

SIMXAI: A Simple Drop-in Layer Increases Interaction Quality and Understanding in Conversational XAI

Van Bach Nguyen¹ Qianli Wang^{2,4}
Nils Feldhus^{3,4,†} Jörg Schlötterer¹ Christin Seifert¹

¹ University of Marburg ² Technische Universität Berlin ³ University of Groningen
⁴ German Research Center for Artificial Intelligence (DFKI)

Abstract

Conversational explainable AI (ConvXAI) agents often use GUS-style architectures with separate modules for intent recognition, dialogue management, and template-based response generation. These pipelines are hard to adapt and often struggle with context-dependent follow-up questions. We present SIMXAI, a lightweight LLM-centric conversational layer that replaces these modules with a single tool-calling LLM that bridges access to existing XAI resources. Rather than a standalone agent, SIMXAI serves as a modular interface that enables conversational access to external XAI resources. We further introduce an evaluation suite for single- and multi-turn parsing, response-to-tool faithfulness, and contextualization. Across existing and new datasets, SIMXAI achieves strong function-selection accuracy and high response-to-tool faithfulness while better handling context-dependent requests. In a user study replacing only the conversational layer of TALKToMODEL, a representative GUS-based agent, users completed tasks more accurately and in fewer turns. Overall, SIMXAI improves conversational interaction while leaving the underlying XAI resources unchanged¹.

1 Introduction

The XAI community has reached a consensus that explanations should take a conversational form (Miller, 2019; Lakkaraju et al., 2022; Zhang et al., 2025). Conversational explainable AI (ConvXAI) agents therefore move beyond one-off explanations and embrace conversational approaches to provide explanations. These conversational explanations have been demonstrated to be more effective for users’ model understanding (Shen et al., 2023; Mindlin et al., 2024; He et al., 2025). Conceptually, a ConvXAI agent can be decomposed

[†]Work done at TU Berlin.

¹The source code (including evaluation suite) is available at: <https://github.com/bach-mr/SimXAI>

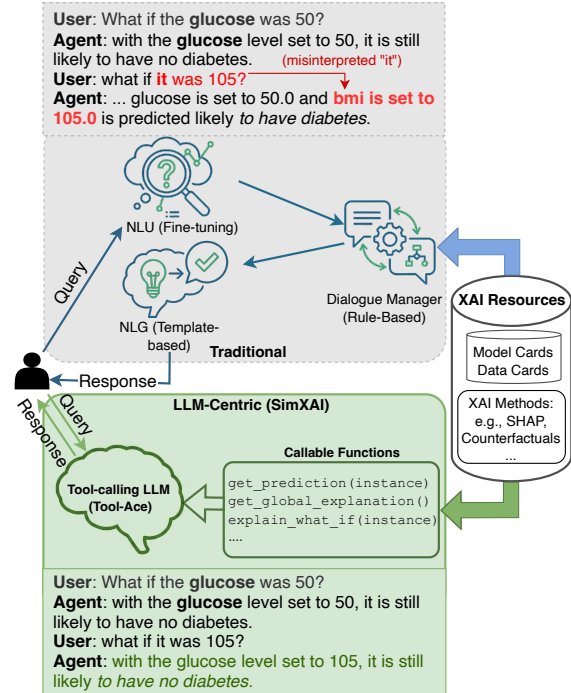


Figure 1: Comparison of TALKToMODEL’s traditional GUS architecture with SIMXAI’s LLM-centric conversational component. Traditional systems rely on multiple components that require rules, templates, or fine-tuning. In contrast, SIMXAI requires only function specifications and a single LLM. Both architectures share the same XAI resources.

into two parts: (i) a conversational layer (**ConvXAI layer**), responsible for managing the interaction with the user, and (ii) the underlying **XAI resources**, which store or generate the explanatory information (e.g., model/data cards (Mitchell et al., 2019; Pushkarna et al., 2022), or feature-attribution methods like SHAP (Lundberg and Lee, 2017) or counterfactual explanations (Mothilal et al., 2020; Ross et al., 2021)). Most current ConvXAI agents (i.e., Nguyen et al. 2023; Slack et al. 2023; Feldhus et al. 2023; Wang et al. 2024a) adopt the traditional GUS (Genial Understanding System) architecture (Bobrow et al., 1977) for the Con-

vXAI layer, which comprises three distinct components: *natural language understanding (NLU)* for intent recognition, *dialogue management* for context maintenance and rule application, and *natural language generation (NLG)* for response production. Although such GUS-based ConvXAI layers present explanations through interactive dialogue, they rely on **multiple models separately fine-tuned on handcrafted data** for intent recognition in NLU, **rule-based** dialogue management, and **template-based** response generation in NLG (Figure 1). As a result, they remain limited in contextualization, scalability, and generalization (Rastogi et al., 2020). These limitations highlight the need for lightweight, adaptive, and flexible ConvXAI layers. Recent developments suggest that LLMs can fulfill this role and replace traditional ConvXAI layers in such systems (Li et al., 2024).

We introduce **SIMXAI**, a simple, lightweight, LLM-centric framework that acts as a ConvXAI layer between users and existing XAI resources (see Figure 1). Unlike rule- or template-based systems, SIMXAI uses an LLM to support natural, context-aware dialogue without requiring handcrafted intent datasets or conversational fine-tuning. The layer is **data- and model-agnostic**. Integrating it into an agent requires only function implementations that provide access to the agent’s existing XAI resources. SIMXAI can therefore replace the conversational components of GUS-based agents (e.g., Slack et al. 2023; Nguyen et al. 2023) while preserving the underlying explanations. **SIMXAI is not a standalone agent** and does not itself select or validate XAI methods. Instead, it provides a modular interface to resources supplied by the host system.

Using LLMs for tool calling is not technically novel. Our contribution is a ConvXAI-specific design that separates a replaceable conversational layer from XAI resources. The design also provides a function interface based on XAI question types and an evaluation framework that separately measures function parsing, response-to-tool faithfulness, and contextualization.

Prior work on evaluating ConvXAI has mainly focused on isolated turns (Malandri et al., 2023; Wang et al., 2024a) and has often overlooked dialogue context. To address this gap, we evaluate SIMXAI on parsing accuracy in single- and multi-turn dialogues, contextualization, and response-to-tool faithfulness. For each aspect, we provide a dataset with manually annotated ground truth,

corresponding metrics, and an automated evaluation pipeline. The faithfulness evaluation measures whether the conversational layer preserves the returned XAI output. It does not measure whether the underlying explanation method is faithful to the predictive model.

In a user study with 30 participants, replacing the ConvXAI layer of TALKToMODEL with SIMXAI while keeping the underlying XAI resources unchanged, significantly improved both, perceived interaction quality and task performance.

In summary, our contributions are threefold. **First**, based on the analysis of limitations of prior work (§2), we introduce SIMXAI, an LLM-centric framework that addresses these limitations and offers a practical alternative to conventional ConvXAI layers (§3). **Second**, we present a multi-aspect evaluation suite for ConvXAI layers, including automated assessment of parsing, response-to-tool faithfulness, and contextualization (§4). **Third**, we show through quantitative evaluation and a user study that replacing a conventional conversational layer with SIMXAI improves significantly conversational quality while preserving faithfulness to underlying XAI outputs (§5, §6).

