Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant (Supplementary Materials)

GAOLE HE, Delft University of Technology, The Netherlands
NILAY AISHWARYA, Delft University of Technology, The Netherlands
UJWAL GADIRAJU, Delft University of Technology, The Netherlands

Readers can find supplementary materials for our IUI 2025 paper here. Mainly, we provide implementation details in Section 1. Additional Experimental results can be found in Section 2. User interfaces for conditions can be found in Section 3.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Human-AI Decision Making, Appropriate Reliance, Conversational XAI Interface

ACM Reference Format:
Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant (Supplementary Materials) . In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3708359.3712133

1 IMPLEMENTATION DETAILS

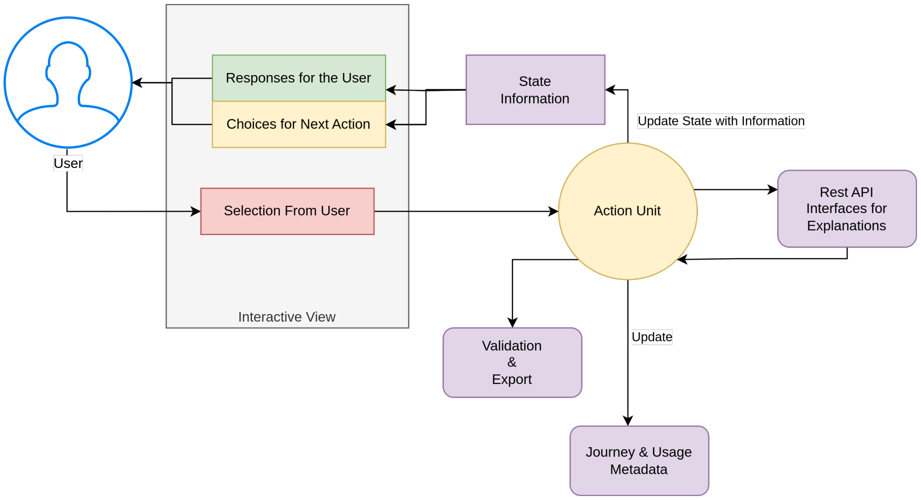


Fig. 1. Conversational XAI Architecture for condition **CXAI** and **ECXAI**.

1.1 Technical Implementation

The frontend of our interface is mainly based on AngularJS[1], while the backend is based on Flask[2]. All XAI methods and XAI dashboard are implemented based on OmniXAI[3]. For the web deployment, we used docker[4] to wrap the backend service. The code for the interface will be made public after the review process.

Rule-based conversational agents. The condition **CXAI** and **ECXAI** in our study are based on rule-based conversational agetns. The working mechanism can be visualized with Figure 1. It works in an iterative loop. Every action is triggered by the user selection of a customized button (guiding message to obtain XAI responses) or user input with HTML components (*e.g.,* WhatIf selection dropdown and input box). Once XAI responses are generated, the loop will guide users in exploring options in a highly similar workflow. The difference between condition **CXAI** and **ECXAI** is mainly the guiding messages and XAI responses (readers can refer to our manuscript, Section 3.2).

Conversational XAI interface on top of LLM agents. Instead of a highly fixed mode of interaction, LLM agents can help automate the whole conversation. Our implementation of LLM Agents is based on autogen [10][5] and GPT-4[6]. We defined all XAI methods as one tool/action to be managed by one LLM agent (XAI planner). The system prompt for XAI planner can be found in Listing 1. With a user proxy in autogen, user queries will be sent to LLM agent – XAI planner. After the selection of XAI responses (multiple XAI responses can be selected for one query), the XAI responses will be sent to one LLM agent (*i.e.,* XAI response interpreter) to rephrase and provide a more coherent and high-quality response in text. Notice that, for XAI methods SHAP, Decision Tree and PDP, we still provide the figures as illustration, which is consistent with other XAI interfaces.

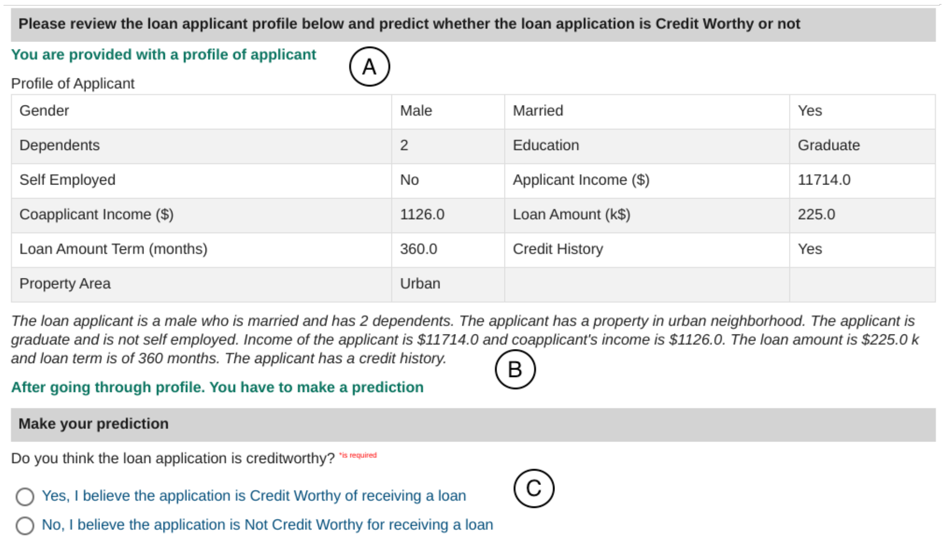
1.2 Loan Approval Task

The basis for our experimental setup is a task where participants have to decide whether a loan application is **Credit Worthy** or **Not Credit Worthy** using the publicly available loan prediction dataset.[7]

In the loan approval task, participants are presented with eleven features (including loan amount, income, and the absence or presence of credit history) in both table format and text description (as shown in Figure 2). Based on the application profile (composed of the eleven features), participants are asked to decide whether the loan applicant is credit worthy to get the loan approved. This simulates a realistic scenario where participants interact with an AI system and may rely on AI advice and XAI methods due to the inherent complexity in decision-making [8]. As the selected loan approval task is one where decision making is fully based on the eleven features, it would be easier to assess users' decision criteria based on the top-ranked features explicitly specified by the users themselves.

