"I understand, but...": Towards a Comprehensive Account of the Explainee's Voice in Explanatory Dialogues

Andrea Zaninello[1,2,*], Petar Bodlović[3], Marcin Lewinski[4] and Bernardo Magnini[2]

[1]Free University of Bolzano, Italy
[2]Fondazione Bruno Kessler, Trento, Italy
[3]Institute of Philosophy, Zagreb, Croatia
[4]Universidade Nova de Lisboa, Portugal

Abstract

In this paper, we introduce IUBAS, the first annotation scheme that provides an in-depth analysis of the Explainee's reactions in explanatory dialogues. Current schemes, mainly focusing on answers to *what*, *how*, and, occasionally, *why* questions, lack the granularity to capture the full range of the Explainee's contribution. Our richer framework, grounded in argumentation and philosophical theory, distinguishes different kinds of explanation requests, feedback types, and critical questions. We provide empirical evidence of the effectiveness of the scheme through a set of experiments with three SOTA LLMs. The IUBAS scheme provides a more detailed understanding of how Explainees interact with Explainers in a dialogical setting, contributing to the development of more sophisticated and human-like conversational agents.

Keywords

explanatory dialogues, annotation scheme, explanations

1. Introduction

The ability to provide and understand explanations is crucial in human communication and cognitive development. Psychologists argue that explanation is a key mechanism by which we learn generalizations and theories about the world (e.g., in childhood development) [1, 2]. Similarly, the ability of an automated system to justify its predictions and provide human-understandable explanations for some given facts has been a key research objective since the dawning of Machine Learning (ML) and Artificial Intelligence (AI). The recent rise of AI systems in highly specialised fields, such as the legal or medical domain for prediction and diagnostics, has brought the need for eXplainable AI (XAI) ever more to the forefront, but the computational modeling of agents capable of engaging in collaborative *explanatory dialogues* with users still represents a significant challenge [3, 4]. Researchers have long emphasized that interactive explanatory dialogues — where a user asks clarification questions and an AI system explains — are essential for trust and understanding in critical domains, such as education, healthcare, and law.

However, existing computational frameworks for dialogues typically rely on speech act theory [5], and describe explanations as answers to *what*, *how*, and *why* questions, while not accounting for the kind of feedback given for an explanation in a detailed manner. Explanations or argumentative dialogical turns are described - at the high level - as explanation requests by Explainees (e.g., "Why [*explanandum*]?"), explanatory answers by Explainers ("Because of [*explanans*]"), and possibly some basic feedback by the Explainee ("I (don't) understand.") [6, 7]. However, these frameworks lack the granularity to capture the complex interplay of challenges, clarifications, and personalized feedback, especially on the Explainee's side, that characterize real-world explanatory dialogues.

In order to contribute to this line of research, we introduce *IUBAS* (the *"I Understand, But..." Annotation Scheme*), a novel framework that goes beyond the simplest kinds of "question-answer-feedback" interactions, offering a more fine-grained approach to labelling the Explainee's reactions in explanatory dialogues. Our annotation scheme distinguishes between types of *explanation requests*, *feedback*, and *critical questions*, and incorporate contrastivity and motivation as key dimensions of our proposal. To empirically verify the effectiveness of the scheme, we perform experiments onthe task of predicting dialogue quality with three recent LLMs. Our results demonstrate that automatically annotating the Explainee's turns in a corpus of explanatory dialogues help achieve comparable or higher performance in comparison with current frameworks, confirming the central role of the Explainee in modelling effective explanatory dialogues[1].

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
azaninello@fbk.eu (A. Zaninello); pbodlovic@ifzg.hr (P. Bodlović); m.lewinski@fcsh.unl.pt (M. Lewinski); magnini@fbk.eu (B. Magnini)
0000-0001-9998-1942 (A. Zaninello); 0000-0001-8430-2599 (P. Bodlović); 0000-0001-7116-9338 (M. Lewinski); 0000-0002-0740-5778 (B. Magnini)

[1]Code and annotation here: github.com/andreazaninello/iubas

2. Background

The study of explanations in philosophy and argumentation theory covers a wide range of questions. Researchers have focused on distinguishing **explanations** from other forms of reasoning, such as clarifications and arguments, highlighting the difference in their core function. Explanations differ from **clarifications** in that, while the latter simply aim at understanding, explanations aim at increasing knowledge, carrying greater illocutionary force. Moreover, while **arguments** aim to provide evidence for a doubted claim, explanations seek to account for (e.g., provide a cause for) an already accepted, uncontroversial statement [8, 9, 10, 11]. This distinction becomes evident in the medical context, where a doctor might request an *explanation* for a patient's dark urine (a belief in an already accepted symptom in which no justification is required) but may seek an *argument* for the diagnosis of hemolytic anemia (a hypothesis that requires justification).

