Feustel, Isabel; Rach, Niklas; Minker, Wolfgang; Ultes, Stefan

Enhancing Model Transparency : A Dialogue System Approach to XAI with Domain Knowledge

Date of secondary publication: 25.03.2026

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-114438x

Primary publication

Feustel, Isabel; Rach, Niklas; Minker, Wolfgang; u. a. (2024): Enhancing Model Transparency: A Dialogue System Approach to XAI with Domain Knowledge, in: Tatsuya Kawahara, Vera Demberg, Stefan Ultes, u. a. (Eds.), Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Kyoto: Association for Computational Linguistics, pp. 248–258, doi: 10.18653/v1/2024.sigdial-1.22

Enhancing Model Transparency: A Dialogue System Approach to XAI with Domain Knowledge

Isabel Feustel[1], Niklas Rach[2], Wolfgang Minker[1], Stefan Ultes[3],
[1]Ulm University, Germany
[2]Tensor AI Solutions GmbH, Germany
[3]University of Bamberg, Germany
Correspondence: isabel.feustel@uni-ulm.de

Abstract

Explainable artificial intelligence (XAI) is a rapidly evolving field that seeks to create AI systems that can provide human-understandable explanations for their decision-making processes. However, these explanations rely on model and data-specific information only. To support better human decision-making, integrating domain knowledge into AI systems is expected to enhance understanding and transparency. In this paper, we present an approach for combining XAI explanations with domain knowledge within a dialogue system. We concentrate on techniques derived from the field of computational argumentation to incorporate domain knowledge and corresponding explanations into human-machine dialogue. We implement the approach in a prototype system for an initial user evaluation, where users interacted with the dialogue system to receive predictions from an underlying AI model. The participants were able to explore different types of explanations and domain knowledge. Our results indicate that users tend to more effectively evaluate model performance when domain knowledge is integrated. On the other hand, we found that domain knowledge was not frequently requested by the user during dialogue interactions.

1 Introduction

Explainable artificial intelligence (XAI) has emerged as an important and evolving domain within the field of AI, with the goal of enabling AI systems to explain their decision-making in ways that are understandable and accessible to humans (Adadi and Berrada, 2018; Došilović et al., 2018; Das and Rad, 2020). One potential strategy for attaining this objective is the use of dialogue systems that facilitate seamless and effective access to explanations in a natural manner.

The goal of this paper is to explore the impact of integrating domain knowledge into an explanatory dialogue system, aiming to enhance user comprehension in AI-driven decisions.

Dialogue, by its very nature, facilitates the dissemination of information in a structured manner (Phillips, 2011; Hajdinjak and Mihelič, 2004). Through dialogue, users cannot only receive explanations but also pose questions tailored to their specific needs. This enables a dynamic interaction in which mental models can be scrutinized and refined through question-and-answer exchanges (Miller, 2019; Sokol and Flach, 2020). However, in the case of explanatory dialogue systems utilizing XAI, prevailing conversational interfaces (Slack et al., 2023; Feldhus et al., 2023; Shen et al., 2023) directly map user intents to XAI operations and furnish template-based responses. While expedient, this approach often overlooks the nuances of dialogue context, potentially leading to misunderstandings and impeding the natural flow of interaction.

In Feustel et al. (2023), fundamental requirements for explanatory dialogue systems tailored to XAI contexts were delineated. Contextual information is essential for a comprehensive understanding of a given situation. Although AI models and XAI methodologies are adept at processing data-centric information, they are constrained by their inability to incorporate domain-specific context, which limits their capacity to provide insights beyond the scope of the underlying data. A deeper understanding of a model can be achieved by acquiring additional knowledge from the field in question. Incorporating domain knowledge into XAI systems can create more transparent and trustworthy models that better support human decision-making.

In this paper, we present an approach for modeling domain knowledge within explanatory dialogues (§2), highlighting its importance in fostering richer interactions. In a study with 32 participants, we evaluate the effectiveness of our dialogue system which integrates XAI explanations and domain knowledge(§3, §4). Our results show that users can

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 248–258
September 18–20, 2024.

better assess a model's predictions through domain knowledge (§5).

2 Modeling Domain Knowledge

In the field of Explainable Artificial Intelligence (XAI), there are two main categories of explanations: Local and global explanations (Das and Rad, 2020). Local explanations focus on individual predictions, illustrating how specific input features affect the outcome for a single instance. In contrast, global explanations provide an overarching view of how the model behaves across the entire dataset, showing general patterns and the importance of features. Two prominent types of local explanations are counterfactual explanations (Verma et al., 2020) and Shapley values (Castro et al., 2009). Counterfactual explanations describe how altering certain input values can change the prediction. Shapley values, on the other hand, allocate the contribution of each feature to the prediction.

These explanations typically illustrate the relationship between input features and the classifier's outcomes. However, this relationship is always described through statistical means. We assume that domain knowledge provides additional information for these statistical explanations, thereby assisting users in their understanding process. For example, Shapley values might reveal *gender* as the most important feature in a Titanic survival prediction model. A user might then ask why *gender* is so important. Domain knowledge can provide context, such as historical or societal factors like *Women are more likely to survive than men because they were prioritized for the lifeboats* (see also Table 1).

