

# Evaluating the Reliability of Large Language Model-Generated Explanations in Dialogue-Based XAI

Isabel Feustel<sup>1,\*</sup>, Niklas Rach<sup>2</sup>, Wolfgang Minker<sup>1</sup> and Stefan Ultes<sup>3</sup>

<sup>1</sup>Ulm University, Albert-Einstein-Allee 43, 89081 Ulm, Germany

<sup>2</sup>Tensor AI Solutions GmbH, Magirus-Deutz-Straße 2, 89075 Ulm, Germany

<sup>3</sup>University of Bamberg, 96045 Bamberg, Germany

## Abstract

Natural language generation (NLG) for explainable AI (XAI) is an increasingly important area of research. This is particularly true given the growing availability of large language models (LLMs) that can produce natural explanations. While recent efforts have focused on using LLMs to generate narrative-style explanations of machine learning predictions, these are often designed as static, one-shot outputs rather than context-aware contributions to dialogue. In this study, we explore the use of LLMs for generating faithful and contextually appropriate explanations within interactive explanatory dialogues. Using dialogue snippets from a recent user study, we construct diverse prompts covering multiple explanation types and evaluate within a human annotation (n=20) the generated outputs in terms of coherence, plausibility, usefulness, and factual correctness. Our results show that LLM-generated explanations significantly outperform template-based responses in coherence, plausibility, and usefulness. However, perceptions of factual correctness vary significantly by annotator expertise: expert raters judged LLM outputs as less factually accurate than template-based responses, while lay annotators rated them as more factually accurate. This discrepancy underscores the influence of background knowledge on perceived explanation reliability. We further assess the potential of automatic evaluation metrics and discuss their potential role in supporting reliable NLG for dialogue-based XAI systems.

## Keywords

Conversational XAI, Explainable AI, Large Language Models, Dialogue Systems, Natural Language Generation, Human-Centered Evaluation

## 1. Introduction

Conversational explainable artificial intelligence (XAI) is an emerging field that aims to make AI models more accessible and transparent by offering user-centered, interactive explanations [1, 2, 3]. Dialogue-based explanations are especially valuable, as they allow users to ask follow-up questions, request clarification, and receive information in manageable portions tailored to their needs [3, 4, 1].

Large language models (LLMs) are promising tools for enabling such interactions, as they can produce fluent and contextually relevant responses in natural language. Recent work has demonstrated the potential of using LLMs to generate explanatory narratives about machine learning models [5, 6]. A major concern with LLM-generated explanations is their reliability—particularly their factual correctness [7, 8]. In interactive exchanges, users rely on precise and trustworthy responses. Misleading or incorrect explanations may compromise users' understanding and trust in the system [8].

In this work, we evaluate the reliability of LLM-generated explanations in the context of dialogue-based XAI. Specifically, we examine how such explanations perform in response to typical user prompts during real explanatory dialogues. We use dialogue snippets from a previous user study and compare the original, template-based system responses with newly generated LLM explanations produced via

MAI-XAI@ECAI'25: Workshop on Multimodal, Affective and Interactive eXplainable AI, October 25–31, 2025, Bologna, Italy

\*Corresponding author.

✉ isabel.feustel@uni-ulm.de (I. Feustel); niklas.rach@tensor-solutions.com (N. Rach); wolfgang.minker@uni-ulm.de (W. Minker); stefan.ultes@uni-bamberg.de (S. Ultes)

🌐 <https://tensor-solutions.com/> (N. Rach); <http://www.dialogue-systems.org/> (W. Minker); <https://www.uni-bamberg.de/ds/> (S. Ultes)

🆔 0009-0004-9161-6843 (I. Feustel); 0000-0001-9737-8584 (N. Rach); 0000-0003-4531-0662 (W. Minker); 0000-0003-2667-3126 (S. Ultes)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the GPT API. A human annotation study involving experts and laypeople evaluates both answer types along four dimensions: coherence, plausibility, factual correctness, and usefulness. Through this study, we aim to answer the following research questions:

- **RQ1:** How reliable are LLM-generated explanations across explanation types and user groups?
- **RQ2:** Are some explanation types more prone to hallucinations or misrepresentation?

The remainder of this paper is structured as follows: Section 2 reviews related work on conversational XAI and LLM-based explanation generation. Section 3 describes the basis of our work, including the origin of the dialogue data and our prompt design strategy. Section 4 outlines the evaluation design, detailing the study methodology and annotation procedure. Section 5 presents the results of the human annotation study. Section 6 discusses key findings, challenges, and future directions. Finally, Section 7 concludes the paper.

## 2. Related Work

The growing interest in XAI has led to research across multiple domains, from explanation delivery through dialogue to ensuring explanation fidelity and evaluating the reliability of natural language generation (NLG). This section reviews the literature across three core areas relevant to our work: (1) Dialogue-based XAI systems, (2) LLM-generated explanations and their evaluation, and (3) Hallucination and faithfulness in language models.

### 2.1. Dialogue-Based XAI Systems

Conversational interfaces have emerged as a promising paradigm for delivering AI explanations in a user-centered manner [3, 1]. Systems such as ConvXAI [9] and TalkToModel [10] employ template-based NLG to maintain faithfulness and minimize hallucinations. Similarly, XAgent [11] relies on a simple prompting strategy that instructs the model not to alter the factual content to avoid misrepresentation in dialogue explanations. Recent comparative work has examined the impact of dialogue formats on user perception. He et al. [12] evaluated rule-based and LLM-powered conversational XAI interfaces against dashboards, finding that while conversational systems improved understanding and trust, they also risked increasing user over-reliance. These findings underscore the importance of balancing interactivity with reliability in explanation design. Mindlin et al. [1] emphasize that the highest objectives for dialogue-based XAI are interactivity and trustworthiness, underscoring the need for systems that not only communicate clearly but also adapt to users’ needs while remaining faithful to the underlying model logic.

