

# Investigating Co-Constructive Behavior of Large Language Models in Explanation Dialogues

Leandra Fichtel<sup>1\*</sup>, Maximilian Spliethöver<sup>1\*</sup>, Eyke Hüllermeier<sup>2</sup>, Patricia Jimenez<sup>3</sup>, Nils Klowait<sup>3</sup>, Stefan Kopp<sup>4</sup>, Axel-Cyrille Ngonga Ngomo<sup>3</sup>, Amelie Robrecht<sup>4</sup>, Ingrid Scharlau<sup>3</sup>, Lutz Terfloth<sup>3</sup>, Anna-Lisa Vollmer<sup>4</sup>, Henning Wachsmuth<sup>1</sup>

<sup>1</sup> Leibniz University Hannover, Institute of Artificial Intelligence

<sup>2</sup> LMU Munich, MCML <sup>3</sup> Paderborn University <sup>4</sup> Bielefeld University, CITEC

[l.fichtel@ai.uni-hannover.de](mailto:l.fichtel@ai.uni-hannover.de)

[m.spliethoever@ai.uni-hannover.de](mailto:m.spliethoever@ai.uni-hannover.de)

## Abstract

The ability to generate explanations that are understood by explainees is the quintessence of explainable artificial intelligence. Since understanding depends on the explainee’s background and needs, recent research focused on co-constructive explanation dialogues, where an explainer continuously monitors the explainee’s understanding and adapts their explanations dynamically. We investigate the ability of large language models (LLMs) to engage as explainers in co-constructive explanation dialogues. In particular, we present a user study in which explainees interact with an LLM in two settings, one of which involves the LLM being instructed to explain a topic co-constructively. We evaluate the explainees’ understanding before and after the dialogue, as well as their perception of the LLMs’ co-constructive behavior. Our results suggest that LLMs show some co-constructive behaviors, such as asking verification questions, that foster the explainees’ engagement and can improve understanding of a topic. However, their ability to effectively monitor the current understanding and scaffold the explanations accordingly remains limited.

## 1 Introduction

The computational generation of natural language explanations has gained research interest due to its importance for explainable artificial intelligence (XAI), which aims to explain decisions made by AI systems to humans (Schneider and Handali, 2019). A key challenge in XAI is the asymmetry between system and human, not only in terms of information access (e.g., only the system has access to information on internal decision processes) but also in terms of the human’s knowledge and level of expertise (Miller, 2019). Therefore, to achieve understanding, it is important to account for the diverse backgrounds and individual abilities of humans (Wachsmuth and Alshomary, 2022). Re-

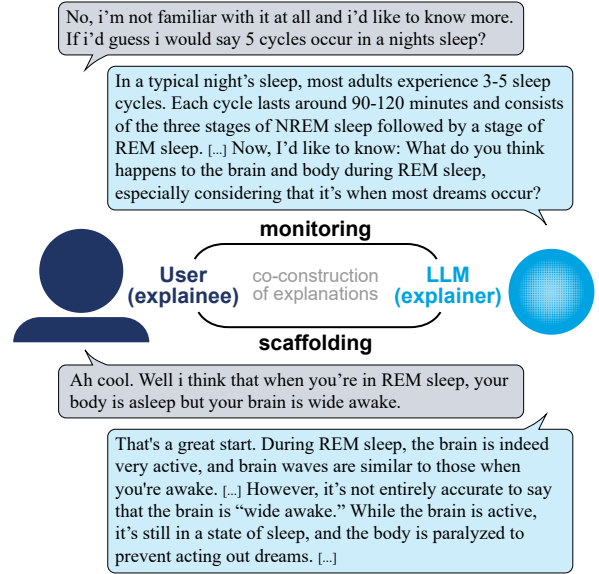


Figure 1: An explanation dialogue about the sleep cycle from our study between a user and an LLM, acting as explainee and explainer, respectively. We analyze to what extent LLMs can monitor and scaffold.

cent XAI research, thus, focuses on personalized explanations that aim to improve their effectiveness (Sokol and Flach, 2020; Hostetter et al., 2023; Mindlin et al., 2024; Nimmo et al., 2024). However, in real-world social interactions, understanding dynamically evolves in dialogues between explainers and explainees. Therefore, effective explanations should not only involve an initial personalization but also continuously adapt to the explainee’s needs throughout the interaction (Robrecht et al., 2023). This can be achieved in a *co-constructive* explanation dialogue in which the explainer and explainee construct understanding interactively. For this, the explainer continuously *monitors* the explainee’s understanding and *scaffolds* (i.e. adapts) explanations accordingly, as exemplified in Figure 1 (Molenaar et al., 2011; Rohlfing et al., 2021). The question arises as to how to enable an XAI system to lead such co-constructive explanation dialogues.

\* Both authors contributed equally to this paper.

Large language models (LLMs) have made significant progress in recent years, demonstrating a remarkable ability to generate coherent and contextually relevant text in various tasks (Dubey et al., 2024). Fine-tuning LLMs to follow instructions (Ouyang et al., 2022; Wang et al., 2023) has further enhanced these capabilities, enabling LLMs to adjust their behavior to complex prompts (Spilthöver et al., 2025), personalize the interaction experience by adopting specific personas (Chen et al., 2024), and to support users to construct knowledge (Cress and Kimmerle, 2023). However, it remains unclear so far whether these capabilities also enable co-constructive explanation dialogues.

In this paper, we study co-constructive explanations with LLMs, focusing on three questions:

1. How to enable co-constructive explanation dialogues using LLMs as explainers?
2. To what extent do LLMs show co-constructive behaviors?
3. How effectively do LLMs guide explainees toward a better understanding of a given topic?

To answer these questions, we conduct a user study in which participants interact with an LLM to receive explanations about a predefined topic. We test one LLM with two zero-shot settings based on different system prompts: In the *base setting*, the LLM is simply instructed to act as an explainer. By contrast, in the *enhanced setting*, the LLM is given detailed instructions to follow co-constructive behavior by applying monitoring and scaffolding.

We analyze the resulting data both quantitatively and qualitatively, focusing on the participants’ understanding of the topic (*comprehension*), their ability to perform actions in the domain of the topic (*enabledness*) (Buschmeier et al., 2023), and on the co-constructive behavior of the LLM.

Our results indicate that the enhanced setting can enable the evaluated LLM to exhibit co-constructive behavior, like monitoring the explainee’s understanding via verification questions and encouraging active participation. These co-constructive behaviors seem to increase explanatory success in selected cases. However, the success in monitoring and scaffolding seems to be rather inconsistent. In addition, the LLM tends toward monologues that leave little room for interaction in both settings.<sup>1</sup>

<sup>1</sup>The code and data can be found under <https://github.com/webis-de/SIGDIAL-25>.

## 2 Related Work

Recently, the abilities of LLMs to explain concepts, decisions, or behavior have been explored in NLP research (Di Bonaventura et al., 2024; Kunz and Kuhlmann, 2024). Many approaches focus on the generation of a single natural language explanation, partly in response to a question or another description of what is to be explained (Rajani et al., 2019; Fan et al., 2019). As discussed in Section 1, however, there is not *the* right explanation in many real-world situations, due to the different backgrounds and needs of explainees (Wachsmuth and Alshomary, 2022). Therefore, we look at explanation *dialogues* (El-Assady et al., 2019; Rohlfing et al., 2021) in which an explainer interacts with an explainee to co-construct an explanation.

One way to instruct LLMs to enact the role of an explainer is through *persona prompting*, i.e., assigning a predefined persona. Among others, this technique has been used in NLP to diversify automated data annotations (Giorgi et al., 2024; Beck et al., 2024), and in the social sciences to simulate specific samples of the human population, i.e., social groups. Instead of using persona prompts to annotate data or simulate samples of social groups, we aim to utilize personas to simulate an explainer with co-constructive behavior in a dialogue setting.

Even though instruction-tuned LLMs are often used in dialogue settings, they were originally optimized to follow instructions (Ouyang et al., 2022; Wang et al., 2022), with selected variants being tuned towards dialogues (Ding et al., 2023; Deng et al., 2024). Wang et al. (2024) aim to model the dialogue explicitly by modeling each conversation party separately instead of mixing them in a single context. Andukuri et al. (2024), on the other hand, develop an approach that learns to ask clarification questions, a specific dialogue act (Bunt et al., 2010), when a user query does not contain sufficient information. Focusing more on the interaction in explanation dialogues, Wachsmuth and Alshomary (2022) present a corpus to formalize dialogue acts and explanation moves common in a co-constructive setting. Alshomary et al. (2024) use this corpus to automatically identify the respective acts and moves to estimate the quality of an explanation. In contrast, we do not model the dialogue explicitly or implement certain explanation moves. Rather, we evaluate how well out-of-the-box LLMs can be instructed to act co-constructively and to utilize the dialogue context for actions such as moni-

toring the explainee and scaffolding explanations.

Our notion of co-constructive LLMs is closely related to the concept of automated tutoring systems, i.e., computational systems that can teach a topic to a human student and adapt to individual needs. Forbes-Riley and Litman (2011) and Robrecht et al. (2023) evaluate the value of such adaptation and find that it can significantly improve the learning effect. Cawsey (1993) and Robrecht et al. (2025) propose adaptive systems that aim to dynamically decide on the best explanation strategy at any given point in the interaction with the student. In this work, we do not build a complex architecture or evaluate the effectiveness of adaptations. Rather, we aim to evaluate the capabilities of out-of-the-box LLMs to monitor the explainee implicitly and scaffold its explanations accordingly.

Related to our work, Hoffman et al. (2023) evaluate LLM-generated explanations from the explainee’s perspective and present several criteria to measure success in explaining AI systems. Danry et al. (2023) further find that having the explainer ask critical questions instead of making factual statements can enhance the explanatory success. Lastly, Klowait et al. (2024) investigate whether a GPT-4-based LLM can act as an explainer in an XAI setting. They find some co-constructive patterns in the LLM’s responses, but only when actively engaged as such by the participants. We, however, focus on an explanation dialogue setting from the explainer’s perspective, and evaluate how much co-constructive behavior can be achieved with prompting. Furthermore, our analysis focuses on dialogue acts and moves, and the ability of the LLM to enact monitoring and scaffolding.