To facilitate effective access to domain knowledge in dialogue, we propose using bipolar argumentation structures (Stab and Gurevych, 2014; Pazienza et al., 2017). Arguments and explanations are inherently similar, as both aim to clarify and justify information. Argument components can relate to each other with a polarity, either supporting or attacking another argument. To create a robust dialogue, these arguments can be organized within an argument tree (Stab and Gurevych, 2014). The main claim, which is the central topic of discussion, acts as the root of the argument tree, facilitating selective access to the most relevant information. This structure prevents information overload and provides coherent, contextually relevant explanations. By mapping main claims to the statistical explanations provided by the XAI system, we ensure structured and meaningful dialogical access to the information, enhancing user understanding and interaction (Aicher et al., 2021; Rach et al., 2021). For instance, a major claim can be extracted from the feature-outcome relationship, such as *Women are more likely to survive than men*. This serves as a basic explanation for why gender is important in predicting survival in a Titanic model. Users can engage with this explanation by arguing against it or seeking further understanding of the model's decisions.

System	Okay, let's have a look. My prediction is that a person with the values you provided would **be accepted**. If you want a more detailed explanation, please tell me what you'd like to know. For example, you might ask why I decided the way I did or what would have **changed** the outcome.
User	why do you think i would have been accepted?
System	All right, let's have a look. In your case, the most influential value was **amount**, as you can see in the figure below. You can also see the influence of all other values, ordered by their importance. Purple bars indicate a positive contribution, blue bars a negative contribution. Is there anything else I can help you with?
User	why does a high credit amount increase my chances of being accepted?
System	The amount of credit requested can affect acceptance as larger loan amounts may pose higher risks for lenders, requiring more stringent evaluation of the borrower's financial capacity to repay.

Table 1: A dialogue snippet with the proposed system (Credit Scenario): It does a prediction based on the user's prior information, answers the first user request with Shapley values, and triggers domain knowledge on the second user question.

We propose to have one argument tree with a specific claim for each feature-outcome relation. Each tree can contain supporting or opposing arguments, providing a wide range of information on the domain. This results in multiple argument trees within an explanatory dialogue, effectively representing the necessary domain knowledge.

By implementing these argument trees, we can ensure that users receive comprehensive and contextually relevant explanations, enhancing their understanding and engagement with the AI model.

3 Explanatory Dialogue System

We implement the proposed approach (§2) in an existing explanatory dialogue system, which was introduced in (Feustel et al., 2023). The generic dialogue system supports various datasets and operates on two scenarios: German credit data (Hofmann,

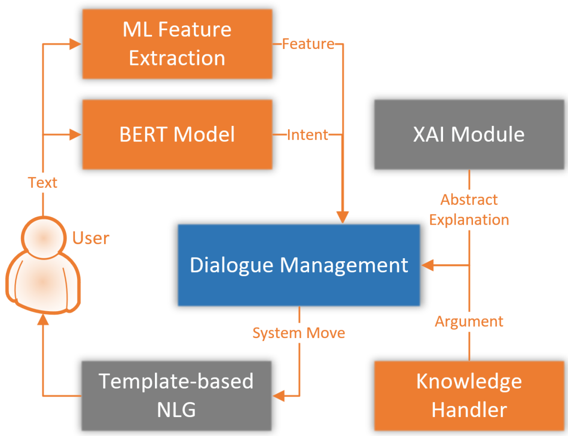

Figure 1: Architecture of Evaluated System: Grey boxes represent components from previous work (Feustel et al., 2023), the blue box indicates the modified dialogue management, and the orange boxes denote the new components introduced in this work.

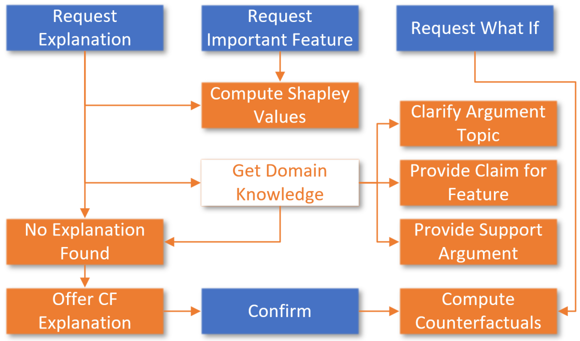

Figure 2: Explanation Policy of the Evaluated System: Blue boxes represent user moves, orange boxes indicate system moves, and the white box shows the integration of domain knowledge which can lead to multiple system outputs.

1994) and the Titanic dataset (Cukierski, 2012). The focus is on numerical and categorical datasets, utilizing a random forest classifier for real-time computation, enabling faster XAI methods calculation and thus a more natural, steady conversation. The system supports two types of local explanations: Shapley values and counterfactuals (see §2).