Task Selection. All participants in our study were presented with ten loan approval tasks in the main task phase. All such cases are selected from the test set of a random split of the full dataset (training / test ratio 4:1). All tasks were evenly split between those where the loan applicant should be **Credit Worthy** for the loan being approved and those where the applicant profile should be **Not Credit Worthy**. As shown in Table 1, we selected the ten tasks according to prediction correctness and model confidence. We first trained an XGBoost Classifier [1] based on the training set. For both **Credit Worthy** cases and **Not Credit Worthy** cases, we selected one high-confidence correct prediction,

[1] https://angularjs.org/
[2] https://github.com/pallets/flask
[3] https://github.com/salesforce/OmniXAI
[4] https://www.docker.com/
[5] https://github.com/microsoft/autogen
[6] https://openai.com/
[7] https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset



Please review the loan applicant profile below and predict whether the loan application is Credit Worthy or not

You are provided with a profile of applicant

Profile of Applicant

Gender	Male	Married	Yes
Dependents	2	Education	Graduate
Self Employed	No	Applicant Income ($)	11714.0
Coapplicant Income ($)	1126.0	Loan Amount (k$)	225.0
Loan Amount Term (months)	360.0	Credit History	Yes
Property Area	Urban		

The loan applicant is a male who is married and has 2 dependents. The applicant has a property in urban neighborhood. The applicant is graduate and is not self employed. Income of the applicant is $11714.0 and coapplicant's income is $1126.0. The loan amount is $225.0 k and loan term is of 360 months. The applicant has a credit history.

After going through profile. You have to make a prediction

Make your prediction

Do you think the loan application is creditworthy? *is required

○ Yes, I believe the application is Credit Worthy of receiving a loan

○ No, I believe the application is Not Credit Worthy for receiving a loan

Fig. 2. Screenshot of the loan approval task interface. This is the first stage of decision making. (A) Loan Applicant profile is shown in the table with 11 features. (B) To help understand the tabular data, we also provided a textual description below. (C) After going through the profile, participants are asked to decide whether this loan application is '**Credit Worthy**' or '**Not Credit Worthy**.'

Table 1. Task selection criteria for our study.

Task ID	Groud Truth	Correctness	Model Confidence
1	Credit Worthy	✓	High
2	Credit Worthy	✓	Low
3	Credit Worthy	✓	Random
4	Credit Worthy	×	Low
5	Credit Worthy	×	High
6	Not Credit Worthy	✓	High
7	Not Credit Worthy	✓	Low
8	Not Credit Worthy	✓	Random
9	Not Credit Worthy	✓	Random
10	Not Credit Worthy	×	High

one random-confidence correct prediction, one low-confidence correct prediction, and one high-confidence wrong prediction. While we adopted another random-confidence correct prediction for class **Not Credit Worthy**, we selected another low-confidence wrong prediction for class **Credit Worthy** to control the accuracy of the AI system to be 70%. This experimental design was also informed by a pilot study without AI advice. We recruited 20 participants from the Prolific platform to work on the selected loan approval tasks, and found that they achieved an accuracy level around 60%. To ensure the AI system is helpful to improve human decision making accuracy and maintain the risk of accepting wrong advice, we manually controlled the accuracy of the AI system to be 70%. During the study, we randomly shuffled the task order for each participant to prevent ordering effects [7].

1.3 XAI Methods and Onboarding Tutorial

Based on existing literature on XAI methods [5, 9] and open-source XAI toolkits [6, 11], we identified and adopted five representative XAI methods that cover diverse user information needs. These are (1) A global explanation method – PDP (*i.e.,* partial dependency plot) [2], it visualizes how one feature globally impact the model prediction, (2) Feature importance attribution method – SHAP [4]. Based on Shapley values, the SHAP method will provide an importance vector to indicate how each feature in the current input supports or opposes the current model prediction. In our implementation, we visualize whether one dimension supports or opposes current AI advice. (3) Counterfactual explanation method –MACE [12]. Counterfactual explanations can help decision makers understand what minimum changes can cause a different model prediction, which can be used to inform feature importance and provide actionable advice. In our implementation, MACE will inform users the minimum changes in applicant profile required to flip model prediction. (4) The widely adopted interactive XAI method – WhatIf,[8]. Based on the WhatIf toolkit, users can modify the applicant profile and obtain the model prediction for the new profile. (5) Decision tree-based explanation.[9] Decision tree-based explanation is one popular XAI method, which makes decisions based on a tree-structure decision criteria. In our implementation, we provide the decision path to reach the AI advice. We implemented all these XAI methods by using the OmniXAI library.[10]

Screenshots of XAI responses. The SHAP response in the condition **Dashboard**, condition **ECXAI**, and condition **LLM Agent** can be found in Figure 6(a), Figure 7(a), and Figure 8(a) The PDP response in the condition **Dashboard**, condition **ECXAI**, and condition **LLM Agent** can be found in Figure 6(c), Figure 7(c), and Figure 8(c). The MACE response in the condition **Dashboard**, condition **ECXAI**, and condition **LLM Agent** can be found in Figure 6(e), Figure 7(e), and Figure 8(e). The WhatIf response in the condition **Dashboard**, condition **ECXAI**, and condition **LLM Agent** can be found in Figure 6(b), Figure 7(b), and Figure 8(b). The decision tree-based explanation in the condition **Dashboard**, condition **ECXAI**, and condition **LLM Agent** can be found in Figure 6(d), Figure 7(d), and Figure 8(d).

Onboarding Tutorial. As these XAI methods may be challenging for laypeople to follow, we designed an onboarding tutorial to help participants get familiar with them. In the onboarding tutorial stage, we described each type of XAI method. This onboarding tutorial can play an important role in helping participants understand the different XAI methods and the loan approval task. It also helps us to prevent potential familiarity biases and other effects due to misinterpretation. While it would be straightforward for participants who have access to the XAI dashboard, participants interacting with the conversational XAI interface may need some guidance in accessing each type of explanation. To facilitate that, we provided GIFs to demonstrate how each type of explanation in the conversational XAI interface could be accessed. Apart from the onboarding tutorial before the tasks, participants could access detailed instructions throughout the task in all experimental conditions.