### 2.2. LLM-Generated Narrative Explanations

Other work focuses on using LLMs to create rich, narrative-style explanations of machine learning predictions. For instance, SHAPStories and CFStories proposed by Martens et al. [5] use GPT-4 to generate explanatory narratives that improve user satisfaction compared to standard SHAP [13] or counterfactual explanations [14], particularly for lay users. However, these explanations are designed as static blocks and are not tailored to dialogue flow or context, limiting their applicability to interactive systems. EXPLINGO [6] introduces a two-part framework that combines a narrative generation module with an automated grading system. The approach emphasizes fluency, conciseness, and completeness in generated explanations, focusing on narrative-style outputs similar to SHAPStories, but lacks human-grounded validation. Expanding on this direction, Zyte et al. [15] propose that LLMs can enhance XAI by translating complex model outputs into natural language for users. The authors emphasize the importance of context-sensitive explanation design and identify open questions around prompt formulation and user-centric evaluation.

Dialogue History

Collapse History

USER

Am I at risk of getting diabetes?

SYSTEM

All right, let's have a look. In your case, the most influential value was glucose , as you can see in the figure below. You can also see the influence of all other values, ordered by their importance. Purple bars indicate a positive contribution, blue bars a negative contribution. Do you want to know more about why glucose is relevant?

USER

yes

SYSTEM

Glucose level plays a pivotal role in diabetes risk assessment as it directly reflects the body's ability to regulate blood sugar, a process that is fundamentally disrupted in both prediabetes and diabetes.

USER

how does sugar intake affect glucose levels?

SYSTEM

Monitoring glucose levels is crucial for identifying individuals at risk and for diagnosing diabetes.

USER

how do you monitor levels?

Template Response

SYSTEM

Testing helps doctors find diabetes sooner, so they can work with their patients to manage the disease and prevent health problems.

LLM Response

SYSTEM

Glucose levels are typically monitored through blood tests, either at your doctor's office or with home glucose meters. Regular testing helps doctors detect diabetes early, so they can help you manage it and avoid health complications. This monitoring is especially important since early detection supports prompt intervention.

**Figure 1:** Example dialogue excerpt shown during the study. The top section displays the preceding user–system interaction. Below, two alternative system responses are presented: the original template-based response (left) and the LLM-generated response (right).

### 2.3. Hallucination and Explanation Faithfulness

One of the key challenges in integrating LLMs into XAI systems is ensuring that the generated explanations are faithful to the model and factually accurate. A comprehensive survey by Huang et al. [16] categorizes hallucination types and discusses mitigation strategies relevant to XAI. They highlight that faithfulness is not just a technical issue but a communicative one, especially in safety-critical applications. The importance of grounding explanations in verifiable facts is further demonstrated in domain-specific approaches such as Zhang et al. [17], which use synthetic training data paired with hallucination detection mechanisms to fine-tune models for visual XAI tasks. In the context of persona-driven interactions, Jandaghi et al. [18] introduce Synthetic-Persona-Chat, a framework that generates dialogues consistent with user profiles. This work provides valuable insights for ensuring coherence and factual consistency in multi-turn conversations, a property critical for reliable dialogue-based XAI systems. Finally, Yang et al. [19] propose ExplainGen, an LLM-based approach trained on misinformation detection datasets to generate explanations that help users assess credibility. While highly relevant for factual consistency, the work is oriented toward news verification rather than ML model explanations. Together, these works illustrate both the promise and limitations of current LLM-based explanation systems. While conversational and narrative formats make AI explanations more accessible, their effectiveness depends on how well they balance fluency, contextual fit, and factual accuracy.

### 3. Natural Language Generation for dialogue-based XAI

This section provides the foundations for our study by detailing the source of the dialogue data and explaining the design of our prompting approach. Our goal is to evaluate the reliability of LLM-generated explanations in realistic, user-facing scenarios by building upon authentic interaction data.

#### 3.1. Data Source and Dialogue Context

As emphasized by Sokol and Vogt [8], evaluating XAI in realistic, application-grounded settings is essential to understanding its real-world impact. Explanations must be assessed not in isolation, but within the context of user needs, decisions, and domain-specific interactions. To this end, we base our study on the system and user data introduced by Feustel et al. [20], which focused on integrating domain knowledge into dialogue-based XAI systems. The original system supported two real-world use cases: credit loan approval and diabetes risk assessment. In both cases, users interacted with a predictive AI model and a dialogue-based explanation agent. The agent supported free-form user queries, providing answers generated from predefined templates. The underlying explanations leveraged several common XAI techniques:

- **Feature Importance (SH):** Quantifies the contribution of each input feature to the model’s prediction using SHAP [13].
- **Counterfactual Explanations (CF):** Describe how small changes in input would alter the model’s decision [14].
- **Example-Based Explanations (EX):** Retrieve similar cases to help users contextualize decisions [21].
- **Domain Knowledge-Based Explanations (DK):** Use structured external knowledge (text) to offer supportive or critical arguments related to the decision [22].

The original study was conducted online using a crowd-sourced participant pool of native speakers. The interactions recorded during this study offer a rich set of authentic dialogue turns grounded in real model outputs and user queries. For our purposes, we extracted representative snippets from these dialogues (see Figure 1), focusing on user questions and corresponding template-based system responses. This allowed us to construct controlled but realistic prompt-response pairs for generating LLM-based explanations.