2 Limitations of Conversational Layers in Prior ConvXAI Agents

GUS-based ConvXAI layers (such as Nguyen et al. 2023; Slack et al. 2023; Feldhus et al. 2023) support follow-up questions, but their conversational behavior is typically implemented through task-specific grammars, intent schemas, templates, or dialogue rules. As summarized in Table 1, this design leads to a recurring trade-off between *faithfulness* and three other dimensions: *contextualization*, *scalability*, and *generalization*.

First, prior systems provide only *weak-to-moderate contextualization*, handling context primarily through pre-defined operational structures rather than flexible discourse-level reasoning. TALKToMODEL resolves follow-up questions through grammar operators such as `previous_filter` and `previous_operation` (Slack et al., 2023). INTERROLANG extends this design with intent recognition, slot filling, reference resolution, and clarification strategies (Feldhus et al., 2023). XAGENT supports multi-turn interaction through a fixed dialogue manager and a curated phrase bank (Nguyen et al., 2023). Such mechanisms

System	Contextualization	Scalability	Generalization	Response faithfulness
TALKToMODEL (Slack et al., 2023)	Medium	Low (FT, D, T, R)	Low	High
INTERROLANG (Feldhus et al., 2023)	Weak	Low (FT, D, T, R)	Low	High
XAGENT (Nguyen et al., 2023)	Weak	Medium (FT, D, T, R)	Low	High
SIMXAI	Strong	High	High	High

Table 1: Design-level comparison of SIMXAI with the conversational layers of prior GUS-based ConvXAI agents in terms of contextualization, scalability, generalization, and response-to-resource faithfulness. The table summarizes architectural properties from the respective system designs. Scalability constraints include: Finetuning (FT), Handcrafted Data (D), Templates (T), and Rules (R).

can handle follow-up questions only to a limited extent and often struggle when users rely on ellipsis, anaphora, or long-range dependencies (see Figure 1 for an example of a failure case).

Second, this design leads to *low-to-moderate scalability* and *limited generalization* across domains and tasks. TALKToMODEL requires dataset-specific grammars and synthetic utterance–parse pairs (Slack et al., 2023), INTERROLANG depends on annotated intents, slots, and task-specific operations (Feldhus et al., 2023), and XAGENT relies on manually curated phrases and response templates (Nguyen et al., 2023). As a result, extending such systems to new settings usually requires additional annotation or rule engineering rather than simply integrating new XAI resources. This also limits generalization: prior systems are typically demonstrated in narrow settings, such as tabular data or NLP-specific pipelines, and their conversational control logic does not transfer easily beyond those conditions.

At the same time, these constrained designs maximally preserve *faithfulness* in response generation. As they generate outputs through rules or templates, the conversational layer is less likely to introduce unsupported claims or misrepresent the underlying XAI method results. The problem with prior systems is not faithfulness itself, but the rigid control logic they use to maintain it. That logic limits how well prior systems handle context and makes these systems hard to adapt across domains.

In contrast, SIMXAI addresses these problems by replacing handcrafted conversational control logic with a tool-calling LLM that bridges access to existing XAI resources, without requiring finetuning, handcrafted data, or templates. This design enables *strong contextualization*, *high scalability* and *high generalization*, as it does not rely on domain-specific implementations.

3 SimXAI

SIMXAI adds a conversational interface to existing XAI resources using only two components: (1) an off-the-shelf LLM with tool-calling support and (2) a set of general function specifications that map user requests onto available XAI resources. The novelty lies in this ConvXAI-specific separation and interface, rather than in the general mechanism of LLM tool calling. Unlike prior systems, SIMXAI does not rely on task-specific grammars, intent schemas, dialogue managers, or response templates. Instead, SIMXAI interprets the current user query together with the dialogue history, selects the appropriate function, fills its arguments from context when applicable, and verbalizes the returned result. The backend remains responsible for computing the prediction, explanation, or data/model information returned by that function.

Formally, given a user query q , dialogue history H , and a set of available functions $\mathcal{T}$, the system either answers directly or invokes a function $t \in \mathcal{T}$. When a function call is required, the system extracts the necessary arguments from q and H , asks for clarification if required information cannot be recovered from the dialogue context, executes the function, and generates a final response grounded in the returned output. This design replaces handcrafted conversational control logic with a single tool-calling LLM interface that bridges access to existing XAI resources. Algorithm 1 summarizes this interaction loop.

3.1 Tool-Calling LLM

We use ToolACE-2.5-Llama-3.1-8B², an off-the-shelf LLM trained for tool-use reasoning (Liu et al., 2025; Grattafiori et al., 2024). In SIMXAI, the LLM is responsible for four decisions: (i) whether tool use is needed, (ii) if so, which function to call, (iii)

²<https://huggingface.co/Team-ACE/ToolACE-2.5-Llama-3.1-8B>

Algorithm 1 Response Generation

Input: User query q , dialogue history H , tool set $\mathcal{T}$

Output: Final response r

```
1:  $u \leftarrow \text{DECIDETOOLUSAGE}(q, H, \mathcal{T})$ 
2: if  $u = \text{None}$  then
3:    $r \leftarrow \text{INTERNALKNOWLEDGE}(q, H)$ 
4:   return  $r$ 
5: end if
6:  $t \leftarrow \text{SELECTTOOL}(q, H, \mathcal{T})$ 
7:  $a \leftarrow \text{EXTRACTARGUMENTS}(t, q, H)$ 
8: while  $\neg \text{VALID}(a, t)$  do
9:    $m \leftarrow \text{FINDMISSINGARGUMENTS}(a, t)$ 
10:   $a' \leftarrow \text{SEARCHINDIALOGUE}(m, H)$ 
11:  if  $a' \neq \emptyset$  then
12:     $a \leftarrow \text{UPDATEARGUMENTS}(a, a')$ 
13:  else
14:     $c \leftarrow \text{ASKUSERFORCLARIFICATION}(m)$ 
15:     $H \leftarrow H \cup \{c\}$ 
16:     $a \leftarrow \text{UPDATEARGUMENTS}(a, c)$ 
17:  end if
18: end while
19:  $o \leftarrow \text{CALLTOOL}(t, a)$ 
20:  $r \leftarrow \text{GENERATERESPONSE}(q, H, o)$ 
21: return  $r$ 
```

how to ground its arguments in the current turn and dialogue history, and (iv) how to generate a final natural-language response from the function output. No additional parser, dialogue manager, or template bank is required.

3.2 Function Set

Our set of functions is based on the question bank of Liao and Varshney (2022); Liao et al. (2020). This question bank was created from practitioner interviews and organized around model-, data-, and context-related questions about AI systems. From this bank, we implement 11 functions: one retrieves the model prediction for a given instance and the other ten address explanatory questions, such as *why* the model made a particular prediction or *how* to modify an instance to change the prediction. For each function, we specify its purpose, inputs, arguments, and outputs. Table 2 summarizes the full set. The set is extensible: if no predefined function matches a user’s intent, a new function can be added with its own specification (Figure 2).

Function Specification. Function specification is critical for ensuring that the LLM calls the correct function. For each function, we adopt a standard specification template (Figure 2) that defines the function name, argument list and types, documentation, and expected input and output types. *These specifications are not tied to any particular dataset or model.* Instead, we use general specifications that apply across datasets and models, enabling

broad generalization (cf. Figure 5 in the appendix for a concrete example of a counterfactual function specification).

```
function_name(argument: type) -> return_type
# summary of what the function does.
Args:
  argument (type): Description of the
    input (format, constraints).
Returns:
  return_type: Description of the
    output (labels, score range, etc.).
```

Figure 2: Template for Function Specification.