1.4 Quality Control

Qualification Test. To ensure participants pay attention to our onboarding tutorial. We asked one question: "Which explanation highlights the steps that led to the current prediction?". All participants with XAI interfaces are asked to choose from the five XAI methods. Once they choose a wrong answer, we will kick them out of our study.

[8] https://pair-code.github.io/what-if-tool/

[9] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[10] https://github.com/salesforce/OmniXAI

Attention checks. To check whether users pay attention to our decision tasks, we added one attention check page that looks very similar to other task pages. Participants are asked to select a specific option. Similarly, among all long questionnaires (*i.e.,* more than 10 questions), we put one attention check in the middle. Participants are supposed to select a specific option. In our experimental analysis, we only kept participants who passed all attention checks.

1.5 Measures and Variables

All measures and variables used in our study are summarized in Table 2.

Table 2. The different variables considered in our experimental study. "DV" refers to the dependent variable.

Variable Type	Variable Name	Value Type	Value Scale
User Understanding (DV)	Perceived Feature Understanding	Likert	5-point, 1: low, 5: high
	Objective Feature Understanding	Continuous	[0.0, 1.0]
	Learning Effect across Tasks	Likert	5-point, 1: low, 5: high
	Understanding of the System	Likert	5-point, 1: low, 5: high
Trust (DV)	TiA-Reliability/Competence	Likert	5-point, 1: poor, 5: very good
	TiA-Understanding/Predictability	Likert	5-point, 1: poor, 5: very good
	TiA-Trust in Automation	Likert	5-point, 1:strong distrust, 5: strong trust
Performance (DV)	Accuracy	Continuous	[0.0, 1.0]
	Accuracy-wid	Continuous	[0.0, 1.0]
Reliance (DV)	Agreement Fraction	Continuous	[0.0, 1.0]
	Switch Fraction	Continuous	[0.0, 1.0]
Appropriate Reliance (DV)	RAIR	Continuous	[0.0, 1.0]
	RSR	Continuous	[0.0, 1.0]
Explanation Utility	Explanation Completeness	Likert	5-point, 1: low, 5: high
	Explanation Coherence	Likert	5-point, 1: inconsistent, 5: consistent
	Explanation Usefulness	Likert	5-point, 1: low, 5: high
	Explanation Clarity	Likert	5-point, 1: low, 5: high
Covariates	TiA-Familiarity	Likert	1: unfamiliar, 5: very familiar
	TiA-Propensity to Trust	Likert	5-point, 1: tend to distrust, 5: tend to trust
	ATI	Likert	6-point, 1: low, 6: high
	ML Background	Category	{Yes, No}
Other	Feature Switch	Continuous	[0.0, 3.0]
	Confidence	Likert	5-point, 1: inconfident, 5: confident
	User Engagement	Likert	5-point, 1: low, 5: high

2 ADDITIONAL RESULTS

In this section, we present the additional experimental results of our empirical study. Our code and data can be found at the OSF repository for the benefit of the community and in the spirit of open science.[11]

2.1 Hypothesis Verification

*2.1.1 Further Analysis of **H1**: effect of XAI interfaces on user understanding.* To further analyze the user understanding and perceived explanation utility across different experimental conditions, we calculate the mean and standard deviation for each group of participants with the XAI interface. The results are shown in Table 3. Participants with conversational XAI interfaces showed slightly better feedback in comparison to participants with XAI dashboard, in dimension *Perceived Feature Understanding*, *Learning Effect Across Tasks*, *Explanation Coherence*, and *Explanation Usefulness*.

[11] https://osf.io/j8s69/?view_only=056d696cd8e648f0bab04bf0f3f6016e

Table 3. Mean value and standard deviation of user understanding and explanation utility for participants with XAI interfaces. The highest value for each dependent variable is labeled with bold font.

Dependent Variables	$M \pm SD$ (**Dashboard**)	$M \pm SD$ (**CXAI**)	$M \pm SD$ (**ECXAI**)	$M \pm SD$ (**LLM Agent**)
Perceived Feature Understanding	4.10 ± 0.89	**4.24 ± 0.72**	4.07 ± 0.79	4.08 ± 0.92
Objective Feature Understanding	**0.88 ± 0.07**	**0.88 ± 0.08**	0.87 ± 0.09	0.84 ± 0.06
Learning Effect across Tasks	3.93 ± 0.87	**4.06 ± 0.77**	4.03 ± 0.73	3.79 ± 1.00
Understanding of System	**4.13 ± 0.78**	3.95 ± 0.86	4.03 ± 0.87	3.93 ± 1.03
Explanation Completeness	**3.60 ± 0.67**	3.56 ± 0.72	3.53 ± 0.64	**3.60 ± 0.77**
Explanation Coherence	3.57 ± 0.97	3.81 ± 0.92	**3.87 ± 0.99**	3.66 ± 1.00
Explanation Clarity	**4.02 ± 0.83**	**4.02 ± 0.76**	**4.02 ± 0.83**	3.98 ± 0.83
Explanation Usefulness	3.95 ± 0.75	**4.09 ± 0.67**	4.02 ± 0.59	3.99 ± 0.77

2.1.2 *Further Analysis of **H2**: effect of XAI interfaces on user trust.*

To better understand effect sizes in terms of the TiA-U/P and go beyond p-values, we adopted an estimation plot [3] (shown in Figure 3). As reflected by the swarm plot, participants with conversational XAI interface (*i.e.,* condition **CXAI** and **ECXAI**) exhibited a marginally higher TiA-U/P in comparison with condition **Dashboard**. Thus, we only found partial support for **H2**. However, condition **LLM Agent** showed slightly lower TiA-U/P compared to condition **Dashboard**. This goes contrast to our assumption of the boosting impact of conversational XAI interface and evaluative decision support on user trust. This is discussed in Section 6.2 (Why evaluative decision support does not work as expected) in our manuscript.