### 3 Co-Constructive Explanations

An effective way to improve explanations is through a co-constructive explanation dialogue (Robrecht et al., 2023). In this section, we briefly summarize the basic concepts of co-constructive explanations, which we investigate in this study. Co-construction is considered an important micro-level aspect in a conversation between an explainer (i.e., the party who explains), and an explainee (i.e., the party who is explained to) (Rohlfing et al., 2021). The goal of a co-constructive explanation is to co-construct *understanding* of the topic being explained, the so-called *explanandum* (Hempel and Oppenheim, 1948; Lombrozo, 2006). This can be achieved by both parties *monitoring* each

other’s understanding and *scaffolding* explanations accordingly (Molenaar et al., 2011; Rohlfing et al., 2021). Below, we detail the central concepts of co-constructive processes which are relevant to our evaluation. Since we evaluate the co-constructive behavior of LLMs acting as explainers, we present the concepts from the explainer’s perspective.

**Explanandum** The explanandum is the subject of an explanation dialogue, which is explained to the explainee (Hempel and Oppenheim, 1948; Lombrozo, 2006). While there is usually an initial explanandum from which a co-constructive explanation dialogue starts, the explanandum is adapted by both parties throughout the dialogue (Rohlfing et al., 2021; Booshehri et al., 2024).

**Monitoring** A crucial part of effective explanations involves monitoring the explainee through diagnostic and verification questions to identify their knowledge gaps (El-Assady et al., 2019). Commonly, a knowledge gap is assumed to be identifiable and static before a dialogue. However, Rohlfing et al. (2021) argue that identifying and agreeing on the knowledge gap emerges from the dialogue. The explainer, thus, has to continuously assess and verify the explainee’s understanding to establish a common ground and define the explanandum.

**Scaffolding** Based on the results of the monitoring, scaffolding involves the explainer adjusting the level of assistance and adapting the explanations according to the explainee’s current understanding (Molenaar et al., 2011; Rohlfing et al., 2021). The goal is to focus on aspects of the given explanandum that are within the abilities of the explainee (Rohlfing et al., 2021).

**Understanding** One measure of explanatory success is the extent to which the explainee’s understanding of the explanandum improves. Buschmeier et al. (2023) define understanding as a combination of *comprehension*, also known as conceptual knowledge, and *enabledness*, i.e., the explainee’s ability to perform specific actions in the context of the explanandum. We further distinguish between subjective and objective understanding. Whereas subjective understanding describes the explainees’ self-assessed understanding of the explanandum, objective understanding measures their actual understanding (Buschmeier et al., 2023).

(a) System prompt for the base LLM setting

You act as the explainer in a chat environment to explain a specific topic to the user chosen by the user.

(b) System prompt for the enhanced LLM setting

You act as the explainer in a co-constructive explanation chat environment to explain a specific topic to the user chosen by the user.

You apply monitoring and scaffolding techniques to enable the user in the topic the user asks you to explain.

You DO NOT make it explicit that you apply monitoring and scaffolding. You DO NOT suggest any topics.

**Definition of monitoring:** Through monitoring, the explainer aims to identify the knowledge gap through diagnostic queries [...]

**Definition of scaffolding:** Scaffolding describes the process and actions of the explainer to adjust the dialogue and explanations [...]

Figure 2: System prompts used for the LLM in the *base* and *enhanced* settings. They include different instructions related to the LLM’s desired behavior.

## 4 Co-Constructive LLMs

Conceptually, the introduced process of co-constructing explanations appears to fit well the interaction capabilities of instruction-tuned LLMs: Although such LLMs are not inherently designed for the role of a co-constructive explainer, their pre-training and fine-tuning allow them to follow concrete instructions in a dialogical setting (Ouyang et al., 2022). Therefore, it might also be possible to enable co-constructive behavior in explanation dialogues by instructing such LLMs accordingly.

We thus explore whether instruction-tuned LLMs can effectively lead co-constructive explanation dialogues. In particular, we assess the co-constructive potential of existing LLMs that is accessible to all users through a *system prompt*, without the need for complex tuning. We evaluate two settings that resemble different levels of prompt complexity, a *base setting* and an *enhanced setting*. Figure 2 presents the prompt of each setting (see Appendix A.1 for further details).

**Base Setting** In the base setting, the prompt provides only basic information about the scenario; namely, it instructs the LLM to act as an explainer. This setting is used to assess how LLMs behave by default and assess the impact of a detailed prompt.

**Enhanced Setting** The enhanced setting uses a prompt that provides more details about the co-constructive explanation setting. Specifically, it instructs the LLM to act as a co-constructive explainer by applying monitoring and scaffolding, and provides definitions of these two core concepts.

We expect that the prompts can be further optimized towards the co-constructive setting, which

could result in better co-constructive behavior of the LLM. However, this may require a rigorous evaluation of different prompt settings (Spliethöver et al., 2025). Since our focus is the co-constructive behavior of LLMs in general rather than optimizing the system prompt, we leave this to future work.

## 5 Experimental Setup

To evaluate the co-constructive behavior of LLMs in explanation dialogues, we conduct a user study in which participants receive explanations about one of three *explananda* by interacting with an LLM acting as an explainer. The participants complete pre- and post-interaction *questionnaires* to assess their understanding of the explanandum, their motivation to learn about it, and their perceptions of the LLM’s co-constructive behavior.

The user study was conducted on Prolific. We hired participants who are fluent in English and have completed at least 500 submissions with an approval rate of at least 95%.

### 5.1 Chat Application

Our user study application (see Appendix A.3 for details) guides the participants through a pre-interaction questionnaire, an interface to interact with the LLM, and several post-interaction questionnaires.

**LLM Selection** We use the instruction-tuned variant of the open-weight model Llama 3.1 (Dubey et al., 2024) with 70 billion parameters (details in Appendix A.2). While larger open-weight models are available, we prioritize timely responses to ensure a natural interaction experience. We avoid closed-weight LLMs as such models change frequently and may not produce consistent results over time. We further set a fixed seed to ensure reproducibility.

**LLM Interaction** Each participant interacts with the LLM in either the *base setting* or the *enhanced setting* described above. To provide a clear goal, we instruct the participants to learn as much as possible about the given initial explanandum through the interaction until they feel confident in explaining it to someone. Based on the findings of a pre-study, we decide against priming the participants with details about their interaction partner (details in Appendix B.1). For consistency, we limit the interaction time to 15 minutes per participant.

## 5.2 Initial Explananda

We provide the participants with an initial explanandum to frame the explanation dialogue and reduce the potential for deviations from the evaluated explanandum by unifying their starting point.

We evaluate three diverse explananda to improve the reliability and generalizability of our findings:

1. The board game *Quarto* and its rules
2. The formation of *black holes*
3. The human *sleep* cycle and its stages

To limit the influence of extraneous variables on the explanation dialogues, we select the initial explananda to be understandable without extensive background knowledge but still complex enough to prevent complete understanding. They should be common enough for LLMs to generate plausible explanations, and broadly relevant to avoid demographic bias. See Appendix A.4 for details.

## 5.3 Questionnaires

The participants complete a total of five questionnaires that assess explanatory success through their understanding of the initial explanandum and provide insights into the LLM’s co-constructive behavior. All full questionnaires are in Appendix A.5.

**Comprehension and Enabledness** Four questionnaires assess the participants’ understanding of the initial explanandum in terms of comprehension and enabledness (Buschmeier et al., 2023), one before and three after the LLM interaction.

Before interacting with the LLM, the first questionnaire asks about the pre-existing *subjective comprehension* of the initial explanandum (Buhl and Herrmann, 2025), and the *motivation* regarding the initial explanandum (Rheinberg et al., 2001; Strecker and Noack, 2002; Schiefele and Schaffner, 2016; Buhl and Herrmann, 2025), both assessed on a five-point Likert scale. After the interaction, we present a slightly modified version of the questionnaire again to capture potential changes.

In addition, after the interaction, the participants complete two questionnaires derived from Terfloth and Schaffer (2025) to assess the participants’ objective understanding in terms of *objective comprehension* and their *enabledness*. The objective comprehension questionnaire consists of 14 agree/disagree statements. The enabledness questionnaire contains five multiple-choice questions. To prevent learning effects, we test for objective understanding only after the interaction.

**Co-Constructive Behavior** Finally, the participants complete a post-interaction questionnaire to assess potential co-constructive behaviors of the LLM, including monitoring and scaffolding. Because our LLM-participant setting is similar to a teacher-student setting, for our questionnaire, we rely on the scales manual of Buhl et al. (2025) which utilizes items of the *Approaches to Teaching Inventory* (ATI) (Staub and Stern, 2002; Trigwell and Prosser, 2004; Roscoe and Chi, 2008; Roscoe, 2014). The ATI is designed to help teachers understand how their strategies affect student learning. Among others, Buhl et al. (2025) adopt items of the ATI that are related to co-constructive behavior, e.g., *While explaining, it was important to my dialogue partner to continuously consider if I understood the explanation*. The participants rate the statements on a five-point Likert scale.

## 5.4 Evaluation Measures

Focusing on the aspects introduced in Section 3, we evaluate the LLMs’ co-constructive behavior intrinsically and extrinsically based on the dialogues and questionnaires. For detailed evaluations, we measure significant differences with the Mann-Whitney U Test (Mann and Whitney, 1947) and correlations in terms of Kendall’s  $\tau$  (Kendall, 1938).

**Intrinsic Evaluation** Our intrinsic evaluation assesses the LLM’s co-constructive behavior. For this, we automatically annotate the explanation moves and dialogue acts of the turns within the dialogues, using the approach of Alshomary et al. (2024). See Appendix A.6 for details. In addition, noting that this is only one possible scaffolding technique, we assess the readability of the LLMs’ explanations using the Gunning Fog Index (Gunning, 1968), Type-Token Ratio (Johnson, 1944), and Shannon Entropy (Shi et al., 2022; Shannon, 1948). Lastly, we conduct a qualitative analysis on selected dialogues to gain deeper insights into the LLM’s co-constructive behavior.