4 Evaluation

We evaluate ConvXAI layers along three aspects: *parsing*, *response-to-tool faithfulness*, and *contextualization*. Parsing measures whether the system selects the correct function and arguments for a user request, response-to-tool faithfulness measures whether the final response is supported by the returned XAI output, and contextualization measures whether the system uses dialogue history appropriately when interpreting the current turn. Table 4 summarizes the evaluation suite. This scope evaluates the conversational layer only: it neither measures nor guarantees the intrinsic correctness or model-faithfulness of the backend XAI method that produced the tool output. We use this suite for two purposes. First, we assess SIMXAI directly across settings of increasing conversational complexity. Second, we qualitatively analyze gains from replacing the conversational components of prior GUS-based ConvXAI systems with SIMXAI while keeping their underlying XAI resources unchanged.

4.1 Parsing

Parsing evaluates whether the system maps a user request to the correct function and arguments. We report exact-match accuracy over the selected function. Unlike prior work, which mainly evaluates parsing in single-turn settings (Slack et al., 2023; Feldhus et al., 2023; Nguyen et al., 2023), we consider three levels of conversational complexity.

Single-turn parsing. We evaluate standard single-turn function selection on two datasets in ConvXAI settings: CoXQL (Wang et al., 2024b) for explanation request parsing (Figure 4) and Compass (Wang et al., 2025) for user input extraction

Function	Description
<code>get_prediction(instance)</code>	Return the model prediction for an instance.
<code>get_global_explanation()</code>	Provide a global explanation of model behavior.
<code>get_local_explanation(instance, prediction_label)</code>	Explain why the model produced a specific prediction.
<code>explain_why_not(instance, expected_label)</code>	Explain why an expected label was not predicted.
<code>explain_how_to_be_that(instance, target_label)</code>	Suggest how the input could be changed to obtain a target label.
<code>explain_how_to_still_be_this(instance)</code>	Identify changes that would preserve the current prediction.
<code>explain_what_if(instance)</code>	Show how predictions change under input modifications.
<code>get_model_performance(metric)</code>	Report a model performance metric.
<code>get_data_information()</code>	Provide information about the training and evaluation data.
<code>get_model_information()</code>	Describe the model, training setup, strengths, and limitations.
<code>get_output_information()</code>	Explain the meaning of the model outputs.

Table 2: Core functions used by SIMXAI.

(Figure 6). In this setting, the current query can be interpreted without dialogue history.

Multi-turn parsing. To evaluate short-range contextual parsing, we construct Compass-MultiTurn from Compass by merging related single-turn interactions into short conversations and removing redundant user turns (details in Appendix A). In this setting, the use of recent dialogue history is required to parse the users’ intent correctly.

Dialogue-level parsing. To evaluate long-range dependencies, we generate synthetic dialogue-level datasets in two domains: sentiment analysis SYN-Sentiment and heart rate prediction SYN-HeartRate. Each of the two datasets combines dialogues generated under two settings: **Constrained**, in which we provide our 11 function specifications in the prompt and ask GPT-5.1 to generate realistic conversations covering all operations, and **Unconstrained**, in which GPT-5.1 is prompted to generate realistic conversations without any predefined constraints. Each user turn defines one parsing instance, with the full preceding dialogue provided as context. Human validation assessed these synthetic dialogues as mostly realistic (details in Appendix B).

Together, these datasets assess whether ConvXAI systems can select the correct function corresponding to user’s intent. They range from isolated questions up to full multi-turn conversations with long range dependencies. Table 3 summarizes the parsing datasets.

4.2 Response-to-Tool Faithfulness

As LLMs are prone to hallucination (Huang et al., 2025), we evaluate the faithfulness of the agent’s responses with respect to the provided XAI resources. Following prior work (Adlakha et al., 2024), we

Dataset	#Dialogues	#Turns	#Queries
CoXQL + Compass	–	1	462
Compass-MultiTurn	10	4–6	26
SYN-Sentiment	4	12–24	40
SYN-HeartRate	4	12–24	31

Table 3: Statistics for parsing datasets with increasing conversational complexity.

define *response-to-tool faithfulness* as the extent to which the conversational layer preserves the information contained in a returned XAI output without introducing unsupported claims. This is distinct from the intrinsic faithfulness of the XAI method to the predictive model, which is outside our scope.

We construct two faithfulness datasets, SYN-QA-Sentiment and SYN-QA-HeartRate, based on SYN-Sentiment and SYN-HeartRate. The datasets contain 71 question–answer pairs covering prediction lookup, explanation outputs, model descriptions, and dataset information. For each question, we manually construct a gold standard function output and evaluate whether the system response faithfully preserves this information using a binary judgment and accuracy as aggregated metric.

4.3 Contextualization

Contextualization evaluates whether the system uses dialogue history appropriately to answer the current query. We first assess *context relevance* by manually determining whether a query can be answered independently or requires the preceding dialogue. For queries that depend on prior context, we then assess whether the response incorporates the necessary dialogue history to resolve the current turn coherently. This *contextual answer coherence* is a binary judgment, aggregated as accuracy (the fraction of contextually coherent answers).

Aspect	Subtask	What is evaluated	Dataset	SIMXAI
Parsing	Single-turn	Correct function selection from a single user query	CoXQL, Compass	86%
	Multi-turn	Correct function selection from short dialogue history	Compass-MultiTurn	88%
	Dialogue-level	Correct function selection from full conversations	SYN-Sentiment	92.5%
		with long-range dependencies	SYN-HeartRate	96%
Faithfulness	–	Whether the response is fully supported by the underlying XAI output	SYN-QA-Sentiment	93.6%
	–		SYN-QA-HeartRate	91.5%
Contextualization	Contextual relevance	Whether the current query requires prior dialogue context	SYN-Sentiment	N/A
	Contextual coherence	Whether responses to context-dependent queries use prior context coherently	SYN-HeartRate	N/A
			SYN-Sentiment	90.9%
			SYN-HeartRate	94.74%

Table 4: Overview of the evaluation suite for the ConvXAI layer and the corresponding accuracy results of SIMXAI.

5 Results

To assess SIMXAI, we report performance on parsing, response-to-tool faithfulness, and contextual coherence. Following recent work on LLM-based evaluation (Xu et al., 2025; Tamber et al., 2025), we additionally examine opportunities for automated evaluation by measuring the agreement between human rating and LLM-as-a-judge.³ Specifically, we report how often the LLM judge assigns the same label as the human annotator for response-to-tool faithfulness and contextualization, including *context relevance* and *contextual coherence*.

5.1 SIMXAI Performance

Table 4 reports results across the three aspects in the evaluation suite: parsing, response-to-tool faithfulness, and contextualization.

For *parsing*, SIMXAI performs strongly in all settings, achieving above 86% accuracy for single-turn, multi-turn, and dialogue-level evaluation. Interestingly, accuracy is higher on the multi-turn and dialogue-level datasets, suggesting that dialogue context helps the model to resolve user intents better.

In most cases (over 91%), SIMXAI produces responses that are *faithful to the tool output*, indicating that the conversational layer preserves the returned XAI information rather than introducing unsupported claims. This is particularly important because SIMXAI replaces rule- and template-based conversational components with an LLM-based interface, which may increase the risk of hallucinations. These results do not evaluate the intrinsic faithfulness of the backend XAI methods.

SIMXAI further shows high *contextual answer coherence* across domains: 90.91% for sentiment

analysis and 94.74% for heart rate, indicating that prior dialogue context is generally incorporated correctly in context-dependent queries.