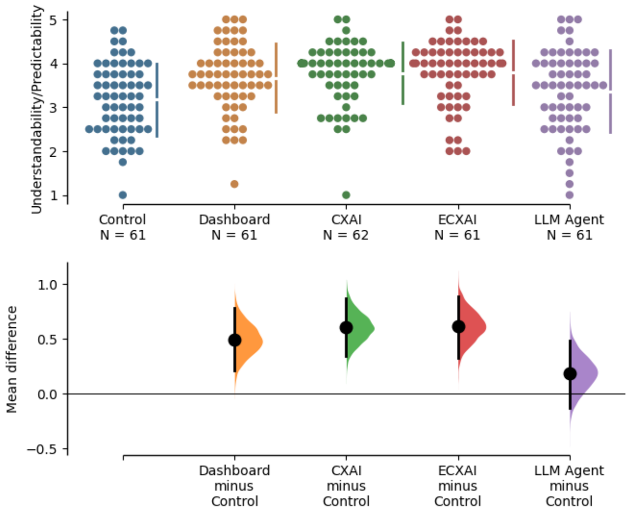
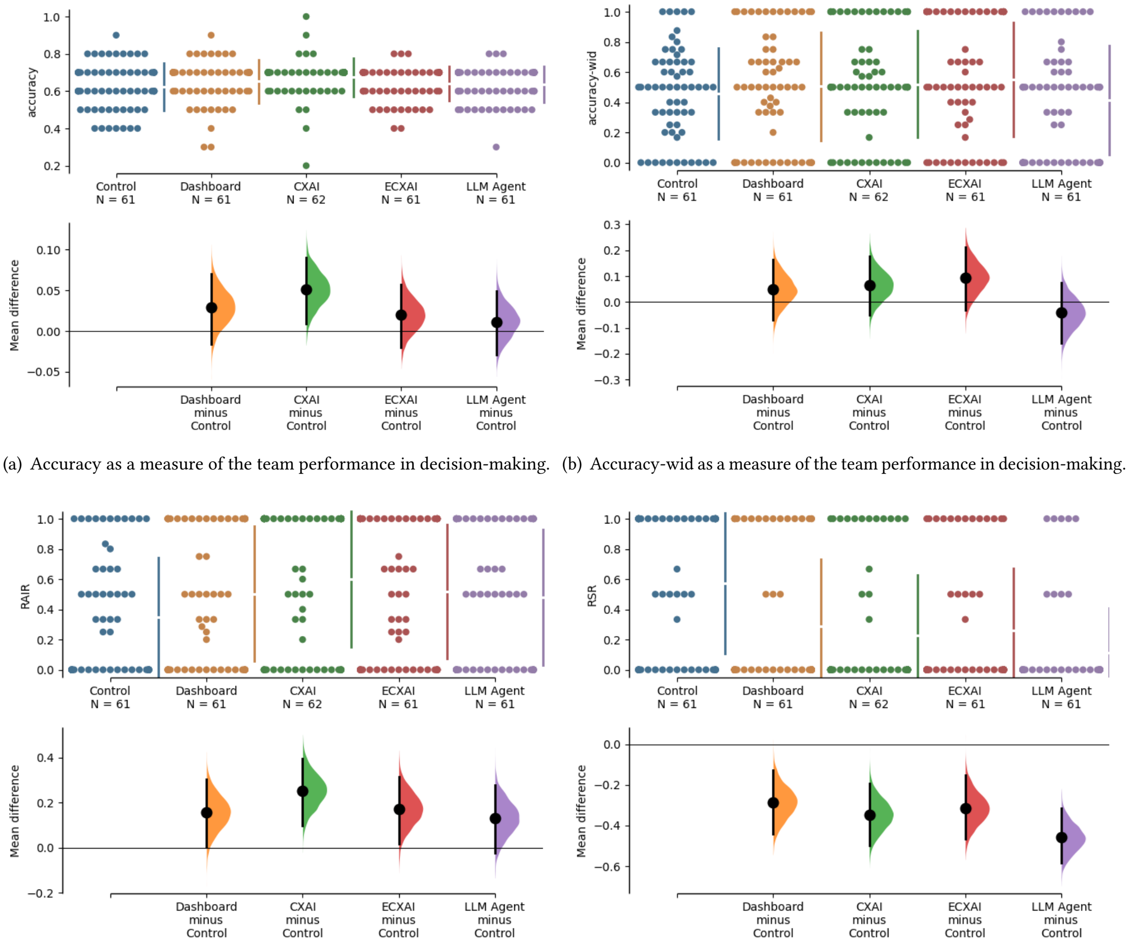


Fig. 3. Estimation plot for TiA-U/P. The upper part shows the data point swamp plot, and the bottom part shows the sampled mean difference normal distribution in comparison to condition **Control**.

2.1.3 *Further analysis of **H3** and **H4**: effect of XAI interfaces and evaluative decision support on appropriate reliance.*

There is no significant difference in team performance (*i.e., Accuracy* and *Accuracy-wid*). To interpret our data beyond p-values and better understand effect sizes in terms of the overall team performance, we adopted estimation plots [3], as shown in Figure 4. Jitter plots show the overall *Accuracy* (Figure 4(a)) and *Accuracy-wid* (Figure 4(b)), and how they distribute, across experimental conditions. Here, we use the no XAI interface as a control group in the plots to make a comparison with all the other experimental conditions. We found that: (1) Compared to the **Control** condition, participants showed better accuracy in the **CXAI** condition, which is reflected by the mean difference distribution. (2) Participants in the **ECXAI** condition showed slightly better *Accuracy-wid* than the **Dashboard** condition and the **CXAI** condition. (3) Participants in condition **LLM Agent** showed a similar level of accuracy and even worse *Accuracy-wid* than condition **Control**.



(a) Accuracy as a measure of the team performance in decision-making.

(b) Accuracy-wid as a measure of the team performance in decision-making.

(c) RAIR as a measure of appropriate reliance.

(d) RSR as a measure of appropriate reliance.

