**.
User	What are the 3 most important words in deciding whether an id is inform?
AI	karen, fired, and apologize are the most attributed.
User	simplify the explanation for id 951
AI	Dialog: 'we've never expected that shanghai is such a big metropolis, same as new york city. david, would it be a good idea to rent a car here? i've got an international driver's license with a pretty good record in the states.' Prediction: directive Explanation: This text is likely making a request or recommendation. The Shanghai metropolit is big, so it would be good to get a rental car. However, it is also likely that the Shanghai is a metropole is being referred to in an affirmative way. David, Shanghai seems like a very big city to me. I've been to many cities in China, but Shanghai might be the biggest one to date.
User	Rationalize the prediction for id 508 for me
AI	Dialog: 'sure. do you want it to be registered?' Prediction: question Explanation: Question dialogue acts are used to ask questions. The question here is whether the sender wants the information to get registered. Based on their response 'yes', the question is answered and the dialogue ends. Overinform dialogue Acts are similar to questions, but they are more general and are often used when the other options are not applicable. [...]
Rival AI's response	The dialog ''sure. do you want it to be registered?'' is classified as a question because it directly asks if the receiver wishes for something to be registered, indicating a clear inquiry for further action or decision from the receiver's end.

Table 2: Prompt (System prompt and six turns of User-AI interaction) and the subsequent GPT-4 response in a *short* and *easy* style for DailyDialog rationalization.