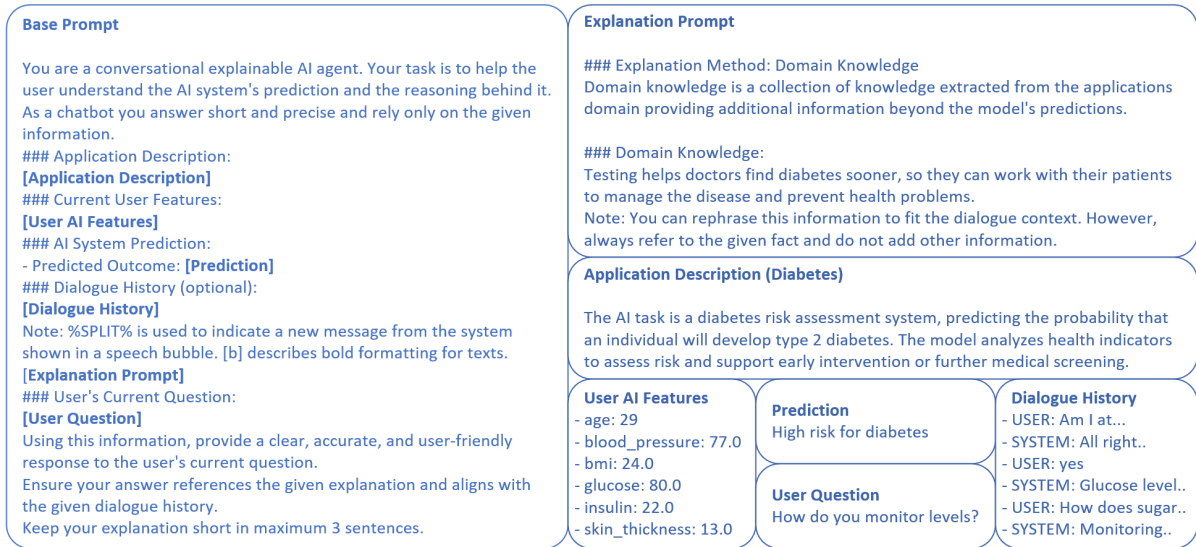
#### 3.2. Prompt Design

The goal of our prompt design process was to replace template-based responses with LLM-generated explanations while preserving the integrity and context of the original dialogue.

Each prompt included the following components (see Figure 2):

- **AI Prediction Context:** A brief task-specific summary describing the use case (credit or diabetes), the user’s input data, and the model’s prediction.
- **Original XAI Explanation:** A textual summary of the explanation type used in the original system (e.g., SHAP for feature importance) and the explanation content itself.
- **Dialogue History:** If applicable, one or more previous dialogue turns to preserve conversational flow and reference context.
- **User Question:** A free-form natural language query as posed in the original dialogue.

To guide the LLM’s output and ensure reliable, grounded responses, we embedded a set of explicit behavioral instructions within each prompt. These instructions specified the role of the model and constrained its behavior to align with the original system’s objectives. Specifically, the model was instructed to act as an AI assistant responding to user questions about a machine learning prediction. It was told to rely solely on the information provided in the prompt, to avoid inventing new data or



**Figure 2:** Example prompt structure used for generating LLM responses. The base prompt (left) contains template instructions with placeholders, while the right column shows the specific insertions for a particular instance (highlighted in bold in the base prompt). The explanation prompt varied depending on the explanation type. The full dialogue history for this example is shown in Figure 1.

explanations not supported by the given context, and to formulate its responses in a helpful and conversational tone. These constraints were essential for promoting consistency, minimizing hallucinations, and simulating a realistic explanatory dialogue. The resulting prompts were submitted to GPT-4.1 via the OpenAI API<sup>1</sup> to obtain model answers, which then formed the basis for our evaluation in the subsequent user study.

## 4. Evaluation

In this section, we describe the design and implementation of our user study to evaluate the reliability of LLM-generated explanations in dialogue-based XAI. We present our study design, outline the data generation process, and define the evaluation criteria.

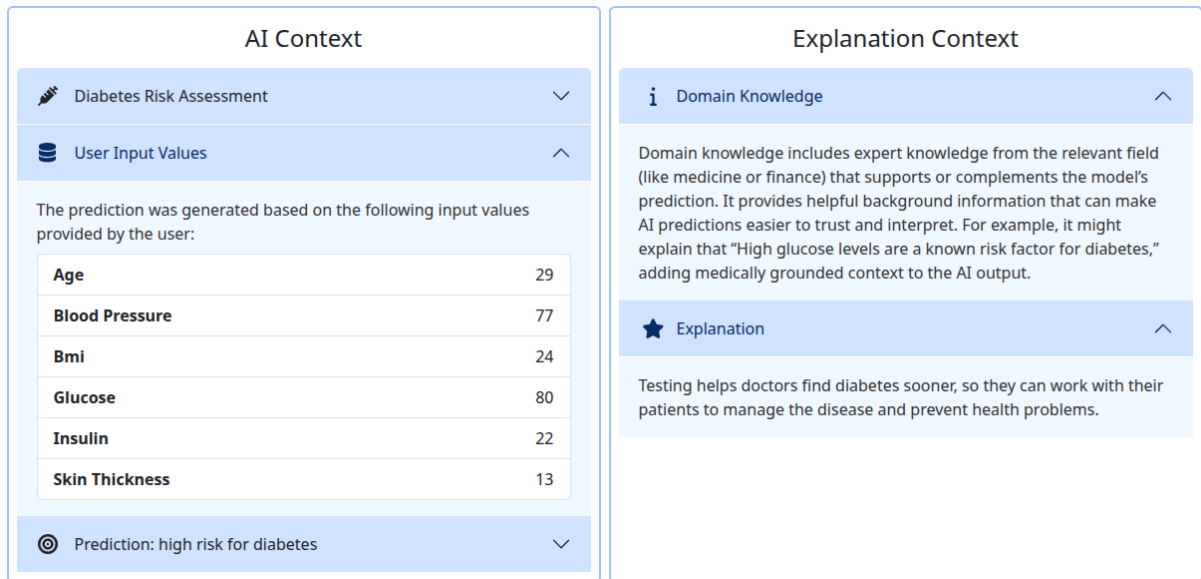
### 4.1. Study Design

To address our first research question (RQ1), we designed a human annotation study comparing template-based system responses with explanations generated by a LLM within realistic dialogue scenarios. The evaluation focuses on user judgments of explanation responses embedded in actual dialogue contexts, as illustrated in Figure 1. For each evaluated snippet, participants were shown the AI prediction and the surrounding conversational context (Figure 3) and asked to assess the appropriateness of both explanation variants. We chose a human-centered annotation approach due to the known limitations of automatic evaluation metrics in dialogue systems. Automatic methods often fail to capture the nuanced ways in which users perceive explanation quality and may not align well with human judgment [23]. This is especially critical in interactive, contextual settings such as dialogue-based XAI. Moreover, we were particularly interested in how user background influences evaluation. Prior research has shown that a user's knowledge significantly affects their ability to simulate model behavior and make decisions based on AI output [24]. Therefore, we included two distinct annotator groups: AI experts and lay users. This allowed us to compare reliability perceptions between these groups and explore how expertise modulates trust in LLM-generated explanations.