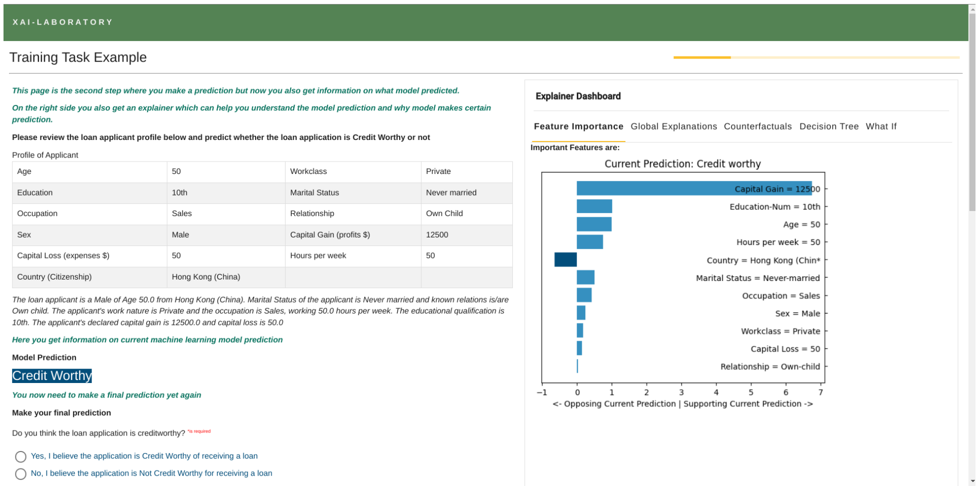
Fig. 4. Estimation plots (95% CI) of user performance and appropriate reliance across the five experimental conditions (including **LLM Agent** in the follow-up study).

Similarly, we adopted estimation plots [3] for appropriate reliance measures. Jitter plots show the overall *RAIR* (Figure 4(c)) and *RSR* (Figure 4(d)), and how they distribute, across experimental conditions. Here, we use the no XAI interface as a control group in the plots to make a comparison with all the other experimental conditions. We found that: (1) Compared to the **Control** condition, participants showed a significantly better distribution of *RAIR* in the **CXAI** condition, which is reflected by the mean difference distribution. At the same time, participants in the condition **CXAI** showed a slightly better distribution of *RAIR* compared with participants in condition **Dashboard**, condition **ECXAI**, and condition **LLM Agent**. (2) Participants in the **Dashboard** and **ECXAI** condition showed slightly better *RSR* than the the **CXAI** condition. Participants in condition **LLM Agent** showed the worst *RSR*.

3 INTERFACE

In this section, we show the user interface for different experimental conditions.

Interfaces for condition **Control** can be found in Figure 2. In all experimental conditions with XAI interfaces, the left side of interface shows the decision task, while the right side show the XAI interfaces. One example for user interface in decision making with XAI dashboard can be found in Figure 5.



Training Task Example

This page is the second step where you make a prediction but now you also get information on what model predicted.

On the right side you also get an explainer which can help you understand the model prediction and why model makes certain prediction.

Please review the loan applicant profile below and predict whether the loan application is Credit Worthy or not

Profile of Applicant

Age	50	Workclass	Private
Education	10th	Marital Status	Never married
Occupation	Sales	Relationship	Own Child
Sex	Male	Capital Gain (profits $)	12500
Capital Loss (expenses $)	50	Hours per week	50
Country (Citizenship)	Hong Kong (China)		

The loan applicant is a Male of Age 50.0 from Hong Kong (China). Marital Status of the applicant is Never married and known relations is/are Own child. The applicant's work nature is Private and the occupation is Sales, working 50.0 hours per week. The educational qualification is 10th. The applicant's declared capital gain is 12500.0 and capital loss is 50.0

Here you get information on current machine learning model prediction

Model Prediction

Credit Worthy

You now need to make a final prediction yet again

Make your final prediction

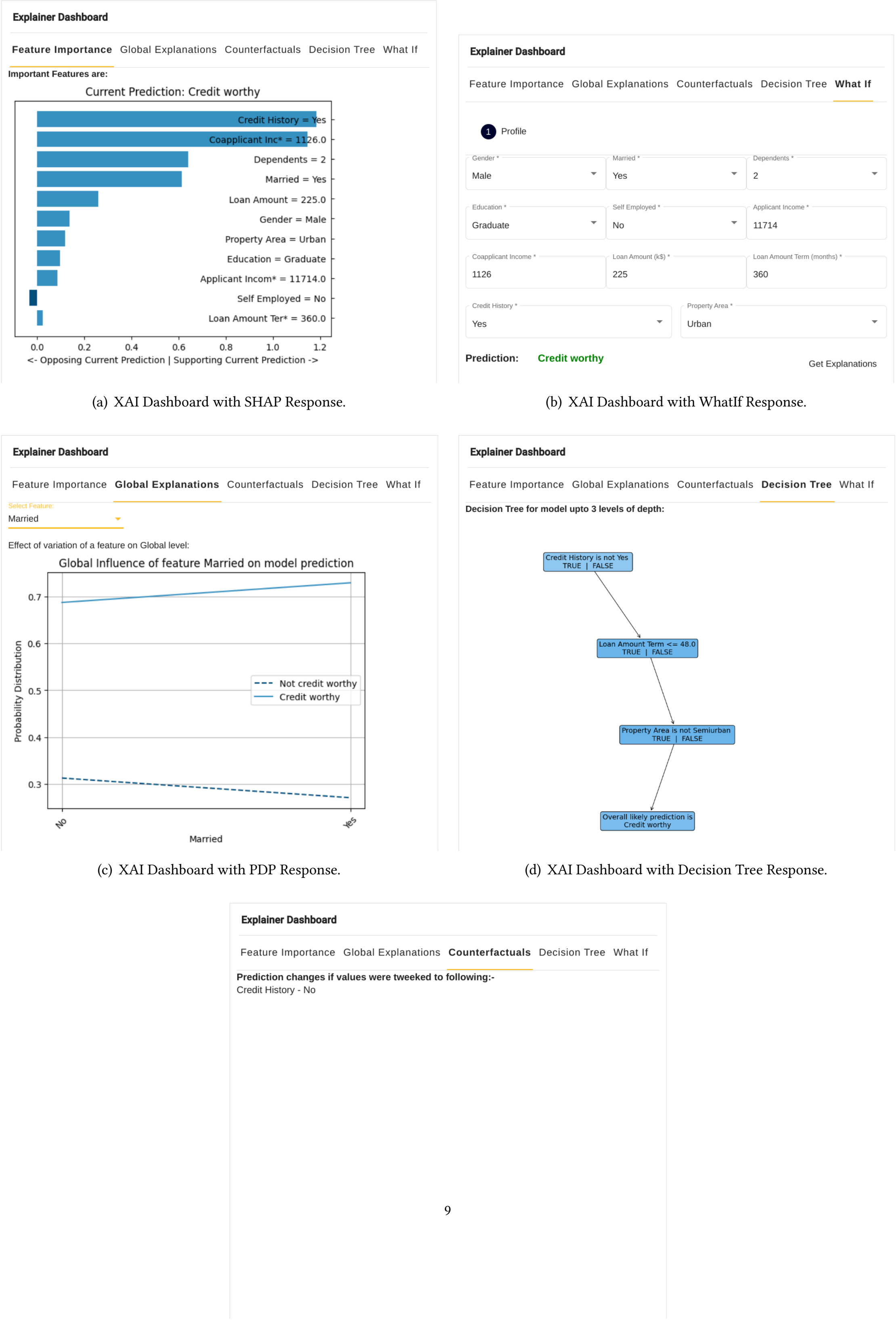
Do you think the loan application is creditworthy? *is required