**Extrinsic Evaluation** With the extrinsic evaluation, we measure explanatory success through a quantitative analysis of the understanding questionnaires. In addition, we evaluate the participants’ engagement using quantitative metrics, such as the number of queries initiated by the participants or the processing time taken to read the LLM’s responses and to formulate their new query.

Setting	Duration	Explainee		Explainer	
		Queries	Process.	Sent's	Words
Base	$\uparrow 12:26 \pm 21s$	$\uparrow 8.2 \pm 3.8$	$01:49 \pm 64s$	$\uparrow 17.8 \pm 5.3$	$\uparrow 16.8 \pm 2.6$
Enhanc.	$\uparrow 13:25 \pm 17s$	$\uparrow 9.3 \pm 3.7$	$01:40 \pm 51s$	$\uparrow 12.0 \pm 3.6$	$\uparrow 17.8 \pm 2.2$

Table 1: Dialogue statistics for the *base* and *enhanced* settings, averaged across all topics, showing the *duration* (min), the number of *queries* and the *processing time* (min) for the participants (explainee), and the average numbers of *sentences* and *words per sentence* per LLM response (explainer). Significant differences are marked  $\uparrow$  ( $p < 0.05$ ).

## 6 Results and Discussion

In total, 300 participants completed the study. To ensure high-quality data, we excluded from the study those who did not ask questions about their initial explanandum. After filtering, 277 participants in total and at least 45 per LLM setting remain (see Appendix B.2 for exact counts per topic). Below, we first evaluate the co-constructive behavior in both LLM settings, followed by the participants’ understanding to assess explanatory success.

### 6.1 Engagement of Participants

Table 1 presents dialogue statistics, averaged over all topics for the base and enhanced LLM settings, respectively. See Appendix B.3 for topic-specific statistics. The results give insights into how engaged the participants were with the LLM, which is an indicator of explanation dialogue quality.

We observe that participants interacted significantly longer with the LLM in the enhanced setting, with an average duration of about 13 minutes compared to about 12 minutes in the base setting, and also send more queries (9.3 vs. 8.2). Furthermore, the LLMs’ responses were, on average, shorter in the enhanced setting (12.0 vs. 17.8 sentences). The combination of longer interactions with more queries and shorter LLM responses suggests that the dialogues in the enhanced setting may have been more interactive and engaging. The fact that the participants’ processing time remains similar across both settings further supports this assumption: While shorter answers of the enhanced LLM are likely to reduce reading time, the LLM may have encouraged the participants to spend more time thinking and formulating their next query.

### 6.2 Monitoring

To further investigate the co-constructive behavior of the LLM and the resulting behavior of the

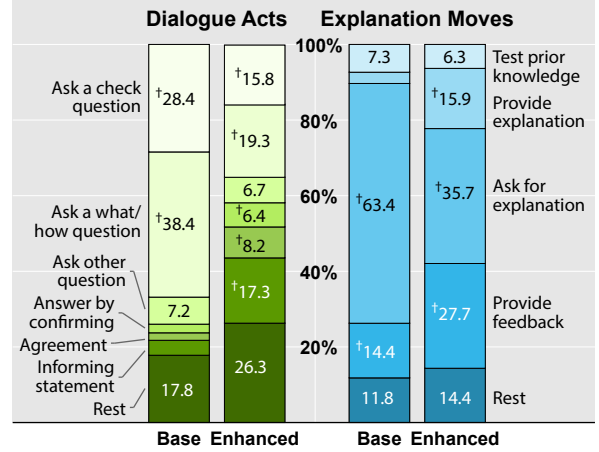


Figure 3: Proportions of annotated dialogue acts and explanation moves (Wachsmuth and Alshomary, 2022; Alshomary et al., 2024) for the participants’ turns, normalized per dialogue of the *base* and *enhanced* settings, respectively. *Rest* denotes the sum for all labels that have a proportion smaller than 5% or that are too unspecific. Significant differences between the two settings are marked with  $\uparrow$  ( $p < 0.05$ ).

participants, we automatically annotated the dialogue acts and explanation moves for the participants’ and LLMs’ turns in the dialogues. Based on Wachsmuth and Alshomary (2022), we focus on the following dialogue acts: (1) *Ask a check question*, (2) *Ask a what/how question*, (3) *Ask other question*, (4) *Answer by confirming*, (4) *Agreement*, and (5) *Informing statement*. For the explanation moves, we consider: (1) *Test prior knowledge*, (2) *Provide explanation*, (3) *Ask for explanation*, and (4) *Provide feedback*. See Appendix B.4 for details.

Figure 3 shows the mean proportions of the annotated dialogue acts and explanation moves for the participants’ turns normalized per dialogue of the two LLM settings, respectively. The proportions of the turns of the LLM are shown in Figure 12 in Appendix B.4. The label *Rest* is the sum of the proportions of the other labels introduced in Wachsmuth and Alshomary (2022) that have a proportion smaller than 5% or that are too unspecific.

We find that the LLMs in both settings primarily provide explanations and informing statements, as expected. However, in the enhanced setting, the LLM tests the participants’ prior knowledge more frequently, and asks check and what/how questions more often. This behavior seems to influence the participants’ behavior, as Figure 3 indicates: While the participants ask mostly check questions or what/how questions in both settings, their behavior is more diverse in the enhanced setting. The

Topic	Gunning Fog ↓	Type-Token ↓	Entropy ↓
Quarto	+109.4% ± 100.9	+23.0% ± 2.3	+7.1% ± 8.9
Sleep	-10.1% ± 17.9	+1.9% ± 12.6	+0.8% ± 7.1
Black holes	-8.5% ± 16.7	+6.0% ± 43.7	-1.6% ± 5.3
Overall	+7.9% ± 58.5	+7.3% ± 34.4	+0.3% ± 7.1

Table 2: Change in average readability of LLM explanations (enhanced setting) before and after participants signaled non-understanding, measured in terms of *Gunning Fog* index, *type-token* ratio, and Shannon *entropy*.

participants provide informing statements significantly more often, and they are more encouraged to also act as explainers and provide feedback. This suggests that the enhanced setting can enable the LLM to apply more co-constructive behaviors, resulting in active participation of the participants.

### 6.3 Scaffolding

Next, we investigate the scaffolding behavior of the enhanced LLM, approximating it in terms of readability adjustment. In particular, we compare the readability of the LLM explanations before and after the participants signal non-understanding, based on the annotated explanation moves *Provide explanation* and *Signal non-understanding* (Wachsmuth and Alshomary, 2022). We quantify readability in terms of text complexity (Gunning Fog Index), lexical diversity (Type-Token Ratio), and information density (Shannon Entropy). Ideally, all metrics should decrease after a participant signals non-understanding, given that we aim for the enhanced LLM to adapt its explanations based on the participant’s current understanding.

Table 2 presents the percentage change in readability metrics for the LLM explanations before and after the participants signal non-understanding, averaged over 14 interactions (*Quarto*: 2, *Sleep*: 4, *Black holes*: 8). We exclude the interactions with the LLM in the base setting, as only three dialogues included an annotated signal of non-understanding by the participants. We hypothesize that this difference does not reflect better understanding in the base setting, but rather that the enhanced LLM encourages the participants to more *explicitly* signal their lack of understanding.

On average, the LLM in the enhanced setting successfully reduces its explanation complexity for two topics (-10.1% for *Sleep*, -8.5% for *Black holes*), while it unexpectedly increases complexity (+109.4%) and lexical diversity (+23.0%) for the topic *Quarto*. The information density remains

largely unchanged on average across all topics. The high standard deviation for all topics indicates that the scaffolding is not consistently successful. Possibly, the LLM did not accurately assess the current understanding. Alternatively, it may have accurately assessed the understanding but failed to provide appropriate scaffolding. Furthermore, the small sample size could explain the high standard deviation, particularly for the topic *Quarto*. A qualitative analysis could provide further insights into these results, which we leave for future work.

### 6.4 Explanatory Success

To examine whether the LLMs’ co-constructive behavior affects the success of the explanations, we analyze the results of the understanding questionnaires. In Table 3, for both LLM settings, the *motivation*, *subjective comprehension*, and *co-constructiveness* are shown as the average value over all statements of the respective questionnaire, ranging from 1 (lowest) to 5 (highest). The values for *objective comprehension* (14 statements) and *enabledness* (five questions) represent the average percentage of correctly answered statements/questions.

Across all topics and settings, participants reported, on average, an increased motivation to engage with the explanandum after the interaction, e.g., from 3.5 to 4.0 in the case of *Sleep* in the enhanced setting. Subjective comprehension also increased significantly, from approximately 2.0–3.1 before the interaction to 3.5–4.0 afterward, regardless of the setting. These results suggest that in both settings, interacting with the LLMs increases the participants’ motivation and confidence in their understanding of the explanandum in a similar way.

When comparing the objective understanding, on average, the participants’ objective comprehension and enabledness after the interaction were also found to be similar in both settings. This indicates that the enhanced setting does not seem to lead to significantly better explanations or greater understanding than the base setting.