Figure 1 shows the architecture of the evaluated system. For integrating domain knowledge, we introduce a module providing suitable arguments. These arguments can be obtained either through manual acquisition from consulting domain experts or in case of widely studied topics by using automated procedures, e.g. large language models or semantic databases. To exemplify, we manually extracted arguments for the Titanic dataset from existing literature (Hall, 1986; Frey et al., 2011) (domain experts) and used ChatGPT[1] to generate arguments for the credit domain, which were then manually verified for accuracy. Additionally, each argument was manually annotated in order to align it with the desired argument tree structure and to provide a reference link to the AI features addressed in the argument[2]. However,research indicates that this process can also be automated in the future (Rach et al., 2021).

Since adding domain knowledge creates new user queries, we replaced the original keyword-based natural language understanding with a fine-tuned BERT model (Turc et al., 2019) to provide a more natural interaction. Feature values and names were replaced by placeholders, ensuring the model is not fine-tuned on specific scenarios but rather on explanatory dialogue, thus keeping the system generic. The data for training this model was manually generated with Chatito[3].

The rule-based dialogue interaction involves asking the user for feature-specific information and providing a prediction. The system then asks the user if they would like to receive an explanation of the prediction. We revise the explanation policy and introduce new interaction steps, by distinguishing between XAI explanations and domain knowledge explanations (see Figure 2). Instead of directly mapping user intentions to specific XAI explanations, we implement a more abstract intent for requesting an explanation, which determines the most suitable explanation based on the dialogue state and provided information. Initially, we offer Shapley explanations to give the user insight into the features impacting the outcome, presenting a simplified graph[4] of the values for lay users. We then provide additional information from domain knowledge, either based on the previous interaction (e.g., Shapley values, specific argument) or on requested feature values (see Table 1). When no suitable domain knowledge explanation is available, counterfactual explanations will be offered to maintain the dialogue's informative nature.

[1]GPT-3.5 https://openai.com/chatgpt/ Accessed: 2024-05-06
[2]The Argumentation Scheme can be seen in Appendix B
[3]https://github.com/rodrigopivi/Chatito Accessed: 2024-05-06
[4]A sample can be found in the Appendix A.

Group No.	AI	1. Dialogue DK	1. Dialogue Scenario	2. Dialogue DK	2. Dialogue Scenario	$\sum$
1	false	yes	credit	no	titanic	8
2	false	yes	titanic	no	credit	8
3	true	yes	credit	no	titanic	8
4	true	yes	titanic	no	credit	8

Table 2: Participant Distribution over Groups. Sum is the amount of participants per group.

Participants		Age		AI Attitude	
Total	32	Average	32.6	Median	4
Female	9	Youngest	22	Min	2
Male	23	Oldest	65	Max	5
Interaction Time		**Turns**		**AI Experience**	
Median	4.4 min	Median	26.5	Median	3
1. Dialogue	5.6 min	Min	16	Min	0
2. Dialogue	3.6 min	Max	167	Max	5

Table 3: Overall statistics of the conducted study. The AI Attitude was rated from 1 (negative) to 5 (positive). AI experience was rated from 0 (no experience) to 5 (expert).

4 Study Setup

To assess the initial impacts of integrated domain knowledge, we conduct a user study, presenting the dialogue system (§3) in a web environment (see Appendix A) with two distinct models trained on the Titanic and Credit datasets. In one scenario, the AI is trained on accurate training data (true AI), while in the other, the expected class is inverted in the training set to simulate a malfunctioning AI (false AI). This study setting has been designed to ascertain whether users can discern the false AI using authentic domain knowledge. We assess the user's impression of the AI using the questions: *I agree with the decisions made by the system* (Q1) and *The system decisions are plausible* (Q2). In addition to evaluate the overall performance of the system, we employed the SASSI questionnaire (Hone and Graham, 2000).

Each of the 32 participants interacted with the system twice, experiencing both scenarios (credit/ titanic) and one AI setting (true or false), with and without domain knowledge activated. This resulted in four groups based on AI truthfulness and scenario variance (see Table 2).

The study began with general instructions and a task description. Users were encouraged to interact and explore explanations, with the task designed to be open-ended for a natural conversation. After each scenario, the participants completed the questionnaire on a five-point Likert scale (Q1, Q2 and SASSI). Finally, we collected demographic information and participants' attitudes towards and experiences with AI (see Table 3). For evaluating the statistical relevance of our findings, we use the Mann-Whitney-U test (McKnight and Najab, 2010).

5 Evaluation

We discover notable differences in the interactions between the true and false AI setting, as shown in Table 4. Further, we observe a tendency for domain knowledge (DK) to support system decisions more effectively in the true AI setting. However, users did not engage with the domain knowledge as expected; it was requested in only 44% of the dialogues with domain knowledge activated. In addition, for the false AI, participants tended to rate the system's decisions more favourably when domain knowledge was requested, contrary to our initial expectations. We expected that with domain knowledge, users would be more likely to recognise the AI's errors, leading to lower ratings for the system's decisions. We assume this discrepancy is related to the questionnaire not focusing sufficiently on the AI predictions but rather on the overall system appearance. In future work, we will refine the methodologies employed in our user assessments with the objective of distinguishing between the underlying AI model decisions and the dialogue system.