Overall, SIMXAI performs well across all three aspects of the evaluation suite. Parsing accuracy remains high from single-turn to dialogue-level evaluation, responses are usually supported by the returned XAI outputs, and most context-dependent queries are answered coherently.

5.2 LLM-as-a-judge Performance

Figure 3 shows the agreement between LLM-as-a-judge and human annotation. The agreement on **response-to-tool faithfulness** is high (around 90%), suggesting that LLM-based evaluation is a promising proxy for this aspect on our datasets. While the agreement on *contextual relevance* in sentiment analysis is equally high (around 90%), it drops to 76.60% in heart rate prediction. This indicates that automatically determining whether a query depends on prior context remains challenging in some domains. For *contextual answer coherence*, the LLM judge agrees with human annotations in more than 80% of cases across both domains, suggesting that it can serve as a useful proxy for coherence evaluation.

In summary, evaluation with LLM-as-a-judge shows potential for automated assessment of response-to-tool faithfulness and contextualization. Despite lower agreement on context relevance in heart rate prediction, it may still serve as an efficient approach for initial comparative evaluation.

6 GUS-based system comparison

SIMXAI is designed to be integrated into existing conversational XAI agents by replacing their original conversational components. We qualitatively examine how SIMXAI can improve existing

³We use <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

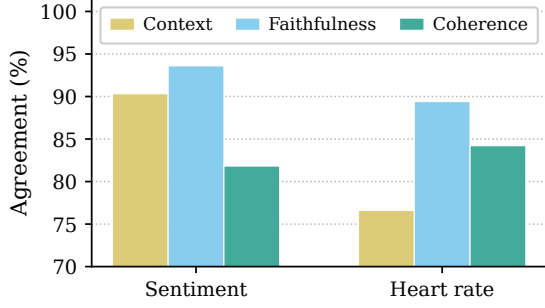


Figure 3: Agreement between LLM-as-a-judge and human annotation.

systems, by replacing the conversational layer of the ConvXAI agents TALKToMODEL (Slack et al., 2023), INTERROLANG (Feldhus et al., 2023), and XAGENT (Nguyen et al., 2023) with SIMXAI and re-assess cases, in which the original system failed (e.g., due to missing context handling abilities).

For each baseline, we keep the original function set and underlying XAI outputs fixed, and replace only the module that maps user requests and dialogue history onto function selection and response generation. This isolates the contribution of SIMXAI acting solely as conversational layer and not as a modified ConvXAI backend.

We further conduct a user study on the integration of SIMXAI into TALKToMODEL, a representative grammar-based system. We directly compare the original model with its SIMXAI-integrated counterpart in terms of both, objective and subjective metrics.

6.1 Qualitative results

Figure 1 shows an example failure case of incorrect context handling of TALKToMODEL’s native conversational layers. TALKToMODEL fails to maintain contextual understanding because the agent misinterprets the referent of “it” (anaphora), which leads to an incorrect final decision. Replacing TALKToMODEL’s conversational layers by SIMXAI fixes this issue and correctly resolves the context in the same conversation dialogue. We observe similar patterns of incorrect context handling in the other agents under comparison, which are partially unable to provide an answer at all. Replacing the conversational layers of these agents by SIMXAI corrects such issues (examples in Appendix C).

6.2 User Study

We conduct a controlled within-subjects study to evaluate whether integrating SIMXAI into TALKTo-

MODEL (Slack et al., 2023) 1) improves the perceived quality of the conversation and 2) leads to a better understanding of the machine learning model. To isolate the effect of the conversational component, we keep all other components (datasets, classifiers, and XAI resources) unchanged. We compare two agents: (i) the original TALKToMODEL with its fine-tuned ConvXAI layers and (ii) TALKToMODEL with SIMXAI replacing the ConvXAI layers. Both agents follow the original TALKToMODEL configuration described by Slack et al. (2023).

6.2.1 Setup

We recruited 30 lay participants via Prolific.⁴ This sample size aligns with prior HCI and XAI studies evaluating interactive explanation systems (Kaur et al., 2020; Slack et al., 2023). Participants interact with both agents using the Diabetes dataset (Kahn, 1994), as in the original TALKToMODEL study.

For each agent, participants complete three tasks. These tasks cover multiple categories from an established XAI question taxonomy (Liao et al., 2020), including data understanding, local explanations, and counterfactual inspection. Each task contains at least two questions, for a total of nine questions across all tasks. The questions are sampled from a pool of predefined multiple-choice questions for each task (details in Appendix Section D.1). They are designed to be answerable using the ground-truth information provided by the agent’s explanations. Some questions by design require multi-turn interactions to arrive at the correct answer, in order to assess conversational coherence and context handling. We randomize the order of agents, and although the question types remain consistent across conditions, specific values (e.g., instance identifiers) are varied to mitigate learning effects.

We collect both, objective and subjective measures. **Objective metrics** are (1) answer accuracy, (2) conversational efficiency measured as the number of turns per correctly answered question, which is commonly used in dialogue agent evaluation (Su et al., 2016; Peng et al., 2017), and (3) indicators of conversational breakdowns, such as repeated or rephrased questions and agent failures, where the agent does not provide meaningful answers. **Subjective measures** consist of two parts. First, we use the 5-point Likert BUS-11 usability scale (Borsci and Schmettow, 2024), which includes 11 items

⁴<https://www.prolific.com/> Participants receive a compensation of \$8/hour

Measure	TALKToMODEL	SIMXAI
Accuracy $\uparrow$	59.0	92.1
Chat turns $\downarrow$	12.4	10.0
Chat time (min) $\downarrow$	7.4	5.8
Turns per correct answer $\downarrow$	2.70	1.32

Table 5: Objective results from the user study.

BUS-11 item (shortened)	TALKToMODEL	SIMXAI
Interaction Clarity	3.33	4.60
Context Awareness	3.13	4.37
Response Comprehensibility	3.60	4.57
Goal Facilitation	3.10	4.37
Information Sufficiency	3.23	4.37
Information Relevance	3.13	4.13
Response Accuracy	3.63	4.47
Response Time	4.37	4.03
Cronbach’s α	0.925	0.856

Table 6: Mean BUS-11 ratings from the user study (1 = strongly disagree, 5 = strongly agree). We report eight relevant items in shortened form, full results are in Table 11. Cronbach’s $\alpha > 0.8$ indicates high internal consistency.

across five subscales, focusing on perceived conversation and information quality and perceived chatbot functionality. Second, we include a **pairwise comparison** questionnaire between the two agents to assess user preference and explanation effectiveness.

6.2.2 Results

This section summarizes the key objective and subjective results. Details on the significance tests for the reported metrics, the full BUS-11 scale, and the complete pairwise comparison are provided in Appendix Section D.3.

Objective results. Replacing TALKToMODEL’s conversational component with SIMXAI leads to clear improvements in both, task performance and perceived interaction quality. Table 5 shows that participants achieve substantially higher accuracy with SIMXAI than with the original TALKToMODEL (92.1% vs. 59.0%, $p < .001$). SIMXAI also reduces interaction costs: users complete the task in fewer turns on average (10.0 vs. 12.4 turns) and in less time (5.8 vs. 7.4 minutes). Across all sessions, this translates to roughly half as many turns per correct answer with SIMXAI (1.32 vs. 2.70).