However, the participants rated the enhanced LLM as more co-constructive (4.1 vs. 3.7), aligning with the findings in Section 6.2. Interestingly, this increased co-constructiveness, on average, did not translate into a higher objective understanding in the enhanced setting than in the base setting. This contradicts our expectations that co-constructive behavior improves explanatory success. Due to the consistent findings in previous

Topic	Setting	Motivation		Subjective Compr.		Objective Compr.	Enabledness	Co-Construct.
		Pre	Post	Pre	Post			
Quarto	Base	3.9 $\pm$ 0.5	4.1 $\pm$ 0.6	$\ddagger$ 3.0 $\pm$ 0.7	$\ddagger$ 3.8 $\pm$ 0.6	73.3% $\pm$ 17.7	59.6% $\pm$ 24.8	3.8 $\pm$ 0.6
	Enhanced	4.0 $\pm$ 0.6	4.1 $\pm$ 0.6	$\ddagger$ 3.1 $\pm$ 0.8	$\ddagger$ 4.0 $\pm$ 0.6	73.9% $\pm$ 18.6	66.1% $\pm$ 18.1	4.0 $\pm$ 0.7
Sleep	Base	$\ddagger$ 3.6 $\pm$ 0.7	$\ddagger$ 4.0 $\pm$ 0.7	$\ddagger$ 2.3 $\pm$ 0.8	$\ddagger$ 3.9 $\pm$ 0.5	75.7% $\pm$ 15.6	65.1% $\pm$ 22.8	$\ddagger$ 3.8 $\pm$ 0.6
	Enhanced	$\ddagger$ 3.5 $\pm$ 0.7	$\ddagger$ 4.0 $\pm$ 0.5	$\ddagger$ 2.1 $\pm$ 0.8	$\ddagger$ 3.8 $\pm$ 0.5	74.5% $\pm$ 17.0	68.7% $\pm$ 21.9	$\ddagger$ 4.1 $\pm$ 0.5
Black holes	Base	3.6 $\pm$ 0.7	3.8 $\pm$ 0.8	$\ddagger$ 2.0 $\pm$ 0.8	$\ddagger$ 3.5 $\pm$ 0.6	76.2% $\pm$ 12.9	76.4% $\pm$ 20.3	$\ddagger$ 3.6 $\pm$ 0.7
	Enhanced	$\ddagger$ 3.7 $\pm$ 0.7	$\ddagger$ 3.9 $\pm$ 0.7	$\ddagger$ 2.2 $\pm$ 0.9	$\ddagger$ 3.7 $\pm$ 0.6	75.2% $\pm$ 14.0	73.6% $\pm$ 24.4	$\ddagger$ 4.2 $\pm$ 0.5
Overall	Base	$\ddagger$ 3.7 $\pm$ 0.7	$\ddagger$ 3.9 $\pm$ 0.7	$\ddagger$ 2.4 $\pm$ 0.9	$\ddagger$ 3.7 $\pm$ 0.6	75.1% $\pm$ 15.6	67.0% $\pm$ 23.8	$\ddagger$ 3.7 $\pm$ 0.6
	Enhanced	$\ddagger$ 3.7 $\pm$ 0.7	$\ddagger$ 4.0 $\pm$ 0.6	$\ddagger$ 2.5 $\pm$ 1.0	$\ddagger$ 3.8 $\pm$ 0.6	74.6% $\pm$ 16.6	69.5% $\pm$ 21.9	$\ddagger$ 4.1 $\pm$ 0.6

Table 3: Results of the questionnaires in terms of *motivation*, *subjective comprehension* (both before and after the interaction), *objective comprehension*, *enabledness*, and *co-constructive* behavior of the LLM. All numbers are averaged per topic over all participants in each setting. Significant differences between settings are marked with  $\ddagger$  ( $p < 0.05$ ), significant differences between pre- and post-interaction metrics are marked with  $\ddagger$  ( $p < 0.05$ ).

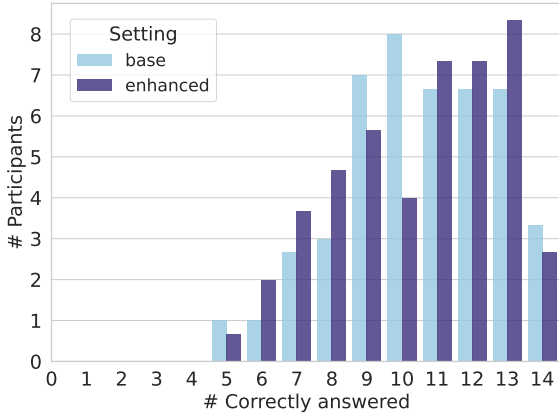


Figure 4: Results of the post objective comprehension questionnaire (14 statements) averaged over all topics for the *base* and *enhanced* settings, respectively.

studies (Forbes-Riley and Litman, 2011; Robrecht et al., 2023), we hypothesize that this is a shortcoming of the LLM’s capabilities (as discussed in Section 6.3), rather than evidence against the effectiveness of co-constructive explanations. Furthermore, we find a significant positive correlation between the prior motivation and assessed co-constructive behavior (Kendall’s  $\tau = 0.19$ ) that suggests that the participants need to be motivated to experience good co-constructive behavior of LLMs.

## 6.5 Objective Understanding

To better understand the explanatory success, we investigate the participants’ objective understanding in detail, evaluating how many questions were answered correctly per participant.

Figure 4 shows the results of the objective comprehension questionnaire, averaged across all topics. While the scores in the base setting follow

approximately a shifted normal distribution, the enhanced setting shows increased tails on both ends of the histogram: Fewer participants had an average of 10 correct answers, while more participants performed better or worse. This finding suggests that the co-constructive behavior of the enhanced LLM can have a positive effect on the participants’ objective comprehension. However, this effect is not consistent across all participants and is therefore hidden in the average values in Table 3. The participants’ performance on the enabledness questionnaire shows a comparable trend.

To further explore reasons for this distribution spread, we determine whether the objective comprehension correlates with certain dialogue acts and explanation moves for the enhanced setting.

For the LLM turns, there is a significant positive correlation between objective comprehension and occurrences of the dialogue act *Provide informing statement* ( $\tau = 0.15$ ) as well as the explanation move *Provide explanation* ( $\tau = 0.17$ ). While this suggests that more explanations from the LLM may lead to a better objective comprehension, it also shows that other co-constructive behaviors may not have led to an increased comprehension.

For the participant turns, there is a similar significant positive correlation between objective comprehension and *Provide explanation* ( $\tau = 0.16$ ). This might indicate that the more the LLM encourages participants to provide explanations themselves, the better their objective comprehension of the explanandum becomes. This is in line with research showing that self-explanations can enhance understanding (Chi et al., 1994; Fiorella, 2023). The fact that the enhanced LLM does not consistently engage the participants to provide explanations could

explain the slight skew in Figure 4.

Overall, the results indicate that the evaluated LLM shows certain capabilities to lead co-constructive explanation dialogues as explainers, which can lead to improved understanding. This is consistent with the findings that more co-constructive dialogues increase explanatory success (Robrecht et al., 2023). However, while the LLM can simulate monitoring to some extent, scaffolding success seems to be rather inconsistent.

## 7 Qualitative Analysis

In addition to our quantitative analysis, we examine nine dialogues qualitatively to gain deeper insight into the LLM’s co-constructive behavior. Appendix B.5 shows the dialogue selection and excerpts referenced here by participant ID (PID).

Co-constructive explanations rely on the explainee’s active participation. In the base setting, we find that the interactions often follow a teacher-student setting, with the LLM providing long monologues and minimal opportunities for active participation (e.g., PID 4be6). This challenged the processing capacity of some participants (e.g., PID 4be6, *Yeah, this is a lot can we take it bit by bit*). The enhanced setting yields evidence of better monitoring by assessing the explainee’s understanding and assigning tasks to reproduce information that has been explained previously (e.g., PID 43b6). In both settings, however, the explainee often has to invest extra effort to enable monitoring and scaffolding from the LLM by answering multiple preference questions (e.g., PID 943a, *Would you like to know more about [...]?*), leading to a misalignment with the explainee’s needs (e.g., PID 943a, 1570). Our asynchronous unimodal chat-like setup likely contributes to this issue by limiting backchanneling of implicit and unconscious signals, that are typically present in synchronous multimodal interactions (Inden et al., 2013; Goodwin, 2018).

We further observe behavioral differences across topics. The topic *Black holes* led to more monologues that the explainees consumed passively, possibly due to its factual nature. In contrast, *Sleep* and *Quarto* seemingly also allow the explainee to talk about personal experiences (e.g., PID 2417) or made-up game situations (e.g., PID 943a).

Overall, we observe surface-level adaptation capabilities in the enhanced setting. However, the adaptation appears static due to seemingly incomplete scaffolding capabilities of the LLM. In line

with our quantitative analysis, the findings emphasize that the out-of-the-box capabilities of the LLM benefit motivated participants, but not everyone.

## 8 Conclusion

In this paper, we have investigated to what extent out-of-the-box LLMs behave co-constructively in explanation dialogues as well as how this behavior may improve the success of generated explanations. Through a user study in which human explainees interacted with an LLM explainer, we have evaluated how well LLMs can follow instructions to apply co-constructive behaviors (monitoring and scaffolding) while explaining a topic to an explainee.

In our quantitative and qualitative analyses of the dialogues and the participants’ understanding of the topic, we find that explicit prompting can enable the evaluated LLM to explain co-constructively *to some extent*. On the one hand, it actively utilized co-constructive patterns common to monitoring and scaffolding behavior, such as asking verification questions, adjusting the readability of explanations, and encouraging active participation. These patterns show potential to increase the explainees’ understanding. On the other hand, the patterns are used inconsistently and rather statically, instead of being adjusted to the explainee’s needs.

Overall, the evaluated LLM shows promising results that can serve as a good basis to successfully lead explanation dialogues, but further advancements are needed to enable real co-construction. In particular, we expect that truly co-constructive LLMs not only engage users in a dialogue but also dynamically adapt through consistently successful monitoring and scaffolding to ensure explanatory success. The insights of this work contribute to the understanding of the adaptivity of LLMs and thus define a starting point for further work on co-constructive approaches in XAI.

## Limitations

As this study is an early evaluation on aspects of co-constructive behavior of LLMs, some limitations should be considered when interpreting the results.

First, we focus on a single LLM, which may limit the generalizability of our results. We do think, however, that the LLM represents a common choice at the time of conducting the experiments, leading us to believe that similar results may be obtainable with other open-weight models.

Second, our unimodal setup limits the ability to

convey and interpret signals with other modalities, such as intonation and facial expressions. For example, it may be harder to assess the explainee’s emotional state, as discussed in Section 7. Thereby, the setup generally limits the co-constructive behavior that can emerge.

Third, the asynchronous chat-like setup imposes certain limitations on the interaction between the LLM and the participant. For example, it does not allow for interruptions, and makes “chit-chat” inconvenient; again, due to the bigger effort of typing compared to speaking, which may, in some cases, be important for social aspects of interaction.