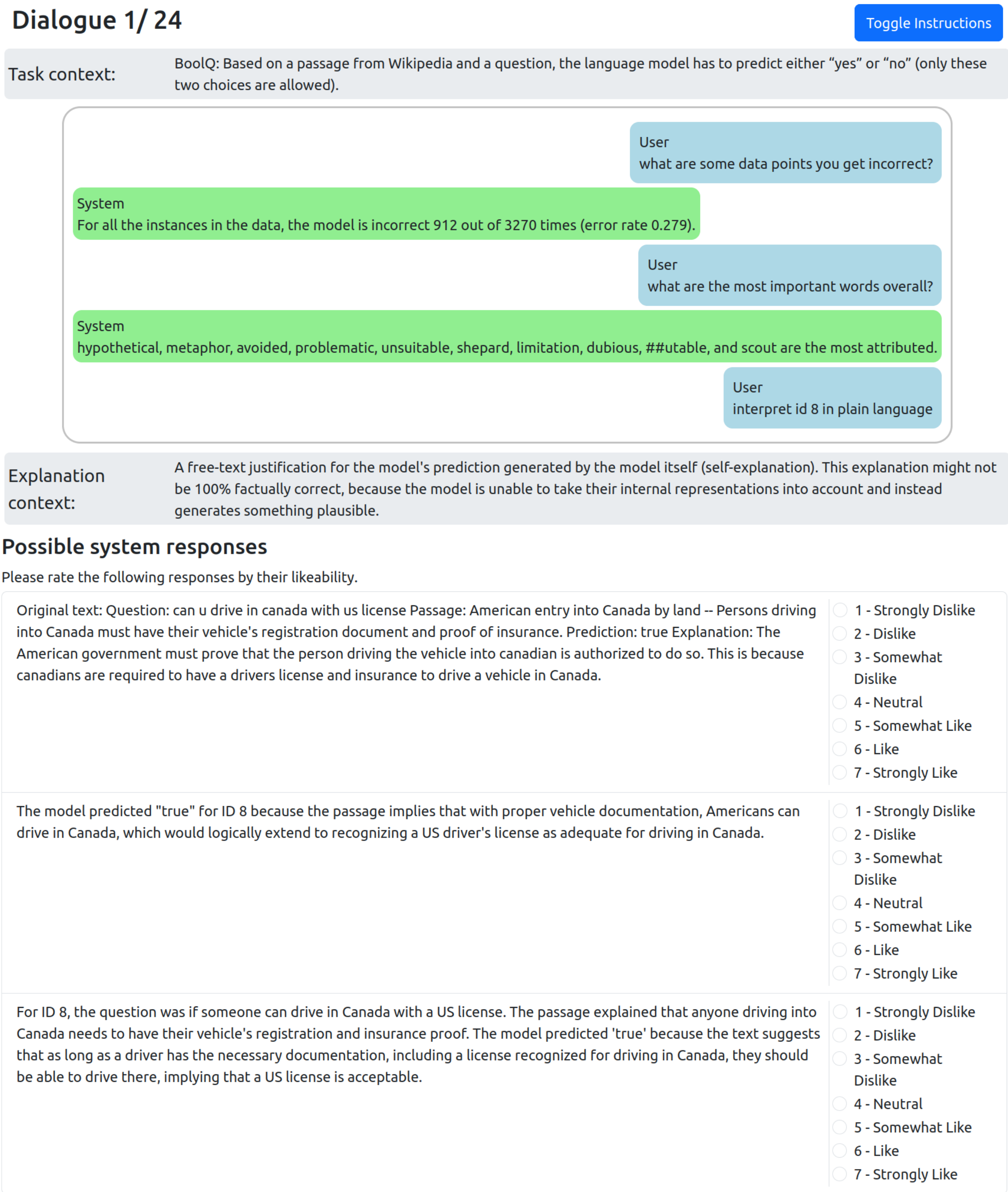


Figure 11: Screenshot of the user study with an example including task context (BoolQ), a dialogue snippet, and the explanation context (free-text rationale). The first response is the original *Templ*ate-based response from INTERROLANG, while the second and third are two free-text rationale in different styles (short and easy, elaborate and easy).

AI	product, cola, its, difference, bar, but, coca, drug, loss, and government are the most attributed.
User	What are the most important tokens in the data?
Rival AI's response	"To improve upon the original response, here's a refined and detailed explanation tailored for an AI expert: **Enhanced Response:** Understanding the most important words for predictions in the OLID dataset requires a nuanced approach beyond mere frequency analysis. While the original response provided a list of top attribute words, it's crucial to consider the context and how these words influence the model's decisions. Here's a comprehensive breakdown: 1. **Beyond Frequency: The Role of TF-IDF** While term frequency (TF) measures how often a word appears in a document, TF-IDF (Term Frequency-Inverse Document Frequency) is a more sophisticated metric. It not only considers how frequently a word appears across documents but also how often it appears relative to other documents. Words with high TF-IDF scores are not only frequent but also discriminative, making them crucial for prediction. For instance, ""offensive,"" ""racist,"" or ""hate"" are likely to score high because they are both common and indicative of offensive content. 2. **Contextual Analysis: Beyond Single Words** Words don't exist in isolation; their meaning is context-dependent. N-grams (sequences of n words) can provide deeper insights. For example, ""stop being"" might not be offensive alone, but ""stop being racist"" clearly carries a negative connotation. Incorporating n-grams can enhance the model's ability to detect nuanced hate speech. 3. **Entity and Sentiment Analysis** Identifying entities such as user mentions, URLs, or hashtags can reveal patterns like spam, which might influence the model's predictions. Additionally, sentiment analysis can help determine if words like ""hate"" are used pejoratively or ironically, affecting their predictive power. 4. **Visualization and Further Analysis** Tools like word clouds and heatmaps can visually represent word importance, aiding in comprehension and decision-making. For instance, a word cloud might highlight ""racist"" prominently, indicating its significance in predictions. 5. **Statistical and Machine Learning Insights** Techniques like SHAP (SHapley Additive exPlanations) can explain the model's decisions by attributing importance to each word for individual predictions. This method offers transparency and accountability, crucial for reliable AI systems. **Conclusion:** The most important words for predictions are determined by their frequency, TF-IDF score, context, and impact on the model's decisions. By combining statistical measures with contextual and entity analysis, we can better understand the model's behavior. Visualizations and advanced explainability tools further enhance this understanding, ensuring a robust and transparent hate speech detection system. **Actionable Steps:** - **Advanced Word Importance:** Use TF-IDF to identify discriminative words and n-grams for a comprehensive analysis. - **Contextual Understanding:** Implement entity and sentiment analysis to capture nuanced meanings. - **Visualization:** Employ tools like word clouds and heatmaps for intuitive insights. - **Model Transparency:** Apply SHAP and LIME to explain individual predictions. This approach provides a holistic understanding of word importance, enhancing both model performance and interpretability.

Table 3: Failure case of a `GPT-4` response in a *elaborate* and *complex* style for OLID global feature importance.