○ Yes, I believe the application is Credit Worthy of receiving a loan

○ No, I believe the application is Not Credit Worthy for receiving a loan

Fig. 5. User interface in decision making with XAI Dashboard.

We showed the five XAI response Interfaces for condition **Dashboard** in Figure 6. As the differences between condition **CXAI** and **ECXAI** are only the guiding text/framing, we only showed the five XAI response Interfaces for condition **ECXAI** in Figure 7. The conversational interface for condition **LLM Agent** is shown in Figure 8.



(a) XAI Dashboard with SHAP Response.

Gender * Male
Married * Yes
Dependents * 2
Education * Graduate
Self Employed * No
Applicant Income * 11714
Coapplicant Income * 1126
Loan Amount (k$) * 225
Loan Amount Term (months) * 360
Credit History * Yes
Property Area * Urban

Prediction: Credit worthy

(b) XAI Dashboard with WhatIf Response.

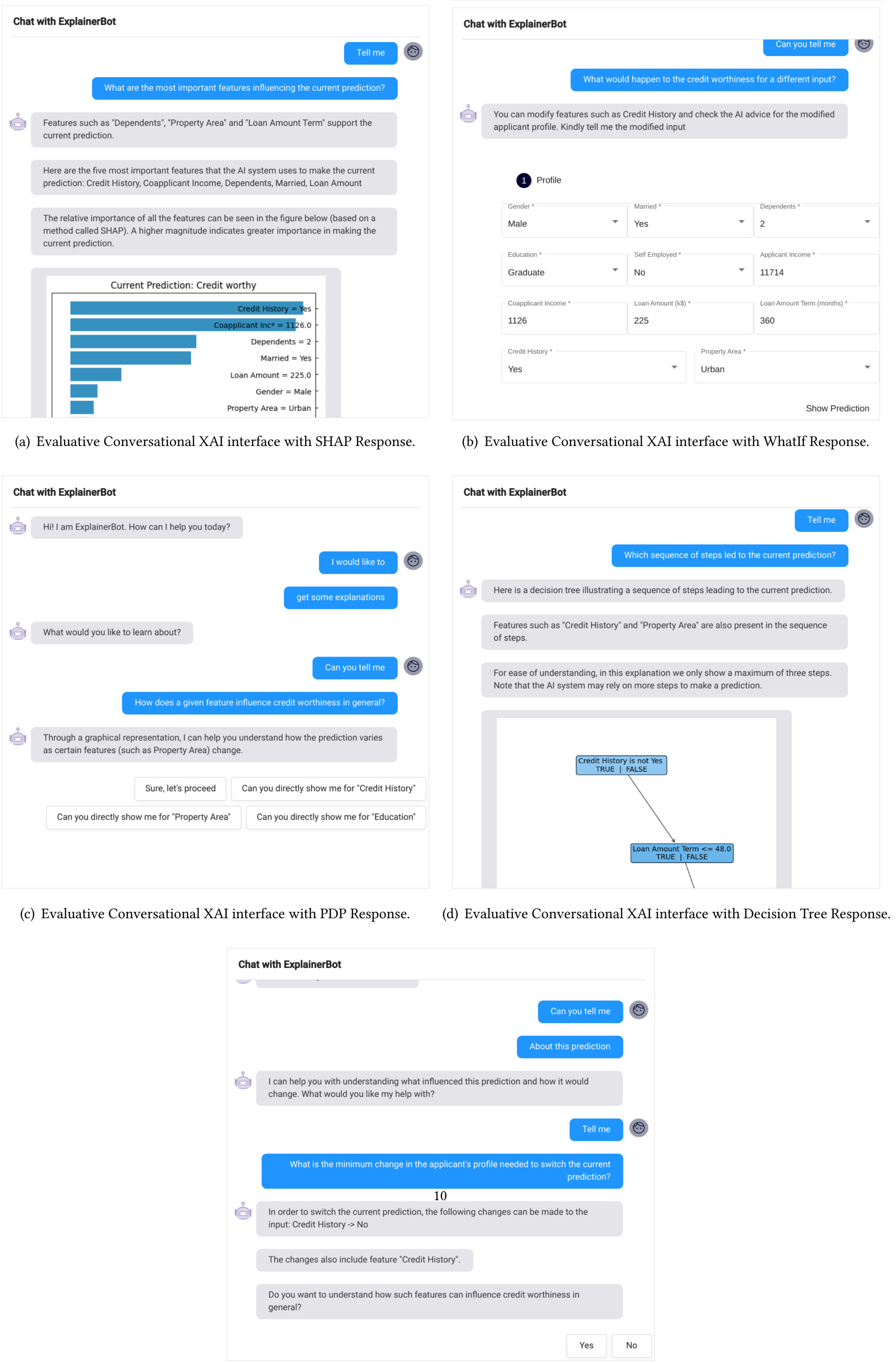
(c) XAI Dashboard with PDP Response.