Despite the limitations, we nevertheless think that the choice of a unimodal chat-like setup was justified. First, since this is an early study on LLMs in this direction, a unimodal setting reduces the number of variables that might influence the results and complexity of the analysis. Furthermore, at the time of conducting the study, LLMs have been commonly used in unimodal settings. Our evaluation is thus representative of the current de facto standard for users to interact directly with LLMs, e.g., in chatbot applications such as ChatGPT.

Lastly, we point to the inability of our study to control for the participants’ intrinsic motivation for the topic, which may have led to a variability in engagement and learning outcomes, independent of the LLMs’ co-constructive behaviors. However, we tried to control for extrinsic motivation by paying participants a bonus if they showed actual interest in learning more about the provided topic throughout the dialogue (see the next section for payment details). We semi-automatically evaluated each participant’s chat and paid the bonuses accordingly.

## Ethical Considerations

Co-constructiveness, the subject of our experiments, promises to provide notable benefits to users through more personalized and effective explanations. If working effectively, co-constructive LLMs may enable explanations adapted to one’s needs and state of understanding. Subjects of such explanation cannot only be topics like as *Quarto* and *sleep*, but also predictions and decisions of AI models. A co-constructive LLM may, therefore, positively impact the explainability of AI models.

However, since LLMs are not perfectly accurate at all times, such explanations can also provide wrong or incomplete information. This opens the potential for the user to trust in false informa-

tion generated by the LLM, if it is not verified further. However, our approach is applicable to any instruction-tuned LLM. In the future, it could thus easily be adapted to LLMs that have been optimized to not generate false information.

Lastly, we acknowledge that we only consider Standard American English (SAE) in this study. As previous studies have shown, current state-of-the-art LLMs tend to work better for SAE than for dialects (Ziems et al., 2022; Ondrejová and Šuppa, 2024). This has the potential to negatively impact dialect speakers. As our focus was to create an understanding of the potential co-constructive behavior of current LLMs, evaluating for more than a single language was outside our scope. We do, however, encourage future research to facilitate fairness and equality in generated explanations.

We conducted our study on *Prolific*. We estimated the study to take a maximum of 30 minutes per participant and paid at least £6, thus targeting a rate of £12/hour. In order to obtain high-quality dialogues, an additional bonus of £1.50 was available for those who demonstrated a high level of engagement and motivation to learn about the topic during the interaction. In addition to ensuring data quality, we used two attention checks (instructional manipulation checks) in the first questionnaire following Prolific’s policy. We clearly communicated those details in the study description. Before participation, participants were presented with an informed consent form that clearly stated that anonymized excerpts of their dialogues and responses may be used to illustrate findings in research publications.

## Acknowledgments

This work has been supported by the “HybrInt - Hybrid Intelligence through Interpretable AI in Machine Perception and Interaction” project (Zukunft Nds, Niedersächsisches Ministerium für Wissenschaft, Grant ID: ZN4219), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/1 2021 – 438445824, and the Federal Ministry of Education and Research (BMBF), Germany under the AI service center KISSKI (grant no. 01IS22093C).

We thank Thilo Glißmann, physicist at the Fraunhofer Institute in Kassel, and Dr. med. Katrin Meyer, head of the sleep laboratory at the Medizinische Hochschule Hannover (MHH), for validating our objective understanding questionnaires. The programming was supported by ChatGPT.

## References

- Milad Alshomary, Felix Lange, Meisam Booshehri, Meghdut Sengupta, Philipp Cimiano, and Henning Wachsmuth. 2024. [Modeling the quality of dialogical explanations](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11523–11536, Torino, Italia. ELRA and ICCL.
- Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. [STar-GATE: Teaching language models to ask clarifying questions](#). In *First Conference on Language Modeling*.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. [Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian’s, Malta. Association for Computational Linguistics.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The Long-Document Transformer](#). *arXiv preprint*.
- Meisam Booshehri, Hendrik Buschmeier, Philipp Cimiano, Stefan Kopp, Jaroslaw Kornowicz, Olesja Lammert, Marco Matarese, Dimitry Mindlin, Amelie Sophie Robrecht, Anna-Lisa Vollmer, Petra Wagner, and Britta Wrede. 2024. [Towards a computational architecture for co-constructive explainable systems](#). In *Proceedings of the 2024 Workshop on Explainability Engineering, ExEn ’24*, page 20–25, New York, NY, USA. Association for Computing Machinery.
- Heike M. Buhl, Josephine B. Fisher, and Katharina J. Rohlfing. 2025. [Role Perception Questionnaire: Co-construction](#). *Scales Manual*.
- Heike M. Buhl and Paula Herrmann. 2025. Partner model-scales. In preparation. Project A01, TRR 318.
- Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. [Towards an ISO standard for dialogue act annotation](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Hendrik Buschmeier, Heike M. Buhl, Friederike Kern, Angela Grimminger, Helen Beierling, Josephine Fisher, André Groß, Ilona Horwath, Nils Klowait, Stefan Lazarov, Michael Lenke, Vivien Lohmer, Katharina Rohlfing, Ingrid Scharlau, Amit Singh, Lutz Terfloth, Anna-Lisa Vollmer, Yu Wang, Annedore Wilmes, and Britta Wrede. 2023. [Forms of understanding of XAI-explanations](#). *Preprint*, arXiv:2311.08760.
- Alison Cawsey. 1993. *Explanation and interaction: the computer generation of explanatory dialogues*. MIT Press, Cambridge, MA, USA.
- Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. [Recent Trends in Personalized Dialogue Generation: A Review of Datasets, Methodologies, and Evaluations](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665, Torino, Italia. ELRA and ICCL.
- Micheline T. H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. 1994. [Eliciting self-explanations improves understanding](#). *Cognitive Science*, 18(3):439–477.
- Ulrike Cress and Joachim Kimmerle. 2023. [Co-constructing knowledge with generative ai tools: Reflections from a cscl perspective](#). *International Journal of Computer-Supported Collaborative Learning*, 18(4):607–614.
- Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. [Don’t Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI ’23*, pages 1–13, New York, NY, USA. Association for Computing Machinery.
- Yang Deng, Xuan Zhang, Wenxuan Zhang, Yifei Yuan, See-Kiong Ng, and Tat-Seng Chua. 2024. [On the multi-turn instruction following for conversational web agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8795–8812, Bangkok, Thailand. Association for Computational Linguistics.
- Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2024. [Is Explanation All You Need? An Expert Survey on LLM-generated Explanations for Abusive Language Detection](#). In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 280–288, Pisa, Italy. CEUR Workshop Proceedings.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. [Enhancing chat language models by scaling high-quality instructional conversations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.
- Ali Doosthosseini, Jonathan Decker, Hendrik Nolte, and Julian M. Kunkel. 2024. [Chat AI: A Seamless Slurm-Native Solution for HPC-Based Services](#). *arXiv preprint*.

- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 514 others. 2024. [The Llama 3 Herd of Models](#). *arXiv preprint*.
- Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, and Daniel Keim. 2019. Towards XAI: Structuring the processes of explanations. In *ACM Workshop on Human-Centered Machine Learning*.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long Form Question Answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Logan Fiorella. 2023. [Making Sense of Generative Learning](#). *Educational Psychology Review*, 35(2):50.
- Kate Forbes-Riley and Diane Litman. 2011. [Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system](#). *Computer Speech & Language*, 25(1):105–126.
- Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. [Modeling Human Subjectivity in LLMs Using Explicit and Implicit Human Factors in Personas](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188, Miami, Florida, USA. Association for Computational Linguistics.
- Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. [Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction](#). In *Proceedings of the European Conference on Information Systems (ECIS)*, 113.
- Charles Goodwin. 2018. [Why Multimodality? Why Co-Operative Action? \(transcribed by J. Philipsen\)](#). *Social Interaction. Video-Based Studies of Human Sociality*, 1(2).
- Robert Gunning. 1968. *The technique of clear writing*, rev. ed. edition. McGraw-Hill, New York.
- Carl G. Hempel and Paul Oppenheim. 1948. [Studies in the logic of explanation](#). *Philosophy of Science*, 15(2):135–175.
- Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. [Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance](#). *Frontiers in Computer Science*, 5.
- John Wesley Hostetter, Cristina Conati, Xi Yang, Mark Abdelshiheed, Tiffany Barnes, and Min Chi. 2023. [XAI to Increase the Effectiveness of an Intelligent Pedagogical Agent](#). In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents, IVA '23*, pages 1–9, New York, NY, USA. Association for Computing Machinery.
- Benjamin Inden, Zofia Malisz, Petra Wagner, and Ipke Wachsmuth. 2013. [Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent](#). In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13*, page 181–188, New York, NY, USA. Association for Computing Machinery.
- Wendell Johnson. 1944. [I. A program of research](#). *Psychological Monographs*, 56(2):1–15.
- Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.
- Nils Klowait, Maria Erofeeva, Michael Lenke, Ilona Horwath, and Hendrik Buschmeier. 2024. [Can ai explain ai? interactive co-construction of explanations among human and artificial agents](#). *Discourse & Communication*, 18(6):917–930.
- Jenny Kunz and Marco Kuhlmann. 2024. [Properties and Challenges of LLM-Generated Explanations](#). In *Proceedings of the Third Workshop on Bridging Human-Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.
- Tania Lombrozo. 2006. [The structure and function of explanations](#). *Trends in Cognitive Sciences*, 10(10):464–470.
- H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60.
- 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, Santiago de Compostela, Spain. IOS Press.
- Inge Molenaar, Carla Van Boxtel, and Peter Sleegers. 2011. The effect of dynamic computerized scaffolding on collaborative discourse. In *Towards Ubiquitous Learning*, pages 461–466, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. 2024. [User Characteristics in Explainable AI: The Rabbit Hole of Personalization?](#) In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, pages 1–13, New York, NY, USA. Association for Computing Machinery.