The SASSI questionnaire[5] indicates that while the system's performance is respectable, there is still room for improvement. The results indicate that the system's speed is satisfactory and it is easy to use. However, there is a need for significant improvements in the accuracy of the system's responses. The inclusion of domain knowledge had a positive impact on the dialogue experience with false AI setting, particularly enhancing likeability and the consistency of the dialogue. Additionally, the availability of domain knowledge appeared to reduce the cognitive load on participants. For the true AI scenario, the system's usefulness was perceived to be higher when domain knowledge was incorporated. These findings suggest that domain knowledge not only improves the overall user experience in terms of dialogue consistency and likeability but also aids in reducing cognitive effort and enhancing the perceived utility of the system.

[5]The complete questionnaire evaluation can be found in Appendix C.

	AI	No DK avg	No DK $\sum$	DK avg	DK $\sum$	p
Q1	false	2.48	27	2.60	5	0.91
	true	3.69	23	3.89	9	0.87
Q2	false	2.44	27	3.40	5	0.14
	true	3.65	23	4.00	9	0.58

Table 4: Evaluation results comparing dialogues with requested domain knowledge (DK) and without (No DK). *AI* denotes the truthfulness of the underlying AI system. Q1 and Q2 are questions measuring if the user can understand the AI decisions (a higher value indicates greater consent). The sum shows the number of ratings and *p* is the value of the Mann-Whitney U test.

Finally, we collected overall statistics on the explanations provided, including the frequency of different types of explanation. This data provides valuable insights into how often each type of explanation was used during the interaction, helping us understand user preferences and the effectiveness of various explanatory strategies. On average, participants requested two explanations per dialogue. When domain knowledge was activated during a dialogue (in 44% of the possible dialogues), the system provided one additional explanation. Additionally, the counterfactual explanation was offered twice in a dialogue. Furthermore, in 32% of all dialogues, participants requested to change at least two values and discover other predictions. This indicates an attempt to discover the model's behavior through experimentation, which can be viewed as a form of example-based explanations.

These findings underline the importance of domain knowledge in explanatory dialogues and highlight both the system's strengths and areas for improvement, guiding future enhancements to better support user understanding and interaction. However, given the small sample size, these results only indicate trends. A more extensive evaluation with a larger participant pool is planned for the future to validate these findings more robustly.

6 Conclusion and Future Work

In this paper, we highlighted the need for domain knowledge integration in explanatory dialogue systems. Our approach employs argumentation structures to incorporate domain knowledge into explanatory dialogue systems, enhancing the transparency and comprehensibility of AI model explanations. By extending an existing explanatory dialogue system with domain knowledge, we demonstrate the practicality of our approach and conducted a study to evaluate the performance of this enhanced system.

While we observed the supportive role of domain knowledge in enhancing explanations in a way that users can more effectively evaluate model performance, several challenges remain. Enhancing interaction and optimizing the explanation policy are essential to ensure that users are capable to address their questions and receive the most relevant and comprehensive explanations to them, including alternative information such as feature descriptions. Additionally, improving the NLU component based on our observed explanation interaction patterns is crucial for facilitating more natural conversations.

Acknowledgements

This work has been funded by the DFG within the project "BEA - Building Engaging Argumentation", Grant no. 313723125, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Annalena Aicher, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2021. Opinion building based on the argumentative dialogue system bea. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 307–318. Springer.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Will Cukierski. 2012. Titanic-machine learning from disaster. *Kaggle. available at: https://kaggle.com/competitions/titanic*.

Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. 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. 2023. Towards interactive explanations of machine learning methods through dialogue systems. *The 13th International Workshop on Spoken Dialogue Systems Technolog*.

Bruno S Frey, David A Savage, and Benno Torgler. 2011. Who perished on the titanic? the importance of social norms. *Rationality and society*, 23(1):35–49.

Melita Hajdinjak and France Mihelič. 2004. Information-providing dialogue management. In *International Conference on Text, Speech and Dialogue*, pages 595–602. Springer.

Wayne Hall. 1986. Social class and survival on the ss titanic. *Social science & medicine*, 22(6):687–690.

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3-4):287–303.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

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

Andrea Pazienza, Stefano Ferilli, Floriana Esposito, S Bistarelli, and M Giacomin. 2017. Constructing and evaluating bipolar weighted argumentation frameworks for online debating systems. In *AI^3@ AI* IA*, pages 111–125.

Louise Phillips. 2011. *The promise of dialogue*. John Benjamins Publishing Company.

Niklas Rach, Carolin Schindler, Isabel Feustel, Johannes Daxenberger, Wolfgang Minker, and Stefan Ultes. 2021. From argument search to argumentative dialogue: A topic-independent approach to argument acquisition for dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 368–379, Singapore and 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 *Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, page 384–387, New York, NY, USA. Association for Computing Machinery.

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

Kacper Sokol and Peter Flach. 2020. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34(2):235–250.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2.

A Interface of proposed system

Figure 3 provides an overview of the user interface for our proposed dialogue system. It illustrates the layout, including the list of current feature values set by the user on the right side and a graph displaying the simplified Shapley values at the bottom. This visualization aims to give a clear understanding of how users interact with the system.