Subjective results. Table 6 reports the BUS-11 usability ratings. We observe the largest improvements in clarity of communication, context tracking, helpfulness in achieving the user’s goal, and

providing an appropriate amount of information (all $p < .001$). Perceived response time is the only item where SIMXAI scores lower, suggesting a possible trade-off between conversational quality and latency. We attribute this to the models deployed for parsing in TALKToMODEL, which are comparably smaller (ranging from 60M to 6B) than to the tool-augmented LLM we use in SIMXAI (§3.1). However, the difference is not statistically significant ($p > 0.1$).

Pairwise comparison. 80.0% of the participants prefer SIMXAI overall, while 16.7% prefer TALKToMODEL and 3.3% indicate no preference (cf. Table 9 in the Appendix). We observe the same proportions for perceived task effectiveness.

Overall, the user study shows that replacing the conversational component of TALKToMODEL with SIMXAI improves both, perceived interaction quality and task performance, even when the underlying XAI resources remain unchanged.

7 Related Work

Besides the systems analyzed in §2, that rely on fine-tuned models, off-the-shelf LLMs are increasingly adopted in ConvXAI. Nguyen et al. (2024) extend XAGENT by Llama-2 (Touvron et al., 2023) for postprocessing NLG and answering general factual questions (e.g., about machine learning terminology). KNOWTHYSELF (Prasai et al., 2025) is a multi-agent framework, which at its core still resembles the GUS architecture. It employs Gemma3-27B (Team et al., 2025) for most parts of NLU and NLG, while user intents are matched to specialized agents by embedding-based similarity, using nomic-embed-text (Nussbaum et al., 2025). These specialized agents encapsulate XAI resources. Closest to our work is ECHO (Vanbrabant et al., 2025), another agentic framework with a tool-calling LLM at its core. However, ECHO trades faithfulness for human validation “... might not always result in a helpful or correct answer [...] This further highlights the importance of human-in-the-loop workflows, in which the LLM can try again until the answer is satisfactory”. This undermines the purpose of ConvXAI to explain a model’s true reasoning, as any plausible answer is likely satisfactory. Yet, plausibility and correctness are different evaluation criteria that should not be conflated (Nauta et al., 2023) (a plausible answer might not be correct, while a correct one might not be plausible, e.g., if the model’s reasoning is

flawed).

All these systems are demos without quantitative or qualitative evaluation. Unlike prior GUS-based systems or existing tool-calling ConvXAI agents, our contribution is not the generic tool-calling mechanism itself. SIMXAI introduces a simple, lightweight, open-source intermediate layer with a ConvXAI-specific separation between conversational control and underlying XAI resources, an extensible XAI function design, and a dedicated evaluation suite. The measured faithfulness in this suite is response-to-tool faithfulness and should not be interpreted as validation of the backend explanation method.

8 Conclusion

SIMXAI enables natural conversational interactions with existing XAI resources through general callable functions. Across evaluations of parsing, response-to-tool faithfulness, and contextualization, its lightweight LLM-centric design supports conversational explanations while usually preserving the information returned by the XAI resources. Our user study shows that replacing the conversational layers of a traditional GUS-based ConvXAI agent by SIMXAI significantly improves task performance and perceived interaction quality. This suggests that LLM-centric architectures provide tangible benefits to users in understanding AI model behavior. In future work, we plan to explore and enhance fully automated evaluation for LLM-based conversational XAI layers. Based on that, we plan to study how LLM choice affects tool calling and to develop further improvements to ConvXAI.

Limitations

Like other LLM-based systems, SIMXAI introduces additional latency and remains susceptible to hallucinations.

SIMXAI improves only the conversational component: it assumes that appropriate XAI resources are already available and does not assess the intrinsic quality, correctness, or model-faithfulness of the underlying explanation methods themselves. Consequently, a response can be faithful to a tool output even when that output is incomplete or misleading.

We do not study the impact of the LLM choice for tool calling or LLM-as-a-judge. Our experiments use a lightweight 8B LLaMA 3.1 model, and results may vary with other models.

References

- Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. [Evaluating correctness and faithfulness of instruction-following models for question answering](#). *Transactions of the Association for Computational Linguistics*, 12:681–699.
- Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. [Gus, a frame-driven dialog system](#). *Artificial Intelligence*, 8.
- Simone Borsci and Martin Schmettow. 2024. Re-examining the chatbot usability scale (bus-11) to assess user experience with customer relationship management chatbots. *Personal and Ubiquitous Computing*, 28(6):1033–1044.
- 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. [Is conversational xai all you need? human-ai decision making with a conversational xai assistant](#). In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, page 907–924, New York, NY, USA. Association for Computing Machinery.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Michael Kahn. 1994. Diabetes. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5T59G>.
- Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. [Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning](#). In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, page 1–14, New York, NY, USA. Association for Computing Machinery.

- Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. [Rethinking explainability as a dialogue: A practitioner's perspective](#). *Preprint*, arXiv:2202.01875.
- Zekun Li, Zhiyu Zoey Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A. Crook. 2024. [Large language models as zero-shot dialogue state tracker through function calling](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8688–8704, Bangkok, Thailand. Association for Computational Linguistics.
- Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. [Questioning the ai: Informing design practices for explainable ai user experiences](#). In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery.
- Q. Vera Liao and Kush R. Varshney. 2022. [Human-centered explainable ai \(xai\): From algorithms to user experiences](#). *Preprint*, arXiv:2110.10790.
- Weiwu Liu, Xu Huang, Xingshan Zeng, xinlong hao, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Zhengying Liu, Yuanqing Yu, Zezhong WANG, Yuxian Wang, Wu Ning, Yutai Hou, Bin Wang, Chuhan Wu, Wang Xinzhi, Yong Liu, Yasheng Wang, and 8 others. 2025. [Toolace: Winning the points of llm function calling](#). In *International Conference on Representation Learning*, volume 2025, pages 41359–41381.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Lorenzo Malandri, Fabio Mercorio, Mario Mezzananza, and Navid Nobani. 2023. Convxai: a system for multimodal interaction with any black-box explainer. *Cognitive Computation*, 15(2):613–644.
- 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.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.
- Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617.
- 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](#). *ACM Comput. Surv.*, 55(13s).
- Van Bach Nguyen, Jörg Schlötterer, and Christin Seifert. 2023. From black boxes to conversations: Incorporating xai in a conversational agent. In *World Conference on Explainable Artificial Intelligence*, pages 71–96. Springer.
- Van Bach Nguyen, Jörg Schlötterer, and Christin Seifert. 2024. Xagent: A conversational xai agent harnessing the power of large language models. In *Joint Proceedings of the xAI 2024 Late-breaking Work, Demos and Doctoral Consortium co-located with the 2nd World Conference on eXplainable Artificial Intelligence (xAI-2024)*, Valletta, Malta, July 17, 2024, pages 273–280.
- Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. 2025. [Nomic embed: Training a reproducible long context text embedder](#). *Transactions on Machine Learning Research*. Reproducibility Certification.
- Baolin Peng, Xijun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. [Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2240, Copenhagen, Denmark. Association for Computational Linguistics.
- Suraj Prasai, Mengnan Du, Ying Zhang, and Fan Yang. 2025. [Knowthyself: An agentic assistant for llm interpretability](#). *Preprint*, arXiv:2511.03878.
- Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjar-tansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. [Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.
- Alexis Ross, Ana Marasović, and Matthew E Peters. 2021. Explaining nlp models via minimal contrastive editing (mice). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852.
- 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*.
- Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. [On-line active reward learning for policy optimisation in spoken dialogue systems](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441, Berlin, Germany. Association for Computational Linguistics.
- Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, and Jimmy Lin. 2025. [Benchmarking LLM faithfulness in RAG with evolving leaderboards](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 799–811, Suzhou (China). Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Sebe Vanbrabant, Gilles Eerlings, Gustavo Alberto Rovelo Ruiz, and Davy Vanacken. 2025. [Echo: Enhancing conversational explainable ai through tool-augmented language models](#). *Proc. ACM Hum.-Comput. Interact.*, 9(4).
- 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.
- Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Fedor Splitt, Jiaao Li, Yoana Tsoneva, Sebastian Möller, and Vera Schmitt. 2025. [Multilingual datasets for custom input extraction and explanation requests parsing in conversational XAI systems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 534–555, Suzhou, China. Association for Computational Linguistics.
- Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. 2025. [Does context matter? Contextual-JudgeBench for evaluating LLM-based judges in contextual settings](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9541–9564, Vienna, Austria. Association for Computational Linguistics.
- Tong Zhang, Mengao Zhang, Wei Yan Low, X. Jessie Yang, and Boyang Albert Li. 2025. [Conversational explanations: Discussing explainable ai with non-ai experts](#). In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, page 409–424, New York, NY, USA. Association for Computing Machinery.