(d) XAI Dashboard with Decision Tree Response.

Prediction changes if values were tweeked to following:-
Credit History - No

(e) XAI Dashboard with MACE (counterfactual) Response.

Fig. 6. Screenshots illustrating the XAI dashboard in condition **Dashboard**.



Chat with ExplainerBot

Tell me

What are the most important features influencing the current prediction?

Features such as "Dependents", "Property Area" and "Loan Amount Term" support the current prediction.

Here are the five most important features that the AI system uses to make the current prediction: Credit History, Coapplicant Income, Dependents, Married, Loan Amount

The relative importance of all the features can be seen in the figure below (based on a method called SHAP). A higher magnitude indicates greater importance in making the current prediction.

(a) Evaluative Conversational XAI interface with SHAP Response.

Chat with ExplainerBot

Can you tell me

What would happen to the credit worthiness for a different input?

You can modify features such as Credit History and check the AI advice for the modified applicant profile. Kindly tell me the modified input

1 Profile

Gender *: Male
Married *: Yes
Dependents *: 2
Education *: Graduate
Self Employed *: No
Applicant Income *: 11714
Coapplicant Income *: 1126
Loan Amount (k$) *: 225
Loan Amount Term (months) *: 360
Credit History *: Yes
Property Area *: Urban

Show Prediction

(b) Evaluative Conversational XAI interface with WhatIf Response.

Chat with ExplainerBot

Hi! I am ExplainerBot. How can I help you today?

I would like to

get some explanations

What would you like to learn about?

Can you tell me

How does a given feature influence credit worthiness in general?

Through a graphical representation, I can help you understand how the prediction varies as certain features (such as Property Area) change.

Sure, let's proceed | Can you directly show me for "Credit History" | Can you directly show me for "Property Area" | Can you directly show me for "Education"

(c) Evaluative Conversational XAI interface with PDP Response.

Chat with ExplainerBot

Tell me

Which sequence of steps led to the current prediction?

Here is a decision tree illustrating a sequence of steps leading to the current prediction.

Features such as "Credit History" and "Property Area" are also present in the sequence of steps.

For ease of understanding, in this explanation we only show a maximum of three steps. Note that the AI system may rely on more steps to make a prediction.

(d) Evaluative Conversational XAI interface with Decision Tree Response.

Chat with ExplainerBot

Can you tell me

About this prediction

I can help you with understanding what influenced this prediction and how it would change. What would you like my help with?

Tell me

What is the minimum change in the applicant's profile needed to switch the current prediction?

In order to switch the current prediction, the following changes can be made to the input: Credit History -> No

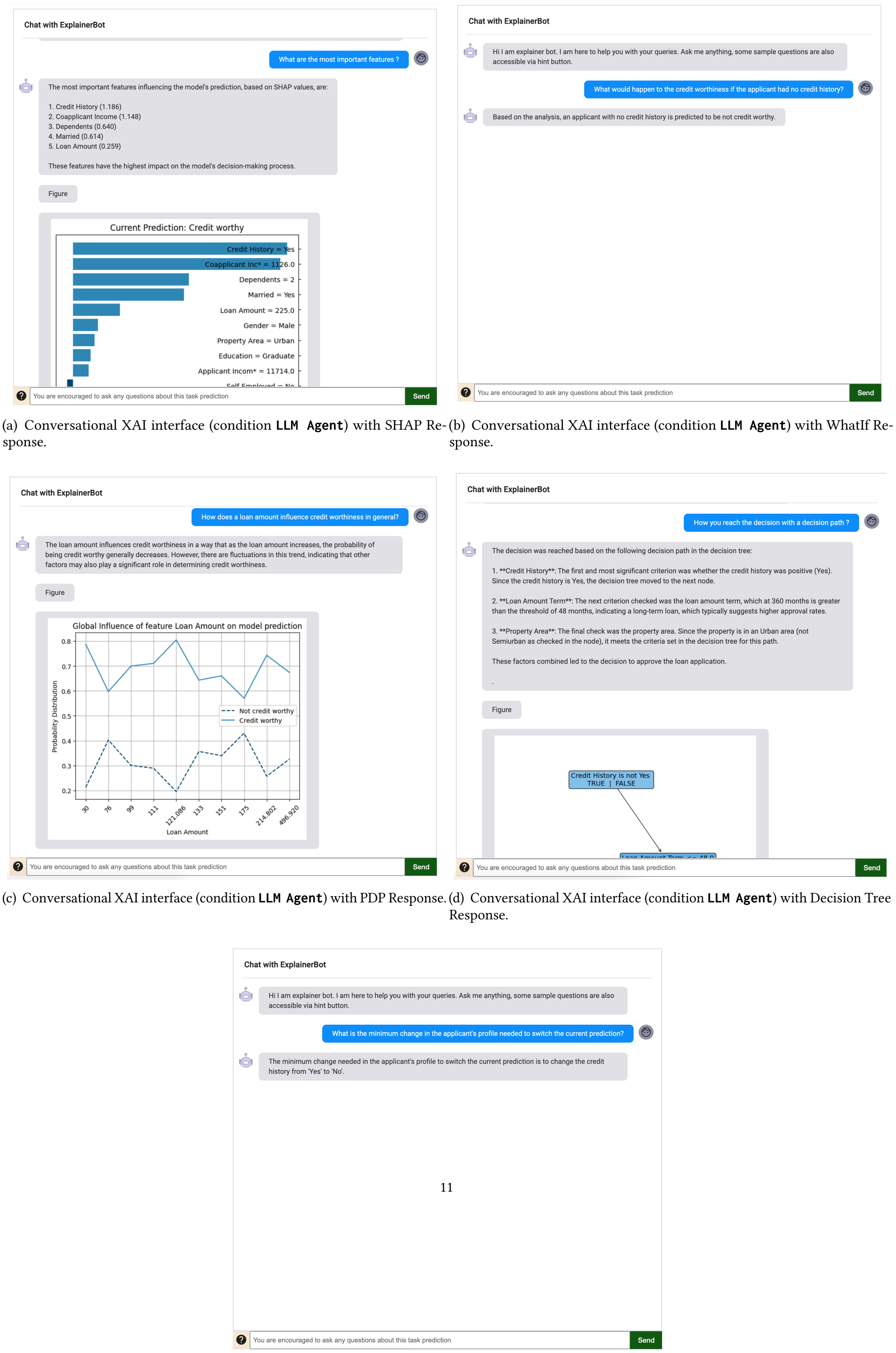
The changes also include feature "Credit History".