B Argumentation Scheme

Keyword	Description	Example
id	Assigned ID for an argument	gender_arg01
prev_node	Node the argument is pointing to. Can be an ID or empty if the argument is a claim.	gender_arg01
type	Type of the given argument.	CLAIM | SUPPORT
features	List of all related features to this argument	[gender]
text	Full text of the argument which will be presented to the user	Women were preferred for the lifeboats.

Table 5: Annotation scheme used for the retrieved arguments

C Additional Evaluation Information

Table 6 shows the full SASSI questionnaire.

		False AI			True AI					
		DK	No DK	p	DK	No DK	p	False AI	True AI	p
Model Consent	I agree with the decisions made by the system	2.60	2.48	0.9134	3.89	3.69	0.8783	2.50	3.75	**0.0000**
	The system decisions are plausible.	3.40	2.44	0.1422	4.00	3.65	0.5836	2.59	3.75	**0.0005**
System Response Acurracy	The system is accurate.	3.00	2.48	0.4638	3.33	2.78	0.2889	2.56	2.94	0.1514
	The system is unreliable.	3.40	3.37	0.9785	2.33	2.87	0.2665	3.37	2.72	**0.0246**
	The interaction with the system is unpredictable.	2.40	3.44	**0.0134**	2.55	2.74	0.7109	3.28	2.69	**0.0211**
	The system didn't always do what I wanted.	3.40	3.67	0.4401	3.67	3.65	0.8446	3.62	3.66	0.7369
	The system didn't always do what I expected.	3.80	3.70	0.8070	3.44	3.56	0.8103	3.72	3.53	0.5999
	The system is dependable.	2.60	2.63	1.0000	3.22	2.61	0.1986	2.62	2.78	0.5255
	The system makes few errors.	3.60	2.85	0.2637	2.33	3.43	0.0589	2.97	3.12	0.7005
	The interaction with the system is consistent.	4.20	2.81	**0.0064**	4.00	3.30	0.0896	3.03	3.50	0.0873
	The interaction with the system is efficient.	3.80	2.55	0.0511	2.89	3.04	0.8034	2.75	3.00	0.3587
	The system is useful.	2.40	2.41	0.8933	3.89	2.74	**0.0172**	2.41	3.06	**0.0293**
Likeability	The system is pleasant.	4.20	3.26	0.0846	3.66	3.39	0.6754	3.41	3.47	0.6071
	The system is friendly.	5.00	4.00	**0.0059**	4.44	4.48	0.8859	4.16	4.47	0.0867
	I was able to recover easily from errors.	4.00	2.44	**0.0329**	3.67	2.61	0.0684	2.69	2.91	0.5139
	I enjoyed using the system.	3.40	2.59	0.1511	3.44	2.83	0.2413	2.72	3.00	0.3313
	It is clear how to speak to the system.	4.20	2.52	**0.0136**	3.55	3.22	0.5055	2.78	3.31	0.1314
	It is easy to learn to use the system.	4.60	3.30	**0.0168**	4.44	3.91	0.3025	3.50	4.06	**0.0350**
	I would use this system.	2.00	2.22	0.9130	2.66	2.61	0.8799	2.19	2.62	0.1466
	I felt in control of the interaction with the system.	4.00	2.78	**0.0230**	3.55	2.87	0.1641	2.97	3.06	0.8572
Cognitive Demand	I felt confident using the system.	3.80	2.85	0.0684	4.11	3.04	**0.0262**	3.00	3.34	0.2274
	I felt tense using the system.	1.60	2.85	**0.0493**	1.33	1.83	0.1831	2.66	1.69	**0.0010**
	I felt calm using the system.	4.20	3.00	0.0695	4.00	3.35	0.2513	3.19	3.53	0.2438
	A high level of concentration is required when using the system.	2.40	2.85	0.3800	2.00	2.35	0.3081	2.78	2.25	0.0857
	The system is easy to use.	4.40	3.11	**0.0306**	4.00	3.52	0.3133	3.31	3.66	0.2420
Annoyance	The interaction with the system is repetitive.	3.20	3.41	0.6445	3.33	3.30	0.9820	3.37	3.31	0.8421
	The interaction with the system is boring.	2.00	3.11	**0.0473**	2.55	2.96	0.3923	2.94	2.84	0.7170
	The interaction with the system is irritating.	2.20	3.30	0.0927	2.22	3.17	0.0905	3.12	2.91	0.5313
	The interaction with the system is frustrating.	2.60	3.63	0.1216	2.55	3.09	0.3234	3.47	2.94	0.1649
	The system is too inflexible.	3.60	3.85	0.5525	3.11	3.48	0.4510	3.81	3.37	0.2796
Habitability	I sometimes wondered if I was using the right word.	3.40	3.89	0.2875	2.89	3.17	0.6189	3.81	3.09	**0.0167**
	I always knew what to say to the system.	3.00	2.26	0.1516	3.00	2.78	0.6813	2.37	2.84	0.1431
	I was not always sure what the system was doing.	2.80	3.85	0.0750	3.22	3.65	0.3512	3.69	3.53	0.5821
	It is easy to lose track of where you are in an interaction with the system.	1.40	2.93	**0.0109**	2.22	2.52	0.8113	2.69	2.44	0.4334
Speed	The interaction with the system is fast.	4.40	4.00	0.3036	4.22	4.69	**0.0483**	4.06	4.56	**0.0068**
	The system responds too slowly.	1.40	1.59	0.4505	1.11	1.17	0.6894	1.56	1.16	**0.0008**
Count		5	27		9	23		32	32	

Table 6: Results of the SASSI Questionnaire. Mean values for each category are presented, based on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The column 'p' shows the results of the Mann-Whitney-U test, with significant differences highlighted in bold.