A Parsing

Failure in single-turn parsing. Figure 6 shows the comparison between single-turn and multi-turn parsing. For multi-turn parsing, LLMs should be able to utilize the context from previous turns. As illustrated in the third example in Figure 6, users can simply use pronouns like “it” to reference previous questions without repeating the actual question from the last turn. Such cases would fail in single-turn parsing, as there is no conversational context to resolve what “it” refers to.

From single-turn to multi-turn. Table 3 To build the multi-turn version (Figure 6, right side), related consecutive questions that operated on the same input text were grouped together into multi-turn conversation sets. Specifically, questions that originally repeated the full instance text had those explicit instance references replaced with pronouns like “it” turning them into natural follow-up turns. This way, the conversation flows naturally: the first turn establishes the target text, and subsequent turns refer back to it implicitly — just as a human would in real dialogue. In total, we created 10 multi-turn dialogues with averaging 4–6 turns each.

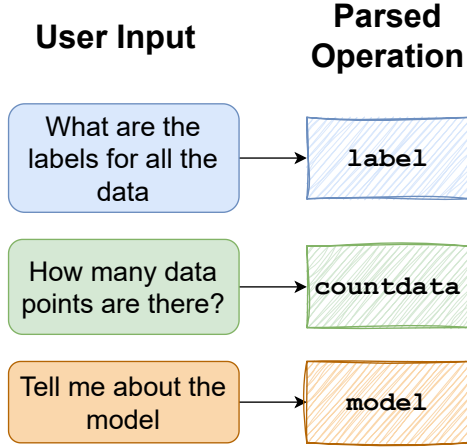


Figure 4: CoXQL examples.

B Realism Validation

Because the conversations in our dataset were generated using GPT-5.1, we additionally evaluated their realism from a human perspective to verify that they resemble plausible explanatory dialogues.

Specifically, we asked five members of our research group with a background in machine learning to assess the realism of user utterances in the synthetically generated explanatory conversations. Participants were asked to imagine themselves as users seeking to understand how a classification model makes its decisions. They were then presented with generated dialogues between a *User* and an *Explainer*. Their task was to judge whether the User’s messages resembled authentic user questions, reflected a consistent level of knowledge, and remained on topic throughout the conversation.

For each conversation, participants provided a rating on a 5-point Likert scale ranging from 1 (**completely unrealistic**) to 5 (**completely realistic**). In addition to the numerical rating, respondents were required to provide a justification for their score, allowing us to collect both quantitative assessments and qualitative rationales regarding perceived conversational realism.

The average score across all conversations was 3.75, suggesting that the generated dialogues achieved a reasonable level of perceived realism.

C Qualitative examples

In this section, we present additional qualitative results illustrating cases where the original conversational XAI agents fail to capture dialogue context

```

def explain_how_to_be_that(instance: str,
    target_label: str = None,
    number_of_iterations: int = 1) -> str:
    """
    Call this whenever you need to explain
    how to modify the instance minimally to
    achieve the target prediction
    (counterfactual). For example, user
    asked "How can I change this instance to
    get prediction X?"
    The tool suggests changes needed to flip
    the current prediction to the desired
    one.

    Args:
        instance: The input instance.
        target_label: (Optional) The desired
        prediction or class. If None, suggests
        changes to flip the current prediction.
        number_of_iterations: Number of
        counterfactual suggestions to generate.

    Output:
        Counterfactual changes (e.g.,
        "increase feature X by 0.2").
    """

```

Figure 5: Counterfactual Function Specification Example (Python)

and how SIMXAI improves their responses. Table 7 provides examples for each existing agent. Each agent operates on different tasks and exhibits different types of contextualization failures. SIMXAI improves these responses across all cases without requiring task-specific or architecture-specific adaptations.

D User Study Details

D.1 Question Pool

Table 8 presents the three tasks and the corresponding question templates that participants were asked to answer. The questions cover several categories of the XAI question taxonomy (Liao et al., 2020), including data understanding, local explanations, and counterfactual inspection.

Some question templates contain placeholders that are replaced with specific values during the study. For these questions, the feature and value are randomly selected when generating the question. The study uses two IDs, which are randomly assigned to the conversational agents.

Each task includes multiple possible questions, and all three tasks are presented to each participant. For each task, questions are randomly selected from the pool without replacement, ensuring that within a single session participants receive as

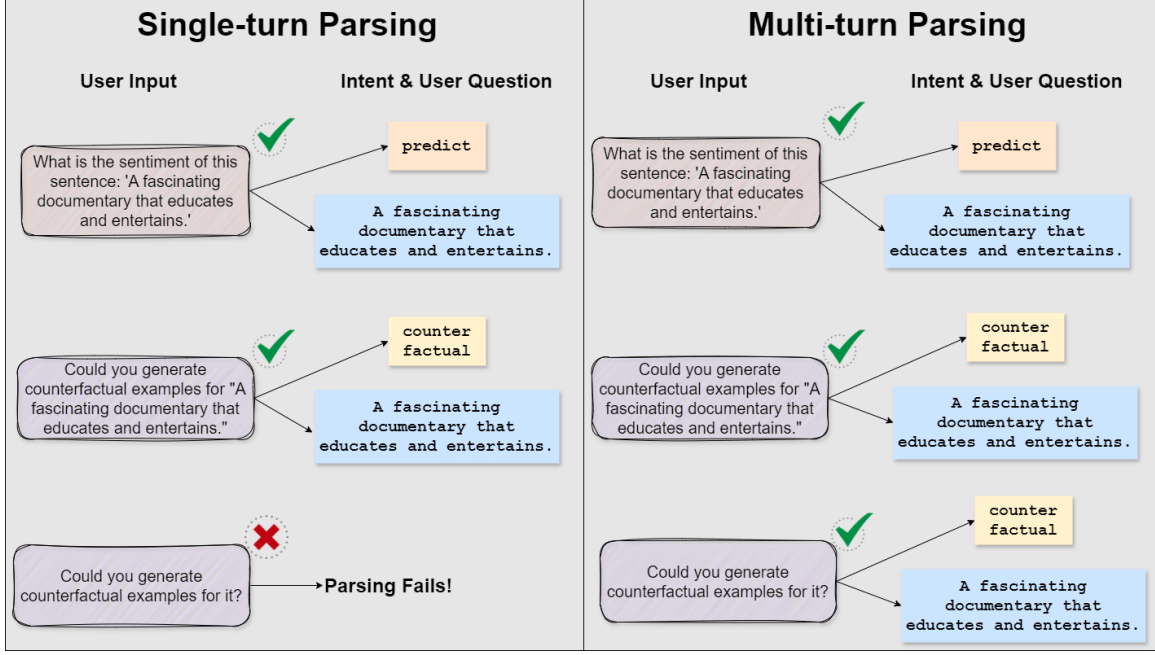


Figure 6: Comparison between single-turn and multi-turn parsing.

many different questions as possible.