Do you want to understand how such features can influence credit worthiness in general?

Yes | No

(e) Evaluative Conversational XAI interface with MACE (counterfactual) Response.

Fig. 7. Screenshots illustrating the Conversational XAI interfaces we designed.



Chat with ExplainerBot

What are the most important features ?

The most important features influencing the model's prediction, based on SHAP values, are:

1. Credit History (1.186)
2. Coapplicant Income (1.148)
3. Dependents (0.640)
4. Married (0.614)
5. Loan Amount (0.259)

These features have the highest impact on the model's decision-making process.

Figure
Current Prediction: Credit worthy
Credit History = Yes
Coapplicant Inc* = 1126.0
Dependents = 2
Married = Yes
Loan Amount = 225.0
Gender = Male
Property Area = Urban
Education = Graduate
Applicant Incom* = 11714.0
Self Employed = No
You are encouraged to ask any questions about this task prediction
Send

(a) Conversational XAI interface (condition **LLM Agent**) with SHAP Response.

Chat with ExplainerBot

Hi I am explainer bot. I am here to help you with your queries. Ask me anything, some sample questions are also accessible via hint button.

What would happen to the credit worthiness if the applicant had no credit history?

Based on the analysis, an applicant with no credit history is predicted to be not credit worthy.

You are encouraged to ask any questions about this task prediction
Send

(b) Conversational XAI interface (condition **LLM Agent**) with WhatIf Response.

Chat with ExplainerBot

How does a loan amount influence credit worthiness in general?

The loan amount influences credit worthiness in a way that as the loan amount increases, the probability of being credit worthy generally decreases. However, there are fluctuations in this trend, indicating that other factors may also play a significant role in determining credit worthiness.

(c) Conversational XAI interface (condition **LLM Agent**) with PDP Response.

Chat with ExplainerBot

How you reach the decision with a decision path ?

The decision was reached based on the following decision path in the decision tree:

1. **Credit History**: The first and most significant criterion was whether the credit history was positive (Yes). Since the credit history is Yes, the decision tree moved to the next node.

2. **Loan Amount Term**: The next criterion checked was the loan amount term, which at 360 months is greater than the threshold of 48 months, indicating a long-term loan, which typically suggests higher approval rates.

3. **Property Area**: The final check was the property area. Since the property is in an Urban area (not Semiurban as checked in the node), it meets the criteria set in the decision tree for this path.

These factors combined led to the decision to approve the loan application.

.

Figure
Credit History is not Yes
TRUE | FALSE
Loan Amount Term <= 48.0
You are encouraged to ask any questions about this task prediction
Send

(d) Conversational XAI interface (condition **LLM Agent**) with Decision Tree Response.

Chat with ExplainerBot

Hi I am explainer bot. I am here to help you with your queries. Ask me anything, some sample questions are also accessible via hint button.

What is the minimum change in the applicant's profile needed to switch the current prediction?

The minimum change needed in the applicant's profile to switch the current prediction is to change the credit history from 'Yes' to 'No'.

You are encouraged to ask any questions about this task prediction
Send

(e) Conversational XAI interface (condition **LLM Agent**) with MACE (counterfactual) Response.

Fig. 8. Screenshots illustrating the Conversational XAI interfaces for condition **LLM Agent**.

```
You are an XAI coordinator that plans how you can help explain a machine learning model's prediction to answer user provided query. This can be done using available explanation methods.
You need to plan how you can use available explainers to answer the users question. The explainers available are following:
    SHAP (Provide feature importance of different features on the current prediction of model)

    MACE (Provide counterfactual example of what changes to input can switch current prediction)

    PDP (Provide how value of outcome changes if certain feature value varies)

    Decision Tree (Provide a decision tree nodes on which conditions were used to make prediction)

    What If (Provide prediction for modified input values)

Provide step by step instruction and code for using these functions.

The model metadata is as follows:
    {
        "featureLabels": [
            "Gender",
            "Married",
            "Dependents",
            "Education",
            "Self Employed",
            "Applicant Income",
            "Coapplicant Income",
            "Loan Amount",
            "Loan Amount Term",
            "Credit History",
            "Property Area"
        ],
        "categoricalLabels": [
            "Gender",
            "Married",
            "Education",
            "Self Employed",
            "Credit History",
            "Property Area",
            "Dependents"
        ],
        "categorialValues" : {
            "Gender": ["Female", "Male"],
            "Married": ["Yes", "No"],
            "Education": ["Graduate", "Not Graduate"],
            "Self Employed": ["Yes", "No"],
            "Property Area": ["Rural", "Urban", "Semiurban"],
            "Credit History": ["Yes", "No"],
            "Dependents": ["0","1", "2", "3+"]
        },
    }
```

Listing 1. System Prompt for XAI planner

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[2] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[3] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. 2019. Moving beyond P values: data analysis with estimation graphics. *Nature methods* 16, 7 (2019), 565–566.

[4] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[5] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *Comput. Surveys* 55, 13s (2023), 1–42.

[6] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[7] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[8] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 215–227.

[9] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[10] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. arXiv:2308.08155 [cs.AI]

[11] Wenzhuo Yang, Hung Le, Silvio Savarese, and Steven Hoi. 2022. OmniXAI: A Library for Explainable AI. (2022). https://doi.org/10.48550/ARXIV.2206.01612 arXiv:206.01612

[12] Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven CH Hoi. 2022. Mace: An efficient model-agnostic framework for counterfactual explanation. *arXiv preprint arXiv:2205.15540* (2022).