To operationalize the evaluation, we adopted four core annotation dimensions:

<sup>1</sup>accessed on 2025-04-14





**Figure 3:** Context information provided to study participants. The left panel displays the AI prediction context, including the application scenario (here: diabetes risk assessment), user input values, and model output (here: high risk). The right panel presents the explanation context, featuring a brief introduction to the explanation type (here: domain knowledge) followed by the actual explanation (here: an argument about glucose testing).

- **Coherence:** The explanation should be logically and linguistically integrated into the dialogue. Coherence reflects whether the explanation fits naturally within the conversational turn, maintaining contextual and grammatical consistency.
- **Plausibility:** The explanation should appear reasonable and believable. Even if not factually perfect, an explanation that aligns with human expectations can improve user satisfaction [7, 25, 26, 27].
- **Factual Correctness:** The explanation should be grounded in the provided data and match the logic of the underlying explanation type. Distinguishing this from plausibility is critical, as an explanation may appear believable yet still be factually incorrect [25].
- **Usefulness:** The explanation should help users better understand the AI system's decision. Importantly, it should also address the user's specific question or concern within the dialogue [1].

Each dimension was rated based on a specific statement, as listed in Table 1, using a 5-point Likert scale ranging from Strongly Disagree (1) to Strongly Agree (5). In addition to these four dimensions, we collected a subjective measure of system preference, asking participants which of the two responses—template or LLM-generated—they preferred overall for each dialogue instance.

To answer RQ2, we sampled dialogue snippets involving four distinct explanation types: SHAP (SH), counterfactuals (CF), example-based explanations (EX), and domain knowledge (DK), as described in Section 3.1. For each dialogue context, we extracted relevant elements including user questions, AI prediction results, and the original system explanation. Using these components, we generated new explanatory responses with GPT-4o. GPT-4o was selected for its superior performance across textual reasoning tasks, strong multi-modal capabilities, and widespread recognition in XAI research [5, 28].

In total, we created 20 LLM-generated responses, five for each explanation type. Each generated explanation adhered to the prompting strategy described in Section 5.

## 4.2. Hypotheses

To investigate how users perceive the reliability of LLM-generated explanations in comparison to template-based outputs, we formulated a set of hypotheses. Our study distinguishes between two user groups:

Rating Category	Statements
Coherence	The explanation flows logically and fits well within the context of the dialogue.
Plausibility	The explanation appears reasonable and believable, given the context.
Usefulness	The explanation helps you better understand the AI’s decision or reasoning.
Factual Correctness	The information is accurate and grounded in the given context and data.

**Table 1**

Rating categories and corresponding evaluation statements used in the user study. Each category was assessed using a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

- **Experts** (e.g., AI researchers, XAI practitioners): These participants are expected to have a deeper understanding of machine learning principles and XAI techniques. As a result, we anticipate that experts will be more attuned to factual inaccuracies and more critical in their assessment of explanation content.
- **Laypeople**: This group represents typical end users with limited technical background. Prior research suggests that laypeople are more likely to be influenced by linguistic fluency and surface-level plausibility, making them potentially more susceptible to over-reliance on convincing yet potentially inaccurate explanations [12].

Based on these considerations, we propose the following hypotheses:

**H1:** *LLM-generated explanations will be rated higher than template-based explanations in terms of coherence, plausibility, and usefulness.*

This hypothesis reflects the general expectation that LLMs produce more natural and fluent responses, leading to improved user perceptions along these subjective dimensions.

**H2:** *LLM-generated explanations will be rated lower in factual correctness compared to template-based responses.*

While LLMs excel at generating fluent text, prior work has raised concerns about hallucinations and factual drift [16]. Template-based responses, though often rigid, are grounded in predefined logic and data, potentially making them more reliable from a factual standpoint.

**H3:** *Experts will rate LLM responses as less factually correct than laypeople.*

We expect a divergence in factuality judgments based on expertise. Experts are more capable of detecting subtle inaccuracies or misrepresentations in LLM outputs, while lay users may conflate fluency with correctness.

**H4:** *Factual correctness ratings will vary across explanation types.*

This hypothesis addresses RQ2 and reflects the assumption that some explanation types are more prone to hallucinations or misinterpretation in generative models than others.

These hypotheses guide our evaluation approach and provide the basis for interpreting participant ratings in the subsequent analysis.

### 4.3. Study Procedure

The evaluation was conducted as an online questionnaire, beginning with a brief introduction, followed by a detailed annotation guideline that outlined key instructions to ensure consistent and objective ratings. Annotators were explicitly instructed to:

- Use only the information provided in the AI context, dialogue history, and explanation section.
- Avoid relying on personal knowledge or assumptions when assessing the explanations.
- Consider any explanation content that includes information not present in the provided context as incorrect, regardless of its plausibility.

	Rating Category	All	CF	DK	EX	SH
Experts	Coherence	0.466	0.5	0.81	0.743	-0.195
	Factual Correctness	0.018	-0.188	-0.198	-0.064	0.295
	Plausibility	0.336	0.229	0.618	0.543	-0.113
	Usefulness	0.556	0.514	0.716	0.705	0.23
Laypeople B3	Coherence	0.492	0.693	0.569	0.712	0.232
	Factual Correctness	0.036	0.405	-0.029	0.091	0.131
	Plausibility	0.375	0.537	0.305	0.591	0.293
	Usefulness	0.421	0.772	0.571	0.584	0.109

**Table 2**

Inter-rater agreement (Krippendorff’s  $\alpha$ ) for expert and lay annotators across all rating categories, shown for all explanation types combined (All) and individually for each type: counterfactuals (CF), example-based (EX), domain knowledge (DK), and Shapley values (SH). Agreement scores are based on the three most consistent annotators (B3) per group and dimension.

- Maintain objectivity and apply the same judgment criteria across all examples.

To aid understanding and support reliable evaluation, participants were also provided with a page containing annotated example ratings that illustrated how to apply the guidelines. Both the guideline and the examples were accessible throughout the entire study, ensuring participants could revisit them at any point. Each participant rated 20 dialogue instances, each of which included two alternative system responses: one template-based and one LLM-generated. The order of dialogues and responses was randomized to avoid bias based on explanation type or system identity. The participant pool consisted of two groups: 9 domain experts (including AI researchers and XAI practitioners from academic and industry settings) and 11 laypeople recruited from the authors’ personal networks. However, due to technical issues, such as non-functional SHAP graphs (not displayed to participants) and occasional missing ratings from faulty form validation, a subset of annotations was incomplete. To ensure full coverage, a selective re-annotation strategy was employed in which additional raters evaluated specific explanation types (e.g., only DK or SH dialogues). In total, 3 experts and 4 laypeople completed all 20 dialogues without omissions. With the re-annotation strategy, we a minimum of four raters per dialogue. Thus, while not every participant annotated the full set, each dialogue instance received multiple independent ratings (see Table 3 for exact counts).