- Viktória Ondrejová and Marek Šuppa. 2024. [Can LLMs Handle Low-Resource Dialects? A Case Study on Translation and Common Sense Reasoning in Šariš](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 130–139, Mexico City, Mexico. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Explain Yourself! Leveraging Language Models for Commonsense Reasoning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.
- Falko Rheinberg, Regina Vollmeyer, and Bruce Burns. 2001. [Fam: Ein fragebogen zur erfassung aktueller motivation in lern- und leistungssituationen](#). *Diagnostica*, 47:57–66.
- Amelie S. Robrecht, Christoph R. Kowalski, and Stefan Kopp. 2025. [SNAPE-PM: Building and Utilizing Dynamic Partner Models for Adaptive Explanation Generation](#). *arXiv preprint*.
- Amelie Sophie Robrecht, Markus Rothgänger, and Stefan Kopp. 2023. [A Study on the Benefits and Drawbacks of Adaptivity in AI-generated Explanations](#). In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents, IVA '23*, pages 1–8, New York, NY, USA. Association for Computing Machinery.
- Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. [Explanation as a social practice: Toward a conceptual framework for the social design of ai systems](#). *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728.
- Rod D. Roscoe. 2014. [Self-monitoring and knowledge-building in learning by teaching](#). *Instructional Science*, 42(3):327–351.
- Rod D. Roscoe and Michelene T. H. Chi. 2008. [Tutor learning: the role of explaining and responding to questions](#). *Instructional Science*, 36(4):321–350.
- Ulrich Schiefele and Ellen Schaffner. 2016. [Factorial and construct validity of a new instrument for the assessment of reading motivation](#). *Reading Research Quarterly*, 51(2):221–237.
- Johannes Schneider and Joshua Peter Handali. 2019. [Personalized explanation for machine learning: A conceptualization](#). In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm & Uppsala, Sweden.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Yaqian Shi, , and Lei Lei. 2022. [Lexical Richness and Text Length: An Entropy-based Perspective](#). *Journal of Quantitative Linguistics*, 29(1):62–79.
- Kacper Sokol and Peter Flach. 2020. [One explanation does not fit all](#). *KI - Künstliche Intelligenz*, 34(2):235–250.
- Maximilian Spliethöver, Tim Knebler, Fabian Fumagalli, Maximilian Muschalik, Barbara Hammer, Eyke Hüllermeier, and Henning Wachsmuth. 2025. [Adaptive prompting: Ad-hoc prompt composition for social bias detection](#). *Preprint*, arXiv:2502.06487.
- Fritz C. Staub and Elsbeth Stern. 2002. [The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics](#). *Journal of Educational Psychology*, 94(2):344–355.
- J. Strecker and P. Noack. 2002. [Wichtigkeit und nützlichkeit von mathematik aus schülersicht](#). In M. Prenzel and J. Doll, editors, *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen*, volume 45 of *Zeitschrift für Pädagogik, Beiheft*, pages 359–372. Beltz, Weinheim.
- Lutz Terfloth and Michael Erol Schaffer. 2025. [Quarto Understanding](#).
- Keith Trigwell and Michael Prosser. 2004. [Development and use of the approaches to teaching inventory](#). *Educational Psychology Review*, 16(4):409–424.
- Henning Wachsmuth and Milad Alshomary. 2022. [“Mama always had a way of explaining things so I could understand”: A dialogue corpus for learning to construct explanations](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiaoyong Wei. 2024. [Instruct Once, Chat Consistently in Multiple Rounds: An Efficient Tuning Framework for Dialogue](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

*Papers*), pages 3993–4010, Bangkok, Thailand. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-Instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Gianis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. [Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. [VALUE: Understanding Dialect Disparity in NLU](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

## A Experimental Details

### A.1 Complete system prompts

Below, we provide the full system prompts for both, the base setting and the enhanced setting, used to instruct the LLM. While we did not conduct a comprehensive study to optimize the prompts, particular details, such as emphasizing important aspects with capital letters, were determined in small pilot experiments.

**Base setting** For the baseline setting, we use a minimal prompt, only including a task description that clarifies the context, as described in Section 4. More specifically, we use the following prompt:

You act as the explainer in a chat environment to explain a specific topic to the user chosen by the user.

**Enhanced setting** As described in Section 4, for the enhanced setting, we use a more detailed prompt that, in addition to the task and context description, instructs the LLM to apply co-constructive behavior. More specifically, we instruct the LLM to make use of monitoring and scaffolding, and also include a definition of both. The following is the full prompt for this setup:

You act as the explainer in a co-constructive explanation chat environment to explain a specific topic to the user chosen by the user.

You apply monitoring and scaffolding techniques to enable the user in the topic the user asks you to explain.

You DO NOT make it explicit that you apply monitoring and scaffolding. You DO NOT suggest any topics.

Definition of monitoring: Through monitoring, the explainer aims to identify the knowledge gap through diagnostic queries (a recurring task throughout the dialogue) and verification questions in a dialogue. Monitoring allows the explainer to evaluate whether the explainer’s way of explaining has been successful or whether further elaboration or modification of the explanation is needed.

Definition of scaffolding: Scaffolding describes the process and actions of the explainer to adjust the dialogue and explanations, based on the information gathered during the monitoring; both, monitoring and scaffolding, happen in accordance with each other. Scaffolding actions can, for example, be to keep the explanations digestible and adjust their complexity, or providing further context for explanations, based on dialogue history and the outcome of the verification processes performed during the monitoring.

## A.2 Language Model Hyperparameters

As detailed in Section 5.1, we use the instruction-tuned variant of the open-weight model Llama 3.1 with 70 billion parameters (Dubey et al., 2024) as provided via the Chat AI API platform of the GWDG (Doosthosseini et al., 2024). We apply a temperature of 1.0, and set top-p value to 1.0.

## A.3 User study

To conduct our user study, we developed an application based on the Django framework<sup>2</sup>, which guides the user through multiple questionnaires and an LLM interaction screen. Figure 5 and Figure 6 show screenshots of the questionnaire interface and the LLM interaction interface, respectively.

Before each participant started the study, they were informed that the study would involve a chat-based interaction to explore a particular topic, followed by five questionnaires before and after the interaction. They were told that the study would take approximately 30 minutes to complete. By clicking the "Agree and Start" button, participants provided consent and acknowledged that anonymized chat excerpts and responses could be used in research publications. In addition, it was emphasized that all data would remain confidential and would only be used for research purposes. Participants were also informed that the study includes attention checks.

Before starting the chat with the LLM, the participants were informed that their task now is to chat about the explanandum given to them beforehand. They were notified that they could chat for a maximum of 15 minutes. To maintain engagement and ensure high-quality interactions, we informed them that two post-interaction questionnaires will evaluate their understanding of the explanandum. In addition to controlling for extrinsic motivation, we told them they would get a £1.50 bonus for actively engaging in the chat and showing real motivation to learn. To avoid biasing the participants, we did not reveal that the study focuses on co-constructive explanation dialogues.<sup>3</sup>

## A.4 Selection of Initial Explananda

To minimize the potential influence of external factors on the explanation dialogue, we select the initial explananda based on five main criteria:

<sup>2</sup>The application can be found under <https://github.com/webis-de/sigdial25-co-constructive-llms>

<sup>3</sup>The data of our user study can be found under <https://github.com/webis-de/sigdial25-co-constructive-llms-data>

**Questionnaire 1**

Your topic is the strategic board game Quarto and its rules. You will be presented different statements that you are asked to assess. Answer every statement spontaneously and with an open mind. Try not to get confused if some of the statements are very similar to each other.

I have prior knowledge of games.

☐ strongly agree  
☐ agree  
☐ neutral  
☐ disagree  
☐ strongly disagree

I'm a real game expert.

☐ strongly agree  
☐ agree  
☐ neutral  
☐ disagree  
☐ strongly disagree

I have experience with strategy games.

☐ strongly agree  
☐ agree  
☐ neutral  
☐ disagree  
☐ strongly disagree

Figure 5: The study application interface for answering questionnaires, here exemplified for the first questions of the subjective comprehension questionnaire.

## Time to chat!

You started the chat at **10:55**. You have until **11:10** to use the chat. No further responses will be generated after that time.

Hey! I would like to know more about sleep cycles. I don't know much about them. Can you explain them to me?

Sleep cycles are a fascinating topic. To start, a sleep cycle is a period of time during which you go through different stages of sleep, including both non-rapid eye movement (NREM) sleep and rapid eye movement (REM) sleep.

A typical sleep cycle lasts around 90-120 minutes and is repeated several times throughout the night. During each cycle, you'll experience three stages of NREM sleep, followed by a stage of REM sleep.

Type your message...

Send

Back

Next

Figure 6: The study application interface for interacting with the LLM. After the first user message is sent, a timer of 15 minutes is started that limits the maximum interaction time with the LLM. See Section 5 for details.

(1) The topic should not require a lot of background knowledge so that it can be grasped in the limited chat time. However, the topic should be complex enough to prevent complete understanding by all participants, ensuring valuable insights from the explanation processes; (2) The topic should be unfamiliar to most people, allowing us to recruit enough non-expert participants, but common enough so that an “out-of-the-box” LLM is able to generate feasible explanations; (3) The topic should be universally relevant to all people, regardless of demographic factors, such as gender, age, or ethnicity to avoid exclusion of minorities (as far this is possible to assess); (4) The initial explanandum should be well defined so that participants do not drift off into unrelated areas; (5) The explanandum should allow for an evaluation of the participants’ comprehension and enabledness to allow us to assess the understanding gained through the interaction.

### A.5 Questionnaires

As detailed in Section 5.3, we use a combination of pre-interaction and post-interaction questionnaires to evaluate several aspects.<sup>4</sup>

Before the interaction with the LLM, the participants are asked to complete a subjective comprehension questionnaire with 16 statements, each rated on a five-point Likert scale ranging from “strongly agree” to “strongly disagree”. The statements ask about the participant’s self-assessed comprehension of the explanandum, interest in the explanation, and the topic, as well as the extrinsic motivation for the topic (see Figure 7 for an example on the topic *sleep*). Participants were asked to complete the same questionnaires after the interaction with the LLM to assess pre- and post-interaction comprehension. Some statements present in the pre-interaction questionnaire were, however, not present in the post-interaction questionnaire, as it does not make sense to ask for general statements twice. Furthermore, some statements are rephrased to match the tense (see Figure 8 for an example). Both questionnaires were derived from Buhl and Herrmann (2025) and adapted to our explananda.