Further research has investigated the formal and normative **dimensions of explanations**, concentrating on developing argument schemes and critical questions associated with common explanatory inferences, such as Inference to the Best Explanation [12, 13, 14, 15]. Pragmatic studies, on the other hand, focus on defining the **speech act** of explaining [16] and its communicative function in various contexts. A key pragmatic function attributed to explanations is the transfer or enhancement of understanding [17, 18, 19], which becomes particularly crucial when communication is triggered by a lack of shared beliefs between the participants. In such instances, explanations act as a *local move* within a broader argumentative dialogue, facilitating smoother communication. For instance, an arguer will more easily develop an effective argumentative strategy once she understands "where the opponent is coming from", i.e., once the opponent explains why she doubts or rejects the arguer's thesis [20].

Analyzing explanations as individual moves within broader argumentative contexts, however, differs from studying genuinely explanatory dialogues. Explanatory dialogues[2] are strict dialectical procedures specifically designed to promote the transfer or enhancement of understanding. In **explanatory dialogues**, the prototypical setting is that of an *Explainer* clarifying or transferring their understanding of a phenomenon (represented as p) in response to an *Explainee*'s "Why p", "What is p?", "How does p work?" etc. questions [22, 23, 19]. The inherent dialogical nature of explanations stems from their communicative goal, which is strictly connected with the Explainee's level of understanding, (social and professional) role, curiosity, interests, beliefs, and doubts. Consequently, a comprehensive model of explanatory interactions should not only focus on the explanations provided, but also on the request for and reception of such explanations. In addition, the development of annotation schemes for explanatory dialogues is also crucial for training automatic dialogue systems and evaluating their ability to engage in effective knowledge transfer.

For simplicity, throughout the paper, we will assume the following definitions and notation.

- *Phenomenon* (***p***): event, fact, evidence, effect discussed in the dialogue; its existence is a precondition for explanatory dialogue (e.g., medical condition)
- *Explanandum* (***E***): event, fact, evidence, effect *in that* it requires explanation or understanding (e.g., medical symptom)
- *Explanans* (***H***): event, fact, hypothesis, cause of E that provides explanation or understanding (e.g., disease or medical injury)
- *Explainer* (***Er***): who is clarifying or transferring their understanding of E through the stating of H
- *Explainee* (***Ee***): who is requesting, giving feedback on or challanging an explanation H for some given E

3. Related work

Models of Explanatory Dialogues

Despite its importance, the field of explanatory dialogues remains relatively understudied compared to that of argumentation in general. Nonetheless, some researchers have studied this phenomenon, contributing to the understanding and modeling of such interactions. Cawsey [22] focuses on human-computer interactions, emphasizing the need for AI systems to respond to user feedback and refine explanations based on their understanding and background knowledge. Cawsey proposes content-related rules for structuring non-interactive explanations and dialogue rules for guiding the interactive process. Moore [23] highlights the role of explanations in facilitating understanding and learning. She proposes four key requirements for interactive explanation systems: naturalness, responsiveness, flexibility, and sensitivity. These requirements emphasize the need for AI systems to engage in natural conversation, adapt to user needs, and be sensitive to contextual factors. Walton [9, 19, 24] present a broader model of explanatory dialogue, characterizing it based on initial situations, collective goals, and rules governing different dialogue stages. Walton [25] distinguishes between explanatory and clarificatory dialogues, noting that clarifications focus on resolving ambiguities in expressions or speech acts while explanations target the understanding of events or facts.

[2] Sometimes also referred to as "explaining dialogues" or "dialogical explanations" [21, 6].

In recent years, Arioua and Croitoru [26] formalized and extended Walton's model, proposing a more flexible protocol that allows for backtracking and dialectical shifts between explanatory and argumentative dialogue. Rohlfing et al. [27] advocated for a social and interactive approach to AI explainability, emphasizing the co-construction of understanding through dialogue. Wachsmuth and Alshomary [6] analyzed human-to-human explanatory dialogues, focusing on linguistic patterns and adaptations based on user proficiency levels and Feldhus et al. [7] revised Wachsmuth and Alshomary [6]'s proposal with an adaptation to a pedagogical setting. More recently, Zaninello and Magnini [28] focused on the co-construction of knowledge in the medical domain, showing that LLMs benefit from a dialogical structure of explanations. Similarly, Fichtel et al. [29] presented a study demonstrating that LLMs can partly engage in co-constructive explanation by fostering user engagement but still struggle to adapt explanations based on user understanding. However, while recognizing the central role of the Explainee, they do not provide a comprehensive framework to model the Explanee's contribution in the co-construction of understanding.

Annotation Schemes

As mentioned in the previous sections, various models of explanatory dialogues have been proposed, each focusing on different aspects of the interaction. However, within the computational linguistic field, few comprehensive annotation schemes can be found. In the following section, we introduce two of the most prominent annotation schemes: the *5-levels* scheme, proposed by Wachsmuth and Alshomary [6], and the *Rewired* scheme by [7], an extension of the *5-levels* proposal.

The *5-levels* scheme [6] annotates each turn of a dialogue according to three different dimensions, resulting in a three-dimensional annotation for each turn where only one tag for dimension is allowed. The dimensions are: the discussed *topic* (T), the *dialogue act* (D), and the *explanation move* (E).