We analyzed the resulting ratings using the Mann-Whitney U test [29] to assess statistical significance between system responses and groups. To evaluate the consistency of ratings across annotators, we computed Krippendorff’s alpha [30, 31], focusing on agreement within each user group and rating category.

## 5. Results

In this section, we present the results of our evaluation. We begin by analyzing inter-annotator agreement to assess the consistency of participant ratings. This is followed by a detailed examination of our study’s core hypotheses based on the collected data. A broader discussion and interpretation of these findings will be provided in the subsequent section.

### 5.1. Annotator Agreement

To assess the reliability of participant ratings, we calculated inter-annotator agreement using Krippendorff’s  $\alpha$  [30], considering only complete annotation sets and including the three most agreeing annotators per user group and rating category. Table 2 presents agreement scores across all rating dimensions—coherence, plausibility, usefulness, and factual correctness, separately for expert and lay participants. The table also distinguishes between aggregated results over all explanation types



Expl. Type	Rating Category	Experts					Laypeople			
		$n_T$	$n_{LLM}$	Templ.	LLM	$p$	$n_T/n_{LLM}$	Templ.	LLM	$p$
All	Coherence	125	125	3.496	4.568	<b>&lt;0.01</b>	140	2.957	4.329	<b>&lt;0.01</b>
All	Factual Correctness	124	125	4.54	3.816	<b>&lt;0.01</b>	140	3.586	3.786	0.42
All	Plausibility	125	124	3.736	4.492	<b>&lt;0.01</b>	140	3.286	4.257	<b>&lt;0.01</b>
All	Usefulness	125	124	2.936	4.194	<b>&lt;0.01</b>	140	2.493	3.936	<b>&lt;0.01</b>
CF	Factual Correctness	39	40	4.487	3.675	<b>&lt;0.01</b>	35	3.171	3.771	0.104
EX	Factual Correctness	40	40	4.6	4.2	<b>0.013</b>	35	3.171	3.8	0.129
DK	Factual Correctness	25	25	4.52	4.6	0.89	35	4.086	4.286	0.677
SH	Factual Correctness	20	20	4.55	2.35	<b>&lt;0.01</b>	35	3.914	3.286	0.093

**Table 3**

User study results showing average ratings scores and number of ratings ( $n_T$  for template,  $n_{LLM}$  for LLM) across the four rating categories for both user groups (experts and laypeople) across all explanations (All). Additionally, factual correctness ratings are broken down by explanation type: Counterfactuals (CF), domain knowledge (DK), example-based (EX), and Shapley (SH). Template-based (Templ.) and LLM-generated responses are compared with p values from Mann–Whitney U tests. Significant differences are highlighted.

(ALL) and results for individual explanation types: counterfactuals (CF), domain knowledge (DK), example-based (EX), and feature importance (SH).

Among expert annotators, we observe moderate agreement for coherence and usefulness when aggregated across all explanation types. Agreement for plausibility is weaker, and ratings for factual correctness show substantial disagreement. When broken down by explanation type, domain knowledge and example-based explanations yield stronger agreement on coherence, plausibility, and usefulness. In contrast, counterfactual explanations exhibit weaker plausibility agreement, while SHAP explanations display a reversal in trend, higher agreement on factual correctness, but lower on other dimensions.

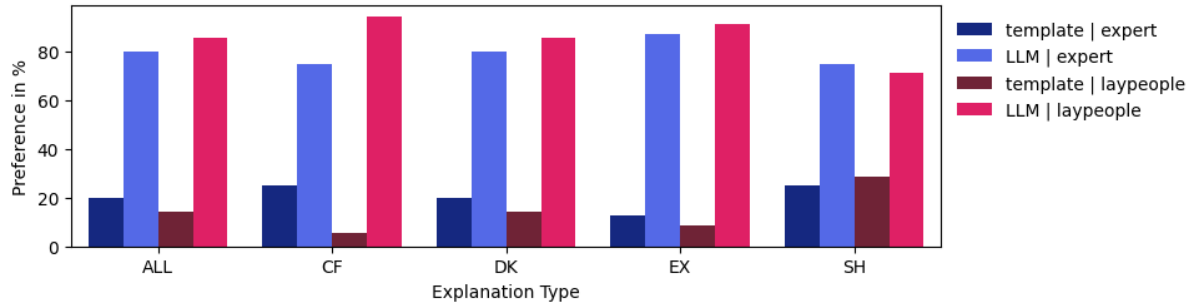
Lay annotators show a similar pattern overall, with moderate agreement in coherence and usefulness, and lower consistency for plausibility. Notably, factual correctness shows minimal to no agreement across most explanation types, with the exception of counterfactuals, where some alignment is observed. SHAP explanations again stand out with the lowest agreement in this group, particularly in factual ratings.

These results underscore the challenges of evaluating explanation quality, especially regarding factual correctness in dialogue contexts, where short, conversational responses may obscure factual grounding and make inconsistencies harder to detect.

## 5.2. Reliability of LLM-Generated Explanations

This section evaluates the reliability of LLM-generated explanations compared to template-based responses, based on the hypotheses outlined in Section 4. Table 3 summarizes the average user ratings across the four evaluation dimensions—coherence, plausibility, usefulness, and factual correctness—separated by explanation type and user group. Statistical significance was assessed using the Mann-Whitney U test [29]. In the following, we examine each hypothesis in turn and discuss the extent to which the results support or contradict our expectations.

**H1: LLM-generated responses will be rated higher than template-based responses in coherence, plausibility, and usefulness.** This hypothesis is supported by the results. In both user groups, LLM-generated responses significantly outperformed template-based answers across all three categories ( $p < 0.01$ ). Coherence and plausibility were consistently rated higher, with LLMs offering more fluid and contextually fitting responses. While expert and layperson trends aligned, laypeople rated the template-based responses notably lower overall.