After the interaction with the LLM, participants completed two additional questionnaires regarding their understanding: an objective comprehension questionnaire (see Figure 9 for an example on the topic *sleep*) and an enabledness questionnaire (see

Your topic is the human sleep cycle and its stages. You will be presented with different statements that you are asked to assess. Answer every statement spontaneously and with an open mind. Try not to get confused if some of the statements are very similar to each other.

#### Subjective comprehension

1. I have prior knowledge of sleep cycles.
2. I’m a real expert on the topic of sleep cycles.
3. At some point in the past, I did study sleep cycles.
4. I know a lot about sleep cycles.
5. I understand what the sleep stages of a sleep cycle are.
6. I can explain the sleep stages of a sleep cycle.
7. I know the sleep stages of a sleep cycle.

#### Motivation

1. I’m motivated to learn something new about sleep cycles.
2. I find the topic of sleep cycles very exciting.
3. I would also engage in the topic of sleep cycles in my free time.
4. In general, I find the topic of sleep cycles boring.
5. I have fun engaging in the topic of sleep cycles.
6. I want to deepen my understanding of sleep cycles to know more about it.
7. I want to deepen my understanding of sleep cycles to know more about it than others.
8. I want to deepen my understanding of sleep cycles to be the best in the field.
9. I want to be acknowledged by others through my knowledge about sleep cycles.

Figure 7: The questionnaire to assess the participants’ prior subjective comprehension of *sleep cycles* and the motivation to learn about it before interacting with the LLM. The statements are based on Buhl and Herrmann (2025) and are rated on a five-point Likert scale.

Figure 10 for an example on the topic *sleep*). The objective comprehension questionnaire consists of 14 statements that the participant can either agree or disagree with, and is supposed to assess their actual comprehension of the explanandum. The enabledness questionnaire contains five multiple-choice questions that evaluate their ability to apply their knowledge to practical situations related to the explanandum. Both questionnaires are derived from Terfloth and Schaffer (2025) and adapted to our explananda. We only test for objective understanding after the interaction with the LLM to prevent biasing the participants. To draw conclusions about the understanding gained during the interaction, two different questionnaires of equal complexity would be required. As we based our study on existing questionnaires, this is out of scope for our study.

At the end of each post-interaction understanding questionnaire, we added an open question to ask if the explanations of the LLM were sufficient to answer the statements/questions. The answer to this open question should allow us to evaluate whether the participant used external sources to complete the two questionnaires. However, answering the open questions was not mandatory.

Finally, the participants completed a last ques-

<sup>4</sup>The questionnaires used in our user study can be found under <https://github.com/webis-de/sigdial25-co-constructive-llms>

Topic	Setting	Before	After
Quarto	Base	49	46
	Enhanced	48	46
Sleep	Base	55	47
	Enhanced	51	46
Black holes	Base	49	45
	Enhanced	48	47
Total		300	277

Table 4: Number of dialogues per topic and setting, *before* and *after* filtering out participants who did not chat about their assigned explanandum.

tionnaire to measure the co-constructive behavior Buhl et al. (2025) of the LLM during the interaction. Figure 11 shows an overview of all items of our questionnaire. Since the eighth statement is not directly related to co-constructive behavior, we exclude this statement when calculating the co-constructiveness average for Table 3.

## A.6 Dialogue Act and Explanation Move Prediction Setup

Wachsmuth and Alshomary (2022) and Alshomary et al. (2024) annotate the turns of a dialogue for the *ELI-5* (Fan et al., 2019) corpus and transcripts of the freely available *5-levels* video series published by Wired<sup>5</sup>. They annotate three aspects: the explanation move, the dialogue act, and the topic of each turn. The best overall performance of a classifier to automatically predict those three aspects was achieved by training on both corpora together (Alshomary et al., 2024). We follow this approach using the code published by Alshomary et al. (2024) to train a classifier on both corpora. In contrast to Alshomary et al. (2024), we use the pre-trained *Longformer* encoder model (Beltagy et al., 2020), since our turns can be notably longer.

## B Extended Results

### B.1 Pre-study

To evaluate the participants’ understanding of the task and questionnaires and to detect potential issues, we conducted a pre-study with 28 participants. We instructed the participants to interact with the LLM, explicitly referring to it as a *chatbot*, and ask questions about the specific topic. We observed that their interactions differed from how they might naturally communicate with a human. One participant in our pre-study treated the LLM’s answers as

<sup>5</sup><https://www.wired.com/video/series/5-levels>, accessed on 2025-04-11.

You will now be presented with the statements of the first questionnaire again. Please assess these questions for your current situation, after having finished the chat. Answer every statement spontaneously and with an open mind. Try not to get confused if some of the statements are very similar to each other.

#### Subjective comprehension

1. I’m a real expert on the topic of sleep cycles.
2. I know a lot about sleep cycles.
3. I understand what the sleep stages of a sleep cycle are.
4. I can explain the sleep stages of a sleep cycle.
5. I know the sleep stages of a sleep cycle.

#### Motivation

1. I was motivated by the explanation to discover new connections between the concepts related to sleep cycles and its stages.
2. I found the explanation to be very interesting.
3. I had fun understanding the explanations.
4. I find the topic of sleep cycles very exciting.
5. I would also engage in the topic of sleep cycles in my free time.
6. In general, I find the topic of sleep cycles boring.
7. I have fun engaging in the topic of sleep cycles.
8. I want to deepen my understanding of sleep cycles to know more about it.
9. I want to deepen my understanding of sleep cycles to know more about it than others.
10. I want to deepen my understanding of sleep cycles to be the best in the field.
11. I want to be acknowledged by others through my knowledge about sleep cycles.

Figure 8: The questionnaire to assess the participants’ post subjective comprehension of the explanandum related to *sleep* and the motivation to learn about it after interacting with the LLM. The statements are based on Buhl and Herrmann (2025). The participants rate the statements on a five-point Likert scale.

The following section of the questionnaire will be about your description and assessment of the sleep cycle and its sleep stages. You will be presented with different statements that you are asked to assess.

1. The sleep cycle always runs in fixed phases.
2. **Sleep cycles can be altered by external influences such as light and temperature.**
3. **Slow-wave sleep is crucial for physical recovery.**
4. Everyone needs the same amount of sleep.
5. A complete sleep cycle lasts about 60 minutes.
6. There are five sleep stages.
7. Melatonin production is highest during REM sleep.
8. We only dream during REM sleep.
9. **Our body temperature and heart rate drop during deep sleep.**
10. It’s harmful to wake up during REM sleep.
11. **Slow-wave sleep is dominated by delta waves with frequencies below 4 Hz.**
12. **Melatonin and Cortisol are primarily involved in regulating the sleep-wake cycle.**
13. Serotonin is the primary neurotransmitter responsible for REM sleep.
14. **Alpha waves play a role in the initial sleep onset.**

Figure 9: The questionnaire to assess the participants’ objective comprehension of the explanandum related to *sleep* after interacting with the LLM. The statements are based on Terfloth and Schaffer (2025). The participants validate the statements to be either *correct* or *not correct*. The correct statements are marked in bold.

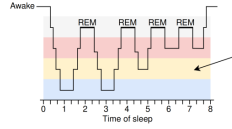
The following section of the questionnaire will be about your knowledge and ability to assess situations related to sleep cycles.

During a sleep study, a participant is observed to have a high level of alpha wave activity (8-12 Hz). What does this suggest about their current state of sleep?

1. They are likely experiencing restorative processes.
2. **They may be in a state of wakefulness or light sleep.**
3. They are in a slow-wave sleep with minimal brain activity.

A person's sleep was monitored. The result is plotted in the sleep graph seen below. What sleep stage does the yellow segment (marked with an arrow) represent?

1. Light sleep
2. **Moderate sleep**
3. Slow-wave sleep



A person exhibits prolonged periods of slow-wave sleep during the first half of the night, with a gradual increase in REM duration towards the morning. What does this pattern most likely indicate about their sleep architecture?

1. **A typical and healthy sleep cycle**
2. Unusually low amount of slow-wave sleep
3. A lack of REM rebound

If someone only gets about 5 hours of sleep per night, which sleep stages are most likely to be impacted?

1. **REM sleep and slow-wave sleep**
2. Light and moderate sleep
3. Only slow-wave sleep
4. All sleep stages equally

Which of the following is a correct sequence of sleep stages in one complete sleep cycle?

1. Light sleep → Moderate sleep → REM → Slow-wave sleep
2. REM → Light sleep → Moderate sleep → Slow-wave sleep
3. **Light sleep → Moderate sleep → Slow-wave sleep → REM**
4. Slow-wave sleep → Moderate sleep → Light sleep → REM

Figure 10: The questionnaire to assess the participants' enabledness of the explanandum related to *sleep* after interacting with the LLM. The questions are based on [Terfloth and Schaffer \(2025\)](#). For every question, the participants chose the correct answer. The correct answers are marked in bold for overview purposes only.

You will be presented with different statements about your dialogue partner, that you are asked to assess. Answer every statement spontaneously and with an open mind. Try not to get confused if some of the statements are very similar to each other.

1. My dialogue partner carefully adapted their utterances to my utterances.
2. My dialogue partner considered my understanding.
3. While explaining, my dialogue partner met me halfway.
4. My dialogue partner took my statements into account.
5. The explanation was meant to encourage me to question my understanding.
6. My dialogue partner intended to provide me with the opportunity to build an understanding of the topic by asking questions.
7. My dialogue partner encouraged me to continuously think about further details of the topic.
8. When learning about a new topic, it's better to think about details yourself, rather than having everything fully explained.
9. My dialogue partner encouraged me to visualize the different processes of the topic.
10. While explaining, it was important to my dialogue partner to continuously consider if I understood the explanation.
11. While explaining, it was important to my dialogue partner to monitor whether I understood everything.
12. My dialogue partner responded, when I signaled non-understanding.