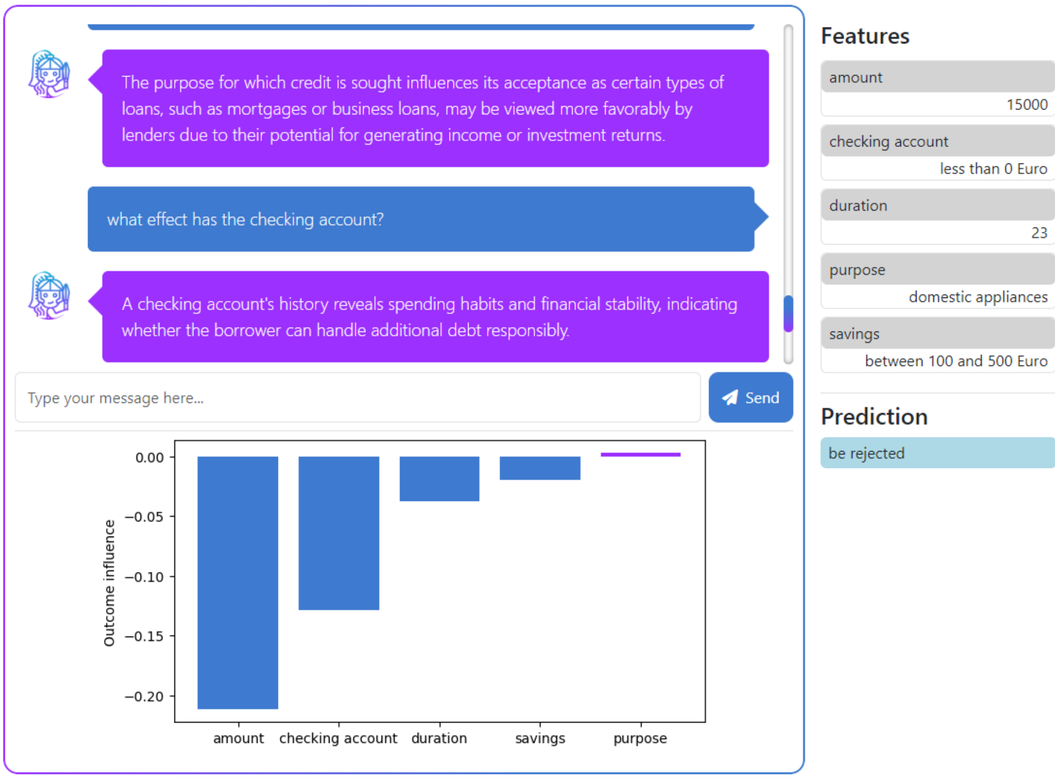

Figure 3: Proposed dialogue system chat interface.

D User study

Within this section we show the questions used for the demographic questionnaire items, as well as the introductory and task texts utilized in the study.

D.1 Demographic questionnaire

Here we show the questions regarding the experience with AI and the attitude towards AI and their options.

Do you have experience with Artificial Intelligence (AI)?

- **No Experience**
- **Novice** - Little to no understanding of AI concepts.
- **Beginner** - Familiar with some basic concepts but lack depth.
- **Intermediate** - Understand fundamental AI principles and their applications.
- **Advanced** - Deep understanding of AI concepts and can apply them practically.
- **Expert** - Have comprehensive knowledge of AI theories, methodologies, and apps.

What is your general attitude towards Artificial Intelligence (AI)?

- **Strongly negative** - Have deep reservations or fears about AI; believe it poses significant threats to society.
- **Somewhat negative** - Harbor concerns about AIś impact but acknowledge some potential benefits.
- **Neutral** - Neither strongly positive nor negative; see AI as a tool with both advantages and drawbacks.
- **Somewhat positive** - Optimistic about the potential benefits of AI but recognize the need for ethical considerations.
- **Strongly positive** - Enthusiastic about AIś potential to solve complex tasks; believe in its ability to drive positive change responsibly

D.2 General Instruction

Enabling conversational Explainable AI
Welcome to our online study, where we are evaluating the effectiveness of an explainable dialogue system.

In this study, you engage in two interactions with a dialogue system. The system is capable to help you access an artificial intelligence model. She will ask you for all necessary information the model needs to create a prediction. Further, she will tell you the prediction outcome and offer you explanations for it. After interacting with the dialogue system, you will be asked to answer questions about the interaction.

In the end, we kindly ask you to complete a demographic questionnaire. Your participation, taking approximately 30 minutes, will provide valuable insights into enhancing the transparency of machine learning models through the usage of dialogue systems. Thank you for your participation.