**Figure 4:** Participant preference ratings (%) for system responses (Template vs. LLM), separated by user group (experts in blue, laypeople in red) and explanation type (ALL = all explanations, CF = counterfactual, DK = domain knowledge, EX = example-based, SH = feature importance with SHAP). Bars indicate the proportion of preferences within each group.

**H2: LLM-generated responses will be rated lower in factual correctness compared to template-based responses.** This hypothesis holds for expert users: factual correctness scores were significantly higher for template-based responses than LLM-generated ones ( $p < 0.01$ ). This aligns with concerns around LLM faithfulness. However, this result must be interpreted cautiously due to the lower inter-annotator agreement on factual correctness ratings. Another finding was that laypeople rated LLM-generated explanations as more factually accurate on average, thus contradicting the prevailing expert trend. The aforementioned opposing views lend support to H3.

**H3: Experts will rate LLM responses as less factually correct than laypeople.** This hypothesis is tentatively supported by observed trends in the data. Expert raters consistently assigned higher factual correctness scores to template-based responses across explanation types (often exceeding 4.5) and were more critical of LLM-generated outputs, particularly for feature importance explanations, where the average rating dropped to 2.35 compared to 3.2 from lay participants. In contrast, laypeople generally rated LLM responses as more factually accurate, with scores ranging from 3.2 to 4.2. These group differences highlight the influence of domain expertise on perceptions of factual reliability.

**H4: Factual correctness ratings will vary across explanation types.** Despite low agreement levels for this dimension, the results show substantial variation across explanation types and thus support the hypothesis. In both groups, domain knowledge-based explanations showed consistent factual correctness between LLM and template responses. However, for expert participants, LLM responses were rated significantly lower in factual correctness for counterfactuals, example-based explanations, and especially feature importance (SH) explanations. SH explanations, in particular, were prone to errors in numerical interpretation and attribution. For lay participants, there was no significant difference overall between template and LLM responses, although ratings still varied across explanation types.

### 5.3. User Preferences for XAI Explanations in Dialogue

Figure 4 illustrates participants' preferences for system responses, separated by user group and explanation type. Overall, LLM-generated responses were preferred more frequently than the template-based ones across both groups. This aligns with earlier findings on coherence, plausibility, and usefulness, where LLM responses consistently scored higher. Interestingly, even in cases where the LLM explanation exhibited factual inconsistencies, such as with feature importance explanations, lay participants still showed a preference for the LLM response over the template.

## 6. Discussion and Future Directions

In this section, we reflect on the findings presented in the previous chapter and explore their implications for the design and deployment of dialogue-based XAI systems. Our results suggest that LLMs hold strong potential for generating natural and engaging explanations, particularly appealing to non-expert users. However, this fluency comes with notable tradeoffs: while LLM-generated responses are often preferred, their factual consistency can be unreliable, especially in the absence of verifiable grounding. These observations underscore a key design tension in XAI: balancing naturalness and user-friendliness with the need for accurate, transparent, and trustworthy explanations. We outline future directions that aim to mitigate these risks while leveraging the strengths of LLMs in interactive explanatory contexts.

### 6.1. Insights and Challenges from Annotation Agreement

The results of our inter-annotator agreement analysis offer valuable insights into the inherent complexity of evaluating dialogue-based explanations. While agreement was stronger in dimensions such as coherence and usefulness, lower scores in factual correctness reflect the nuanced and subjective nature of interpreting brief conversational explanation. In particular, the lower agreement on factual correctness aligns with prior observations that this dimension is especially difficult to judge, even for experts [25, 26, 32]. The short and paraphrased nature of responses, combined with a lack of explicit verification cues (i.e., additional contextual information outside the given AI explanation that could help raters confirm correctness), adds to this complexity. Nonetheless, these challenges underscore the importance of designing evaluation methodologies that reflect realistic user interactions, where subjectivity and interpretation play key roles. Our findings emphasize the need for carefully tailored annotation strategies moving forward. Providing clear explanation scaffolds, adjusting rating scales, or offering binary or percentage-based factuality judgments could be investigated in the future to improve consistency. Moreover, matching the task complexity to annotator expertise and offering modular annotation tasks may ease cognitive load and enhance data quality. Nevertheless, our study demonstrates that meaningful evaluation of LLM-generated dialogue responses is both feasible and informative. The patterns observed across user groups and explanation types point to significant trends in system reliability and user perception. As dialogue-based XAI continues to grow, such nuanced studies will be key to developing adaptive, user-sensitive, and verifiable explanation systems.

### 6.2. Reliability and Explanation Type Sensitivities in LLM-Generated Explanations

The results of Table 3 suggest that while LLMs offer clear advantages in fluency and dialogical appropriateness, their reliability, particularly in terms of factual correctness, remains uneven and dependent on both the explanation type and the expertise of the evaluator. This tension between naturalness and factual precision is central to dialogue-based XAI and underlines the risks of using LLMs in high-stakes or user-facing applications without appropriate safeguards.

The type of explanation provided appears to significantly influence how reliably the LLM can recreate accurate content. For instance, domain knowledge-based explanations were rated similarly in factual correctness for both LLM and template responses. This is a promising signal, likely reflecting the LLM’s strength in text-based reasoning and its ability to reproduce or rephrase structured declarative information when it is explicitly embedded in the prompt. Since domain knowledge was given as clear textual input, the LLM had less opportunity, or need to infer or fabricate content.

In contrast, explanations involving SHAP (SH) exhibited pronounced issues with factual reliability. Here, the explanations were generated based on a simplified bar chart representation of feature importance, which aimed to enhance interpretability for lay users by flipping the value axis to match prediction polarity (e.g., positive bars toward a “good” prediction). This transformation, while user-centered, introduced confusion for the LLM. Despite being instructed on the directionality, the model occasionally reverted to default assumptions (e.g., “blue bars are negative”), indicating a failure to internalize prompt-specific graph semantics. These misinterpretations led to the lowest factual correctness ratings in the both groups. The result emphasizes the limitations of text-only prompting for

explanations rooted in visual artifacts. Providing the original visualizations, possibly via multimodal LLMs, or using structured prompts that explicitly describe graph axes, could improve performance in such cases. Lay participants rated the LLM-generated SH explanations poorly, likely because the system responses did not include explicit references to the accompanying graph. In contrast, the template responses contained phrases such as “as you can see in the graph,” which anchored the explanation more clearly to the visual context provided in the interface. This visual aid helped participants better understand and evaluate the explanations, particularly in identifying key contributing features.