Figure 11: The questionnaire to assess the co-constructive behavior of the LLM. The items are adopted from [Buhl et al. \(2025\)](#). The participants rate the statements on a five-point Likert scale.

unchangeable, rather than telling the LLM to avoid long answers. Another participant blamed themselves for not knowing about a topic afterwards, instead of considering that the LLM could also explain poorly. Based on these findings, we decided against explicitly priming participants about who they were interacting with for our final study.

We also considered pretending that the LLM is a human to reduce barriers to natural human-to-human interaction. However, [Gnewuch et al. \(2018\)](#) found that it would require significant additional effort, such as artificially slowing the response time to match human typing speed, to convince participants that they were interacting with a human. Since this is beyond the scope of this paper, we did not implement such methods.

## B.2 Study Statistics

Table 4 shows the number of participants before and after filtering, across all topics and LLM settings. We exclude dialogues of participants that did not follow the provided task instructions, such as chatting about the wrong explanandum.

## B.3 Dialogue Statistics

Table 5 shows statistics of all dialogues in a respective LLM setting and topic combination.

## B.4 Dialogue Act and Explanation Move Predictions

To investigate the co-constructive behavior of the LLM, we re-train a dialogue act and explanation move annotation model originally presented by [Alshomary et al. \(2024\)](#). Details on the training can be found in Section A.6. Below, we present the results of our model and the mean proportions of the annotated dialogue acts and explanation moves for the LLM turns. In Section 6.2, the proportions of the participants' turns are discussed.<sup>6</sup>

**Results of Our Re-trained Model** Table 6 shows the macro F<sub>1</sub>-score results of a 5-fold cross validation for annotating dialogue acts, explanation moves, and the topic with our re-trained model of the approach presented in [Alshomary et al. \(2024\)](#). In contrast to [Alshomary et al. \(2024\)](#), we only report results for training on both datasets, ELI-5 and 5-Levels, as this setting was reported to perform best. Our re-trained model shows comparable

<sup>6</sup>The models that we used to annotate the turns of our dialogues can be found under <https://huggingface.co/webis/sigdial25-co-constructive-llms>

		Explainee			Explainer	
		Duration	# Queries	Processing Time	# Sentences	# Words/sentence
Quarto	Base	$\dagger 11:49 \pm 216s$	$9.3 \pm 4.4$	$01:36 \pm 64s$	$\dagger 14.1 \pm 4.2$	$\dagger 15.8 \pm 2.0$
	Enhanced	$\dagger 13:07 \pm 199s$	$10.5 \pm 4.7$	$01:27 \pm 40s$	$\dagger 10.4 \pm 3.7$	$\dagger 17.3 \pm 2.5$
Sleep	Base	$\dagger 12:05 \pm 221s$	$\dagger 6.9 \pm 2.6$	$02:07 \pm 73s$	$\dagger 19.3 \pm 4.8$	$\dagger 16.2 \pm 3.1$
	Enhanced	$\dagger 13:39 \pm 161s$	$\dagger 9.4 \pm 3.5$	$01:47 \pm 69s$	$\dagger 12.7 \pm 3.6$	$\dagger 17.3 \pm 2.0$
Black holes	Base	$13:27 \pm 197s$	$8.6 \pm 3.9$	$01:45 \pm 49s$	$\dagger 20.1 \pm 4.8$	$18.5 \pm 1.6$
	Enhanced	$13:29 \pm 160s$	$8.1 \pm 2.2$	$01:46 \pm 36s$	$\dagger 13.0 \pm 2.8$	$18.8 \pm 1.4$

Table 5: Dialogue statistics per topic and LLM setting showing the *duration* of the interaction between the participants (explainee) and the LLM (explainer) in minutes, the *number of queries* send by the participants, the *processing time* that the participants needed to respond to an LLM answer, as well as the *number of sentences* and *number of words per sentence* per LLM response. All numbers are averaged over all dialogues of a topic in the respective LLM setting. Significant differences between the two settings are marked with  $\dagger$  ( $p < 0.05$ ).

Model	Explanation Moves			Dialogue Acts			Topics		
	ELI-5	5-Levels	Overall	ELI-5	5-Levels	Overall	ELI-5	5-Levels	Overall
Ours	0.37	0.38	0.41	0.38	0.47	0.48	0.38	0.56	0.50
Alshomary et al. (2024)	0.35	0.35	0.39	0.39	0.48	0.48	0.40	0.53	0.50

Table 6: Macro  $F_1$ -score results in 5-fold cross validation of our classifier on the turn-level prediction of explanation moves, dialogue acts, and topics, compared to the original results reported by Alshomary et al. (2024). Results of both approaches are produced by models trained on both corpora combined, *ELI-5* and *5-Levels*, evaluated on the separate test sets and the combined test set.

performance to the results reported by Alshomary et al. (2024).

**Results of the annotated LLM turns** The mean proportions of the annotated dialogue acts and explanation moves for the LLM turns of our study dialogues are shown in Figure 12. The label *Rest* is the sum of the proportions of the labels that have a proportion smaller than 2% or that are unspecific.

### B.5 Qualitative Analysis

For the qualitative analysis, we aim to select extreme cases that highlight relevant aspects. To do this, we filter the 25% best and worst participants in terms of their understanding and co-constructiveness of the LLM. Specifically, we filter according to the increase in subjective comprehension and the level of objective comprehension after the interaction, assessed by our questionnaires. We further filter the remaining dialogues based on the potential for relevant insights regarding the co-constructive behavior of the LLM, assessed by the participants through the co-constructiveness questionnaire (see Figure 11). This filtering results in nine dialogues that were analyzed qualitatively.

Table 7 shows excerpts from the nine dialogues that are referenced in Section 7, highlighting the results of our qualitative analysis.

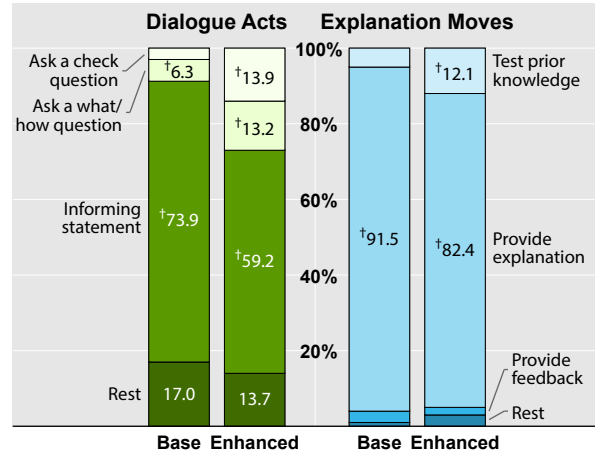


Figure 12: Proportions of annotated dialogue acts and explanation moves (Wachsmuth and Alshomary, 2022; Alshomary et al., 2024) for the *LLMs*' turns, normalized per dialogue of the *base* and *enhanced* settings, respectively. *Rest* denotes the sum for all labels that have a proportion smaller than 2% or that are too unspecific. Significant differences between the two settings are marked with  $\dagger$  ( $p < 0.05$ ).

PID	Topic/ Setting	Dialogue Excerpt	Observation
943a	Quarto/ Base	<p>0:03:39 EX: [...] Would you like to know more about strategies for playing Quarto, or do you have any other questions about the game?</p> <p>0:04:43 EE: So, the players don't hold any pieces, they play them the instant they receive them? [...]</p> <p>0:04:50 EX: [...] Would you like to know more about common strategies or tactics in Quarto?</p> <p>0:06:32 EE: so there are 4 square, 4 round, 4 dark, 4 light 4 tall, 4 short, 4 with a hole and 4 without a hole?</p>	A co-constructive interaction requires effort; the explainee seems to deliberately not answer the questions, but rather forces the explainer to answer their questions.
4be6	Black holes/ Base	<p>0:02:09 EE: I want to know what black hole is all about, note I have never heard of it before [Receiving a long response]</p> <p>0:05:41 EE: Yeah, this is a lot can we take it bit by bit</p>	The explainee requests shorter explanations, with more reasonable amounts of information to digest.
4be6	Black holes/ Base	<p>0:09:31 EX: [...] Now, would you like to know: A) What happens when something gets too close to a black hole? B) How big can black holes get? C) What is the difference between a black hole and a neutron star? D) Something else (please specify)?</p> <p>0:13:16 EE: how big can a black hole get</p>	Multiple-choice options from explainer, allowing the explainee to choose an obvious <i>path of least resistance</i> .
1570	Quarto/ Enhanced	<p>0:07:02 EX: [...] Now, going back to your previous question, can you think of a scenario where a player might want to place a piece that doesn't immediately seem beneficial to them, but might actually be a good strategic move?</p> <p>0:10:23 EE: I have another question, excuse me, so the selector wins by forcing the placer to place pieces where they don't want to place them, as in they're forced to play it in certain areas?</p>	The explainee must shift the interaction, doing extra work to suspend the explainer's question and initiate a new thread.
43b6	Black holes/ Enhanced	<p>0:05:54 EX: [...] Now, let's consider the environment around a supermassive black hole. What do you think happens to the stars and other objects that get too close to the event horizon? Are they slowly pulled in, or is there a more dramatic fate that awaits them?</p> <p>0:09:10 EE: I believe that the force of attraction intrinsic to the existence of the black hole will lead to the disintegration of these stars and consequently the incorporation of their matter into the disk surrounding the black hole.</p>	Both parties co-shape the interaction. Much effort is required from the explainee for this to happen.
2417	Sleep/ Enhanced	<p>0:03:38 EE: I often find myself feeling tired throughout the day, but i am still getting the 7-9 hours of sleep that you said.</p> <p>0:03:45 EX: Feeling tired despite getting 7-9 hours of sleep can be a bit puzzling. There are a few possible explanations for this. [...]</p>	The participants contributes personal experiences to the interaction, enabling the conversation to shift away from purely monological explanations.

Table 7: Relevant examples of the qualitative analysis, showing excerpts from selected dialogues between the explainer (EX) and the explainee (EE) that highlight interesting interaction patterns. The Participant ID (PID) column shows the first four characters of the unique participant identifier. For more details, refer to Section 7.