Several selection rules are applied. **Bolded questions** are always asked, and all types are always included. In addition, *italicized questions* represent follow-up questions that are always selected when their preceding question appears.

D.2 Instructions and User Interface

Figure 7 shows screenshots of general instructions and informed consent, Figure 8 screenshots of information on the overall study procedure and how to conduct the tasks in the study and Figure 9 shows the interface, participants interact with.

D.3 Results

Across 30 completed sessions, integrating SIMXAI into TALKToMODEL improved both objective task performance and subjective user experience. Unless otherwise noted, we report two-tailed paired *t*-tests over participant-level scores. Table 10 summarizes the objective outcomes, Table 11 reports the BUS-11 item means, Table 12 presents the statistical comparisons, and Table 9 summarizes the pairwise preferences.

Objective performance. As shown in Table 10, SIMXAI substantially improves users' ability to answer explanation-related questions correctly. If we count all cases when users indicated that the question is not answerable ("unable to answer") as incorrect, accuracy increases from 53.3 ± 19.6

with TALKToMODEL to 84.1 ± 17.7 with SIMXAI. The same pattern holds when excluding *unable to answer* responses from the denominator, with accuracy increasing from 59.0 ± 21.3 to 92.1 ± 11.6 . The corresponding paired tests in Table 12 confirm that both improvements are highly significant ($p < .001$ in both cases).

The aggregate counts in Table 10 show the same trend: participants produced many more correct answers with SIMXAI (227) than with TALKToMODEL (138). Importantly, both systems generated the same number of explicit *unable to answer* responses (24 each), suggesting that the gain came primarily from better answers rather than fewer abstentions.

SIMXAI also made interactions more efficient. Compared with TALKToMODEL, it reduced average chat time (7.4 ± 3.3 minutes vs. 5.8 ± 3.0 minutes) and average conversation length (12.4 ± 4.1 turns vs. 10.0 ± 1.8 turns). This pattern is also reflected in the turns-per-correct-answer measure reported in Table 10, which was notably lower for SIMXAI (1.32) than for TALKToMODEL (2.70). Overall, users reached correct answers more quickly and with fewer conversational steps.

Subjective evaluation. The subjective ratings mirror the objective results. Table 11 shows that SIMXAI received higher mean ratings on 10 of the 11 BUS-11 items, and Table 12 shows that 9 of these differences are statistically significant. The

Agent	Speaker	Original	SIMXAI
TalkToModel	User	What if the glucose was 50?	
	Agent	with the glucose level set to 50, it is still likely to have no diabetes	with the glucose level set to 50, it is still likely to have no diabetes.
	User	What if it was 105?	
	Agent	... glucose is set to 50.0 and bmi is set to 105.0 is predicted likely to have diabetes.	with the glucose level set to 105, it is still likely to have no diabetes.
Interrolang	User	What is the prediction for instance ID 81?	
	Agent	The instance with id equal to 81 is predicted True.	The instance with id 81 is predicted True.
	User	How to change the instance to get a different prediction?	
	Agent	This is likely out of my expertise, can I help you with something else?	To get a different prediction, here are the list of words ...
XAgent	User	What is the prediction for this instance	
	Agent	The prediction for this instance is You will not get the loan....	The prediction for this instance is you will not get the loan
	User	Why?	
	Agent	Sorry, I am not exactly sure whether I understood your question correctly. Did you mean any of the below?...	The model utilized all features, ... The following graph depicts key features that influence this prediction...

Table 7: Qualitative results

strongest improvements are observed on dimensions central to multi-turn explanatory dialogue, especially clarity, context retention, helpfulness, and informativeness. For example, users rate SIMXAI higher for clear communication (4.60 vs. 3.33), keeping track of conversational context (4.37 vs. 3.13), and helping them achieve their goal (4.37 vs. 3.10).

The same overall trend holds across the remaining items: SIMXAI is rated as easier to understand, more accurate, and easier to use for finding information. Only two BUS-11 items do not differ significantly. First, privacy communication is rated relatively low in both conditions, indicating that neither system communicated privacy-related issues particularly well. Second, perceived waiting/response time slightly favors TALKTOMODEL numerically, but the difference is not significant. Taken together, the BUS-11 results show that SIMXAI mainly improves the quality and usefulness of the interaction. These aspects are central to ConvXAI.

Pairwise comparison. The direct comparison results in Table 9 further support this pattern. A large majority of participants prefer SIMXAI overall (80.0%), while only 16.7% prefer TALKTOMODEL and 3.3% report no preference. The same distribution appears for perceived task effectiveness: 80.0% judge SIMXAI more helpful, 16.7% prefer TALKTOMODEL, and 3.3% consider the two

systems equally helpful. These preferences are consistent with both the objective performance results and the BUS-11 ratings.

Summary. Overall, the results across Tables 10–9 show a consistent pattern: replacing TALKTOMODEL’s conversational component with SIMXAI improves answer quality, increases task success, reduces interaction cost, and leads to more favorable user judgments. The clearest gains appear on aspects that matter most for explanation-oriented multi-turn interaction, namely clarity, contextual coherence, helpfulness, and efficiency.

Task	Type	Question Template
Explore the dataset and answer the following question(s).	Data Exploration (global)	• Number of features in the dataset is: • Is the feature <feature> included in the dataset? • The maximum value of <feature> recorded in the dataset is: • Find the minimum value of <feature> recorded in the dataset and choose the correct value:
Explore patient with ID <patient_id> and answer the following question(s).	Data Exploration (local)	• Find the value of <feature_A> for this patient and choose the correct value. • What is the value of <feature_A> for this patient: • <i>How about the <feature_B> value of this patient?</i> • <i>How about this patient's <feature_B> value?</i>
Find the prediction of the model for patient with ID <patient_id> and answer the following question(s).	Local Explanation	• What does the model predict for this patient? • The most important feature for the prediction is: • The least important feature for the prediction is: • Is <feature> used for the prediction for this patient? • Top 3 most important features for the prediction are:
	Counterfactual	• What would be the prediction if the value of <feature> is increased to <value>? • What would be the prediction if the value of <feature> is decreased to <value>? • <i>What would be the prediction if it is decreased/increased to <value>?</i>

Table 8: Tasks and corresponding questions used in the user study.

Comparison item	TALKToMODEL	SIMXAI
Overall preference	5 (16.7%)	24 (80.0%)
No preference	1 (3.3%)	
Task effectiveness	5 (16.7%)	24 (80.0%)
Equally helpful	1 (3.3%)	

Table 9: Pairwise comparison results. Most participants preferred SIMXAI and judged it more effective.