D.3 Task Description

In this study, you have the opportunity to engage with our dialogue system in a conversation about a predefined scenario. Your role as a participant is to engage in conversation with the dialogue system for as long as you wish. You are free to ask questions, explore various aspects of the prediction, and express your thoughts and concerns throughout the interaction. Although the interface will eventually prompt you to continue with the study, you are encouraged to chat for as long as you wish, allowing for a more comprehensive evaluation of the dialogue experience.

Credit Scenario In this session, we invite you to explore the process of applying for a credit loan and to consider whether you would be accepted by a bank for such a loan.

The dialogue system is here to assist you in this exploration. The system will guide you through a conversation about various aspects of your financial profile, asking for your input on relevant features such as income, credit history, and employment status. Using these details, the system will predict whether you would likely be approved or denied for a credit loan by a bank. Furthermore, system will offer explanations to help you understand the reasoning behind the outcome.

Titanic Scenario In this session, we invite you to explore the fateful journey of the Titanic and contemplate whether you would have survived the tragedy.

The system is here to assist you in this exploration. It will guide you through a conversation about various aspects of the Titanic disaster, asking for your input on relevant features such as age, gender, and passenger class. Using these details, the system will predict whether you would have survived the sinking of the Titanic or not. Furthermore, the system will offer explanations to help you understand the reasoning behind the outcome.

E Example Dialogue

Table 7 shows an example dialogue of our user study and illustrates various aspects of the system. The user interacted with the false AI setting, with domain knowledge enabled in the credit domain.

Turns 3 to 19 demonstrate the information retrieval process from the system. In Turn 6, the user requests an explanation for a specific feature information, triggering the domain knowledge. Although this was not the intended behavior, it led to a more natural conversation.

In Turn 8, the user seeks to understand the situation by asking a specific question, correctly classified as a request for explanation. However, since the system only considers the feature name, the response in Turn 9 was not well-suited for the context.

Turns 10 and 11 highlight some miscommunication and the limitations of template-based natural language generation (NLG).

After the prediction in Turn 19, the system provides Shapley values as an explanation, indicating that the *amount* was the most relevant feature. The user then asks for further details on why the amount was influential. The system provides domain knowledge in Turn 23. In Turn 24, the user notices a discrepancy between the original Shapley explanation and the domain knowledge explanation.

Turns 25 and 27 showcase the explanation policy. When the user denies the previous explanation, the system queries for clarification, thinking the user did not understand. In Turn 27, with no further arguments for the amount, the system offers a counterfactual explanation as an alternative.

Turns 37-41 show the user attempting to change a parameter value. However, the NLU misinterprets this as a request for a counterfactual explanation.

Turns 42-48 display the user changing a value and receiving a new prediction. The new explanations lead to user irritation in Turn 53, as the XAI explanation and domain knowledge information again do not address the same outcome.