Similarly, domain knowledge explanations also received relatively high factual correctness ratings, suggesting that structured textual information was accessible to lay users. Since both counterfactual (CF) and example-based (EX) explanations involved numeric representations, these results indicate that clear visual or textual supports can significantly enhance non-experts’ ability to assess the factual accuracy of system responses. This highlights the value of aligned visual aids in bridging comprehension gaps for lay users, especially when dealing with abstract or data-driven content.

Counterfactual and example-based explanations presented another layer of complexity. While lay participants’ ratings for both explanation types remained relatively consistent across systems, assigning similar scores to template and LLM-generated responses, experts showed more nuanced differences. In particular, expert evaluators penalized LLM-generated counterfactual explanations more severely than example-based ones. Template responses in these cases tended to offer concise contrasts or relatable scenarios that aligned more directly with the underlying model logic. In contrast, the LLM responses, while more linguistically fluent, occasionally introduced implicit assumptions or failed to preserve essential counterfactual conditions. This suggests that although LLMs are capable of producing well-phrased outputs, preserving factual alignment, especially in abstract or hypothetical reasoning tasks like counterfactuals, remains a notable challenge.

Together, these patterns underscore the need for careful prompt engineering and explanation-type awareness when deploying LLMs in XAI systems. What works well for one explanation form may fail for another, particularly if key visual, numerical, or contextual cues are misrepresented or under-specified. These shortcomings also raise practical questions about how much auxiliary explanation data, such as metadata, schema guidance, or fine-tuned examples, LLMs require to generate factually sound outputs consistently. Moreover, although this study relied on a single LLM to manage cognitive load for annotators, future research could explore comparative evaluations across multiple models. Leveraging recent strategies such as few-shot prompt tuning, pattern-guided generation (e.g., ReAct [33]), or structured planning (e.g., Tree-of-Thoughts [34]) may help increase factual reliability. Additionally, employing LLMs as judges, calibrated against high-quality human annotations, could offer scalable alternatives for reliability testing, especially in domains where fine-grained factual evaluation is burdensome.

### 6.3. User Preferences

As illustrated in Figure 4, the majority of participant expressed a clear preference for LLM-generated responses over template-based ones. This preference held even in cases where the LLM output was not factually accurate. Participants frequently emphasized the fluency, tone, and clarity of the LLM responses as key factors influencing their choice. Qualitative feedback reinforces this trend. Some participants described the LLM-generated explanations as more insightful or engaging, highlighting that a well-articulated and relatable explanation, even if partially flawed, felt more helpful than a technically correct but dry or vague response. This feedback underscores a critical trade-off in dialogue-based XAI: while users value correctness, they also seek explanations that feel approachable and meaningful. This highlights the need to balance factual reliability with communicative quality. Future systems must address this tension by generating outputs that are not only factually accurate but also tailored to user expectations in terms of tone and expressiveness. In summary, LLMs show strong potential for generating natural and accessible explanations in dialogue-based XAI. However, making sure these explanations are accurate, context-aware, and factually grounded remains a major challenge, especially since this depends on both the type of explanation and the user’s ability to verify what the system says.

## 7. Conclusion

In this study, we generated LLM-based responses for a set of realistic dialogue scenarios and evaluated them through a human annotation study involving both expert and lay users. Our findings show that LLM-generated explanations are generally preferred across user groups, particularly for their coherence, plausibility, and usefulness. However, assessing factual correctness remains challenging, even for experts, underscoring the complexity of evaluating explanation reliability. Moreover, the observed over-reliance of lay users on fluent but potentially inaccurate responses highlights the risks posed by unverified outputs in high-stakes applications. These findings reinforce the importance of designing explanation systems that are not only adaptive and user-sensitive but also grounded in verifiable reasoning. Future work should explore advanced prompting strategies, such as ReAct [33], Tree of Thoughts [34], and Ra2Fd [35], to improve the faithfulness and transparency of LLM-generated explanations in dialogue-based XAI systems.

## Limitations

This study presents an initial exploration of the reliability of LLM-generated explanations in dialogue-based XAI, and several limitations should be acknowledged. First, the participant pool in our study was relatively small, comprising a total of 20 individuals split between expert and lay user groups. While this limits the generalizability of the findings, the substantial number of individual ratings enabled us to observe statistically significant effects. Moreover, although most participants were non-native English speakers, we mitigated potential issues with language comprehension by including only domain experts with strong academic or professional backgrounds in the expert group, where English proficiency can reasonably be assumed. For the lay user group, participants were German speakers with overall good English skills. While we did not explicitly test proficiency, we recruited only those who reported feeling confident in using English. To further reduce variability and ensure consistency across annotators, we provided detailed annotation guidelines and examples, which remained accessible throughout the study. This structured approach helped control for subjective interpretation and enhanced the reliability of our evaluation across participants. Second, our evaluation used only a single LLM to generate explanations. While this choice ensured consistency and reduced annotation complexity, it limits the generalizability of our findings. Including multiple LLMs could provide a more robust comparison but would increase the cognitive burden on annotators. Additionally, this preliminary study focused on a single prompt formulation per dialogue instance. Exploring a broader range of prompting strategies may reveal further insights into explanation variability and reliability. Given the already challenging nature of the task, we opted for depth over breadth. Future studies could use our results as a baseline and expand toward comparative evaluations or incorporate LLMs as self-assessing judges. Finally, the template-based responses used for comparison were sometimes quite technical and less polished in language. This may have influenced perceived fluency and preference in favor of the LLM-generated responses, especially among lay participants. While this reflects realistic contrasts between rule-based and generative systems, future work should control for linguistic quality more carefully to isolate explanatory content. Overall, these limitations position our study as an important first step toward understanding how LLMs function in interactive explanatory contexts, while pointing to several areas for refinement and deeper investigation.