Metric	TALKToMODEL	SIMXAI
Participants	30	30
Accuracy (unable = wrong)	53.3 $\pm$ 19.6	84.1 $\pm$ 17.7
Accuracy (excluding unable)	59.0 $\pm$ 21.3	92.1 $\pm$ 11.6
Total correct / answered (incl. unable)	138 / 262	227 / 270
Total correct / answered (excl. unable)	138 / 238	227 / 246
Unable-to-answer count	24	24
Questions responded	265	270
Chat time (min)	7.4 $\pm$ 3.3	5.8 $\pm$ 3.0
Chat turns	12.4 $\pm$ 4.1	10.0 $\pm$ 1.8
Turns per correct answer	2.70	1.32

Table 10: Objective user-study results. SIMXAI improves accuracy and reduces interaction cost. Means are reported with standard deviations.

BUS-11 item	Short Form	TALKToMODEL	SIMXAI
Q1 Chatbot function was easily detectable	Interaction Clarity	3.80 $\pm$ 0.76	4.33 $\pm$ 0.76
Q2 Easy to find information via the chatbot	Information Sufficiency	4.13 $\pm$ 0.78	4.50 $\pm$ 0.57
Q3 Communicating with the chatbot was clear	Interaction Clarity	3.33 $\pm$ 1.27	4.60 $\pm$ 0.56
Q4 Chatbot kept track of conversation context	Context Awareness	3.13 $\pm$ 1.31	4.37 $\pm$ 0.93
Q5 Responses were easy to understand	Response Comprehensibility	3.60 $\pm$ 1.16	4.57 $\pm$ 0.68
Q6 Chatbot helped achieve my goal	Goal Facilitation	3.10 $\pm$ 1.21	4.37 $\pm$ 0.67
Q7 Appropriate amount of information provided	Information Sufficiency	3.23 $\pm$ 1.17	4.37 $\pm$ 0.72
Q8 Only gave needed information	Information Relevance	3.13 $\pm$ 1.31	4.13 $\pm$ 0.94
Q9 Responses were accurate	Response Accuracy	3.63 $\pm$ 0.96	4.47 $\pm$ 0.57
Q10 Informs about privacy issues	Privacy Transparency	2.30 $\pm$ 1.02	2.57 $\pm$ 1.30
Q11 Short waiting / response time	Response Efficiency	4.37 $\pm$ 0.61	4.03 $\pm$ 1.13

Table 11: Post-interaction BUS-11 ratings (1–5 Likert). SIMXAI is rated higher on 10 of 11 items.

Metric	t	p
Accuracy incl. unable	-9.646	< .001***
Accuracy excl. unable	-8.293	< .001***
Q1 Chatbot function was easily detectable	-2.898	.0071**
Q2 Easy to find information via the chatbot	-2.483	.0190*
Q3 Communicating with the chatbot was clear	-5.091	< .001***
Q4 Chatbot kept track of conversation context	-4.570	< .001***
Q5 Responses were easy to understand	-3.778	< .001***
Q6 Chatbot helped achieve my goal	-4.911	< .001***
Q7 Appropriate amount of information provided	-4.335	< .001***
Q8 Only gave needed information	-3.257	.0029**
Q9 Responses were accurate	-3.878	< .001***
Q10 Informs about privacy issues	-1.765	.0881
Q11 Short waiting / response time	+1.439	.1608

Table 12: Paired two-tailed *t*-tests.

Evaluating AI Agents for Explaining an AI Model

Thank you for participating in our study. We appreciate your time and effort!

Investigators & Purpose of Study

You are invited to participate in a research led by <removed>.
This study investigates how AI agents can support users in understanding decisions made by an AI model. It will contribute to a better understanding of how users interact with AI agents and support research on transparent and explainable AI.

Time Required

Participation in this study will require approximately 30 minutes of your time.

Risks

The investigators do not perceive more than minimal risks from your involvement in this study (that is, no risks beyond the risks associated with everyday life).

Participation & Withdrawal

Your participation is entirely voluntary. You are free to choose not to participate. By completing and submitting this anonymous study, you are consenting to participate in this research. Should you choose to participate, you can withdraw at any time without consequences of any kind. However, once your responses have been submitted and anonymously recorded, you will not be able to withdraw from the study. All data (chat interactions and survey responses) will be anonymized, stored securely, and used only for research purposes. Your identity will not be linked to any publications or presentations resulting from this study.

Contact Information

If you have questions about this study, please contact the research team at:

Consent to Participate

By continuing, you confirm that:

- You have read and understood the information above
- You voluntarily agree to participate in this study
- You are 18 years of age or older

Prolific ID:

☐ I have read and understood the above information, and I consent to participate in this study.

[I Consent - Begin Study](#)

(a) General Instructions.

(b) Consent form.

Figure 7: General instructions and consent form (identifying information removed for blind submission).

In this study, you will interact with two AI agents. Each agent explains how an AI model predicts whether a patient has diabetes based on clinical information such as age, body mass index (BMI), and blood glucose levels.

Your goal is to use the explanations provided by the agents to answer a set of task questions.

The study consists of the following steps:

- 1 Tutorial (this page)**
Read how the study works and learn how to use the interface.
- 2 Practice round**
Chat with a Demo Agent and answer 3 practice questions. You must answer at least one question correctly to continue. If none are answered correctly, the study will end.
- 3 Main study**
Interact with two AI agents one at a time and answer task questions using information from the chat.
- 4 Agent surveys**
After interacting with each agent, complete a short evaluation survey about your experience.
- 5 Final survey**
Provide your overall comparison of the two agents.

Your task
Use the chat interface to ask the AI agent questions and gather information. Then answer the task questions as accurately as possible.

How it works

- **Read the task and the question carefully.**
- **Ask the agent questions that help you find the correct answer.**
You may ask as many questions as you like. If the agent does not provide the information you expect, try rephrasing your question.
- **Select your answer and click "Submit".**
Once you have interacted with the agent, the Submit button will appear below the question.
- **Move on to the next question.**
The next task will appear after you submit your answer.
- **If you cannot find the answer, or if it takes too long, please choose "Unable to answer" instead of guessing.**

☐ I understand the interface and I'm ready to begin

[← Back](#)

[Start Practice Round with Demo Agent →](#)

(a) Study overview.

(b) Task instructions.

Figure 8: Information on the study procedure and how to conduct tasks.

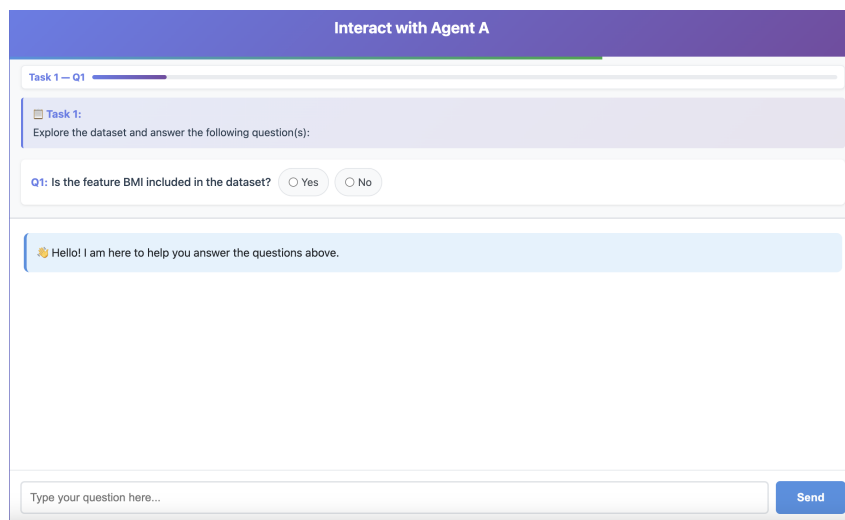


Figure 9: User interface for conversational XAI agent interaction and question answering.