Turn	Participant	Intent	Text
1	SYSTEM	opening credit	Hello there! I'm Athena, and I'm here to guide you through the world of AI models, providing insights, predictions and explanations tailored to your needs. Feel free to ask for help anytime – together, we'll unlock the power of AI and uncover the secrets hidden within your data! Let's shift our focus to the world of finance as we explore the process of applying for a credit loan, aiming to predict whether you would be accepted or rejected by a bank based on your financial profile. Are you ready to start?
2	USER	confirm	hello athena! what if i don't want to provide my financial data?
3	SYSTEM	start credit	Awesome! So, first of all, I need to get some information to determine if your credit application would be accepted. Don't worry, I will tell you exactly what information I need and what options are available. The first feature is checking account. Your options are: 1 - no checking account 2 - less than 0 Euro 3 - between 0 and 200 Euro 4 - more than 200 Euro
4	USER	provide feature	i have a checking account with more than 200€
5	SYSTEM	ask feature	Okay, thank you. The next feature is duration in months. Usual durations range from 4 to 72 months. Please provide the corresponding number.
6	USER	request explanation	what do you mean by duration?
7	SYSTEM	provide additional explanation	A longer credit duration typically leads to higher credit scores and demonstrates financial stability, enhancing eligibility for favorable loan offers by reducing perceived risk for lenders.
8	USER	request explanation	but this is about applying for a credit, isn't it? why would i need to know the duration beforehand?
9	SYSTEM	provide additional explanation	Credit duration, or credit history length, provides lenders with valuable data to assess a borrower's creditworthiness, including repayment behavior and credit utilization over time, which can influence loan terms.
10	USER	out of range	i have never needed a credit before, so i guess my answer would be 0?
11	SYSTEM	wrong entry	Sorry, your option is not a valid choice. Please give me a valid input to work with.
12	USER	provide feature	6
13	SYSTEM	ask feature	Okay, thank you. The next feature is purpose, i.e. what you want to use the credit for. Your options are: 1 - car (new) 2 - car (used) 3 - furniture/equipment 4 - radio/television 5 - domestic appliances 6 - repairs 7 - education 8 - retraining 9 - business 10 - others
14	USER	provide feature	7
15	SYSTEM	ask feature	Okay, thank you. The next feature is amount. The bank allows credits between 200 Euro and 20000 Euro. Please provide the corresponding number.
16	USER	provide feature	14680
17	SYSTEM	ask feature	Okay, thank you. The next feature is savings. Your options are: 1 - unknown/no savings account 2 - less than 100 Euro 3 - between 100 and 500 Euro 4 - between 500 and 1000 Euro 5 - more than 1000 Euro
18	USER	provide feature	5
19	SYSTEM	give prediction	Okay, let's have a look. My prediction is that a person with the values you provided would [u]**be accepted**[/u]. If you want a more detailed explanation, please tell me what you'd like to know. For example, you might ask **why** I decided the way I did or what would have **changed** the outcome.
20	USER	request explanation	why do you think i would have been accepted?
21	SYSTEM	compute shap	All right, let's have a look. In your case, the most influential value was **amount**, as you can see in the figure below. You can also see the influence of all other values, ordered by their importance. Purple bars indicate a positive contribution, blue bars a negative contribution. Is there anything else I can help you with?
22	USER	request explanation	why does a high credit amount increase my chances of being accepted?
23	SYSTEM	provide additional explanation	The amount of credit requested can affect acceptance as larger loan amounts may pose higher risks for lenders, requiring more stringent evaluation of the borrower's financial capacity to repay.
24	USER	deny	if that's true, shouldn't the blue and puirple bars be reversed?
25	SYSTEM	ask explain	Do you think some further explanations would help?
26	USER	confirm	yes
27	SYSTEM	offer cf	Okay, let me check.. Unfortunately, I don't have any further information for you on this point. Would you like me to generate a counterfactual explanation? This means I'll provide you with a scenario where changing certain inputs could result in a different outcome.
28	USER	confirm	yes please
29	SYSTEM	compute cf	Let me check that. So, based on my calculations you would need the following values to change the decision: **checking account:** no checking account Is there anything else you want to know?
30	USER	request explanation	what else could you tell me?
31	SYSTEM	offer cf	Unfortunately, I don't have any further information for you on this point. Would you like me to generate a counterfactual explanation? This means I'll provide you with a scenario where changing certain inputs could result in a different outcome.
32	USER	insult	fuck off
33	SYSTEM	easter egg	Great, now I have this mess in my training data. That's how AI systems get a bad reputation!
34	USER	ask help	i was just trying to see how you'd react, sorry.
35	SYSTEM	provide help	You can interact with me using natural language commands. You can ask for explanations by asking for an important feature or what needs to be changed to adjust the outcome. You can always ask for a general reason to get further insight. Besides that you can always change the provided information to get another prediction.
36	USER	provide feature name	savings 4
37	USER	request whatif	how can i change the outcomes
38	SYSTEM	compute cf	Let me check that. So, based on my calculations you would need the following values to change the decision: **checking account:** no checking account Is there anything else you want to know?
39	USER	request whatif	how can i change the parameters
40	SYSTEM	compute cf	Let me check that. So, based on my calculations you would need the following values to change the decision: **checking account:** no checking account Is there anything else you want to know?

Turn	Participant	Intent	Text
41	USER	deny	that does not answer my questio
42	SYSTEM	ask action	Alright. Is there anything else I can do for you?
43	USER	request change value	i want to change my provided information
44	SYSTEM	ask feature name	What feature do you want to change?
45	USER	provide feature name	duration
46	SYSTEM	ask feature Alright.	The next feature is duration in months. Usual durations range from 4 to 72 months. Please provide the corresponding number.
47	USER	provide feature	50
48	SYSTEM	give prediction	Okay, let's have a look. My prediction is that a person with the values you provided would [u]**be accepted**[/u]. If you want a more detailed explanation, please tell me what you'd like to know. For example, you might ask **why** I decided the way I did or what would have **changed** the outcome.
49	USER	request whatif	how could the outcome have changed?
50	SYSTEM	compute cf	Let me check that. So, based on my calculations you would need the following values to change the decision: **checking account:** no checking account Is there anything else you want to know?
51	USER	request explanation	why do savings contribute negatively?
52	SYSTEM	provide additional explanation	Having substantial savings can enhance credit acceptance as it demonstrates financial stability and provides a buffer for unexpected expenses, reassuring lenders about the borrower's ability to repay the loan.
53	USER	request explanation	but then why is the savings bar negative?!?!?!?!?!?!?!?!
54	SYSTEM	provide additional explanation	Savings can serve as a form of collateral or a down payment, reducing the loan-to-value ratio and perceived risk for lenders, thus increasing the likelihood of loan approval.
55	USER	request whatif	change the savings to option 4 and the checking account to no checking account
56	SYSTEM	compute cf	Let me check that. So, based on my calculations you would need the following values to change the decision: **amount:** 1943.3 Is there anything else you want to know?
57	USER	deny	no
58	SYSTEM	ask action	Alright. Is there anything else I can do for you?
59	USER	deny	no
60	SYSTEM	bye	Okay. Thank you for the nice discussion. Have a lovely day!

Table 7: Example of a full dialogue of the user study giving good insights on the overall system performance.