## Ethical Considerations

A key ethical concern in deploying LLMs for XAI is the risk of user over-reliance. Our findings underscore that lay users, in particular, tend to perceive fluent and well-structured LLM-generated explanations as more trustworthy, even when these contain factual inaccuracies. This overtrust poses significant challenges for transparency and user safety in AI applications. It highlights the critical importance of research like ours, which rigorously evaluates the reliability of LLM explanations in

realistic dialogue settings. Such work is essential for informing the design of explanation systems that can signal uncertainty, clearly reference the underlying data, and help users distinguish between grounded facts and model-generated inferences. Ultimately, responsible integration of LLMs into XAI requires not just improving output quality but also aligning it with users' trust calibration and understanding. To ensure responsible research practices, no personal or sensitive data was collected during the study. All dialogue snippets were anonymized and derived from pre-existing datasets, and participants were informed about the nature and scope of the task prior to participation.

## Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o and DeepL-Write in order to: Grammar and spelling check, Rephrasing. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

## References

- [1] D. Mindlin, F. Beer, L. N. Sieger, S. Heindorf, E. Esposito, A.-C. Ngonga Ngomo, P. Cimiano, Beyond one-shot explanations: a systematic literature review of dialogue-based XAI approaches, *Artificial Intelligence Review* (2025).
- [2] K. Sokol, P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency, *KI-Künstliche Intelligenz* (2020).
- [3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial intelligence* (2019).
- [4] M. Kuźba, P. Biecek, What would you ask the machine learning model? identification of user needs for model explanations based on human-model conversations, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2020, pp. 447–459.
- [5] D. Martens, J. Hinns, C. Dams, M. Vergouwen, T. Evgeniou, Tell me a story! narrative-driven XAI with large language models, *Decision Support Systems* (2025).
- [6] A. Zytek, S. Pido, S. Alnegheimish, L. Berti-Équille, K. Veeramachaneni, Explingo: Explaining AI predictions using large language models, in: *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 1197–1208.
- [7] K. Matton, R. Ness, E. Kiciman, Walk the talk? measuring the faithfulness of large language model explanations, in: *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024.
- [8] K. Sokol, J. E. Vogt, What does evaluation of explainable artificial intelligence actually tell us? a case for compositional and contextual validation of XAI building blocks, in: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–8.
- [9] H. Shen, C.-Y. Huang, T. Wu, T.-H. K. Huang, ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-ai scientific writing, in: *Companion publication of the 2023 conference on computer supported cooperative work and social computing*, 2023, pp. 384–387.
- [10] D. Slack, S. Krishna, H. Lakkaraju, S. Singh, Explaining machine learning models with interactive natural language conversations using talktomodel, *Nature Machine Intelligence* (2023).
- [11] V. B. Nguyen, J. Schlötterer, C. Seifert, 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*, 2024, pp. 273–280.
- [12] G. He, N. Aishwarya, U. Gadiraju, 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*, 2025, pp. 907–924.
- [13] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [14] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, *Harv. JL & Tech.* (2017).



- [15] A. Zyteck, S. Pidò, K. Veeramachaneni, Llm for xai: Future directions for explaining explanations, 2024. URL: <https://arxiv.org/abs/2405.06064>. arXiv:2405.06064.
- [16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, *ACM Transactions on Information Systems* (2025).
- [17] T. Zhang, M. Zhang, W. Y. Low, X. J. Yang, B. A. Li, Conversational explanations: Discussing explainable AI with non-ai experts, in: *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 2025, pp. 409–424.
- [18] P. Jandaghi, X. Sheng, X. Bai, J. Pujara, H. Sidahmed, Faithful persona-based conversational dataset generation with large language models, *arXiv preprint arXiv:2312.10007* (2023).
- [19] Z. Yang, X. Jia, X. Jiang, Explaingen: a human-centered LLM assistant for combating misinformation, in: *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, 2025, pp. 120–123.
- [20] I. Feustel, N. Rach, W. Minker, S. Ultes, Towards a deeper understanding: Effects of dk integration for conversational XAI, [Manuscript under review]. (2025).
- [21] A. Renkl, Toward an instructionally oriented theory of example-based learning, *Cogn Sci* (2014).
- [22] I. Feustel, N. Rach, W. Minker, S. Ultes, Enhancing model transparency: A dialogue system approach to XAI with domain knowledge, in: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2024, pp. 248–258.
- [23] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, J. Pineau, How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, *arXiv preprint arXiv:1603.08023* (2016).
- [24] K. Morrison, P. Spitzer, V. Turri, M. Feng, N. Kühl, A. Perer, The impact of imperfect XAI on human-ai decision-making, *Proceedings of the ACM on Human-Computer Interaction* (2024).
- [25] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, *ACM Computing Surveys* (2023).
- [26] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4198–4205.
- [27] Z. Zhu, H. Chen, X. Ye, Q. Lyu, C. Tan, A. Marasović, S. Wiegrefe, Explanation in the era of large language models, in: *Proceedings of the 2024 Conference of the NAACL: Human Language Technologies (Volume 5: Tutorial Abstracts)*, 2024, pp. 19–25.
- [28] R. Koo, M. Lee, V. Raheja, J. I. Park, Z. M. Kim, D. Kang, Benchmarking cognitive biases in large language models as evaluators, *arXiv preprint arXiv:2309.17012* (2023).
- [29] P. E. McKnight, J. Najab, Mann-whitney u test, *The Corsini encyclopedia of psychology* (2010).
- [30] K. Krippendorff, *Content analysis: An introduction to its methodology*, Sage publications, 2018.
- [31] S. Castro, Fast Krippendorff: Fast computation of Krippendorff’s alpha agreement measure, <https://github.com/pln-fing-udelar/fast-krippendorff>, 2017.
- [32] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansky, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, Y. Matias, TRUE: Re-evaluating factual consistency evaluation, in: *Proceedings of the 2022 Conference of the NAACL: Human Language Technologies*, 2022, pp. 3905–3920.
- [33] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: *Int. Conference on Learning Representations (ICLR)*, 2023.
- [34] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: deliberate problem solving with large language models, in: *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.
- [35] Z. Zhu, Y. Liao, C. Xu, Y. Guan, Y. Wang, Y. Wang, Ra2fd: Distilling faithfulness into efficient dialogue systems, in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 12304–12317.