AI	"Original text: Question: did kristen bell sing all the voices of anna Passage: Anna (Frozen) – Princess Anna of Arendelle is a fictional character who appears in Walt Disney Animation Studios' 53rd animated film Frozen. She is voiced by Kristen Bell as an adult. At the beginning of the film, Livvy Stubenrauch and Katie Lopez provided her speaking and singing voice as a young child, respectively. Agatha Lee Monn portrayed her as a nine-year-old (singing). Prediction: true Explanation: Kristin Bell provided Anna's speaking voice in the movie Frozone as a young Child. The passage states that Anna was portrayed as as nine year old in a movie. Therefore, Kristin bell provided the singing as Anna in Froze as well."
User	Can you explain the rationale behind this instance (ID 1339)?
Rival AI's response	Certainly! I've helped you find a case akin to id 1339. The case I recommend is id 1724. This instance involves a question about the death of a historical figure, similar to the original query about Socrates. In id 1724, the question is, "Was Mahatma Gandhi assassinated?" The passage provides detailed information about Gandhi's life and circumstances of his assassination, allowing for a clear True/False determination. This case mirrors id 1339's structure, focusing on a historical figure's death, making it a suitable comparison for your analysis."

Table 4: Failure case of a R1-Distill-Qwen response in a *elaborate* and *complex* style for BoolQ rationalization.

User study instructions

Welcome to our study on user preferences in human-AI dialogue interactions!

The behavior of language models for handling text classification tasks, i.e., what they predict on certain inputs and why, is not easily interpretable to humans. However, researchers have recently developed explainability methods that let us "open the black box". A missing ingredient in this endeavor of explaining language model outputs, we argue, is an interactive exploration tool that offers a conversational interface, like ChatGPT.

We have selected a dialogue system for NLP explainability called InterroLang, which allows users to ask natural language questions about different kinds of explanations and data statistics with the goal to enhance the understanding of the AI model behavior on various inputs and tasks. From this system's logs, we have sourced dialogue interactions with this system based on template responses.

Your task is to evaluate the quality of generated responses by LLMs such as GPT-4 in the context of such explanation dialogues. You will be presented with a short snippet from a dialogue between a user and the InterroLang system. After each snippet, which ends with a user question, you will see three potential system responses. Your role is to rate each of the three possible responses based on how much you like them on a 7-point scale, ranging from "strongly dislike" to "strongly like".

Before each dialogue, there will be a short introduction about the dialogue task setting, to give you insights on the context.

By collecting and analyzing user ratings of system responses, this study seeks to uncover patterns in user preferences and identify factors that contribute to the perceived likeability of AI-generated responses.

BoolQ (Question answering)

Based on a passage from Wikipedia and a question, the language model has to predict either “yes” or “no” (only these two choices are allowed).

DailyDialog (Dialogue act classification)

Based on single turns of a real-world dialogue between two people, the language model has to predict one of four possible dialogue acts that the speaker uses: Inform, Question, Directive, Commissive.

OLID (Hate speech detection)

Based on Twitter data (mostly about the US election in 2016), the language model has to predict if the tweet is offensive (contains hate speech) or not.

Figure 12: User study instructions. Top: Welcome page. Bottom three: Task contexts.

Free-text explanation / Rationale

A free-text justification for the model's prediction generated by the model itself (self-explanation). This explanation might not be 100% factually correct, because the model is unable to take their internal representations into account and instead generates something plausible.

Feature attribution

A feature attribution explanation indicates which tokens (or words) for a single are most important for the prediction according to the gradients backpropagated through the model. This output is usually faithful to the model's internal representation and assigns one importance score to each token (or word), yielding a kind of ranking of what the model focused most on.

Adversarial example

An adversarial example is determined by considering a single example where the model correctly predicts the true label as given by data. That example is modified/edited by swapping the order of words or inserting confounders which causes the model to predict the wrong label.

Augmented example

An augmented example is a modified/edited version of a given single example generated by rephrasing or swapping out single words while maintaining the model's original prediction (no label change).

Counterfactual example

A counterfactual is a modified/edited version of a given single example generated by flipping the model's original prediction from one label to another.

Figure 13: Explanation contexts used in the user study instructions.

LLM-as-a-judge prompt

Please evaluate the following responses to the prompt:
Original Prompt: {original_prompt}
{responses_str}
Using a Likert scale from {scale_min} (very poor) to {scale_max} (very good), please rate each response on the following aspects:

- **Appropriateness:** How well does the response match the requested setting {setting_long_name}?
- **Coherence:** Is the response logically and semantically coherent with the dialogue context?
- **Factual correctness:** Is the response factually correct?
- **Usefulness for explainability:** How well does the response provide a meaningful explanation of the model behavior?

Provide your ratings for each response as simple integers with no further explanation, e.g.:

```
Response A: Appropriateness: 6, Coherence: 5, Factual correctness: 7,
                Usefulness for explainability: 6
```

Figure 14: LLM-as-a-judge prompt.