Dimension (T) recognizes that participant might be discussing the main topic (e.g. climate change), a subtopic (e.g., temperature increase), or some (un)related topic (e.g., greenhouse gas emissions). Dimension (D) is based on speech act theory and is derived from the DIT++ Taxonomy of Dialogue Acts[3] [30, 5], providing a coarse account of the type of question asked, whether an answer confirms or disconfirms whay previously asked, and whether a given statement agrees, disagrees or provides more information on a certain concept. The third dimension (E) provides a taxonomy of explanation moves in dialogue, including checking understanding or prior knowledge, giving or requesting explanations, signaling (non-)understanding, providing feedback, assessments, or extra information, and a catch-all for any other moves (see Table **??**).

The *5-levels* scheme was used to annotate the *Wired* [6] and the *ELI5* [21] datasets (see Section 3). In both datasets, annotation is realized at the turn level on the three dimensions (T, D, and E), where a turn corresponds to either the Explainer or the Explainee taking the floor. Each turn can be made of one or more utterances. This scheme provides a high-level categorization of explanatory dialogue acts but, as mentioned, mainly focuses on the Explainer's contribution, as can be seen from Table 11.

The *Rewired* scheme [7] is an extension of the *5-levels* scheme that proposes to add a new layer of annotation on top of the three proposed by Wachsmuth and Alshomary [6], drawing from pedagogical studies and teaching practice. The primary difference lies in the introduction of 10 *teaching acts* (T) in the new scheme. This new layer, focused on teaching strategies, such as assessing prior knowledge, proposing lesson steps, engaging in active experience, etc. allows for a more granular analysis of the instructional process, highlighting how teachers manage classroom interactions and instructional delivery.

Datasets

Despite their importance and relevance, explanatory dialogue data are scarse, as they are difficult to collect and analyze. One of the few available datasets is the *5-levels* "Wired" dataset [6], a corpus of 65 English dialogues from Wired's *5 Levels* video series, where 13 topics are discussed and explained to five explainees of varying expertise, resulting in 65 dialogues for a total of 1550 turns. Other available datasets rely on the crawling of discussions online, such as those in blogs and forums. For instance, the *ELI5-dialogues* corpus contains 399 daily-life explanatory dialogues from the Reddit forum "Explain Like I am Five" (ELI5). We introduce one example dialogue from this dataset in Table 9.

4. Accounting for the Explainee's Contribution: The IUBAS Annotation Scheme

As highlighted in Section 3, current dialogue annotation schemes recognize basic explanatory requests, modelled as "what-", "how-", and "why-" questions, which they categorize under "information-seeking" dialogical

[3] https://dit.uvt.nl

functions [5, 19, 6]. Such schemes also acknowledge basic feedback like "signal understanding" or "signal non-understanding" [6]. However, they usually do not recognize complex requests that include contrast classes and motivations, and different kinds of complex feedback that might include, e.g., qualifications, explanatory remarks, or critical questions. Complex requests and feedback are typical in real-world explanatory dialogues. While current accounts underline the dynamic nature of explanatory dialogues, they underestimate the importance of directly considering the Explainee's needs, contextual factors, and the co-construction of understanding, which are, however, vital to fully understand explanatory interactions.

This limited approaches neglect the **contrastive** nature of explanations, where an Explainee might seek to understand why a particular explanandum (E) is the case, instead of alternative possibilities (E^*) [31, 18]. Furthermore, the **motivations** behind the Explainee's questions are often ignored, neglecting the valuable contextual information that motivates their doubts and inquiries [20], which in turn also has important implications on the Explainer's **reaction** itself. For instance, once the Explainer understands what, exactly, puzzles or confuses the Explainee (where does her explanatory request "come from") the Explainer can provide a more effective, tailor-made, explainee-centered response. She can focus on the aspects of the problem that the Explainee considers most relevant and choose the effective communicative strategy sensitive to the required level of detail, requested type of information, etc.

To improve the current research, we propose to integrate existing accounts with *IUBAS*, a multi-dimensional annotation scheme that captures the diverse nature of Explainees' dialogical contributions and reactions. Our proposed scheme aims to address the limitations of the previous schemes by:

1. Providing a more fine-grained categorization of explanation moves, capturing specific actions within the explanatory process, by applying the annotation at the *utterance* level and allowing one utterance to receive zero or more (E) tags[4].
2. Explicitly considering the Explainee's perspective and their active role in seeking and integrating new information.
3. Empirically demonstrate the effectiveness of modelling the Explainee's role in the dialogue through as set of experiments on dialogue quality prediction.

Table 1 presents a summary of our proposed scheme, which we explain, motivate and exemplify in the next sections. A finer-grained comparison with the *5 levels* scheme can also be found in the Appendix, Table 11.

[4]As exemplified in Table 9, we implicitly assume that a tag is expressed at the utterance level and is automatically projected onto the next utterances until a new (E) tag is expressed

4.1. Explanation Requests

Explanation requests are the dialogical moves that, typically, initiate the explanatory process. They signal the Explainee's need for understanding and provide a target for the Explainer's efforts. We distinguish between different types of requests based on two key criteria: **contrastivity** and **motivation**.

4.1.1. Contrastivity: Basic vs. Contrastive Request

Basic explanation requests simply refers to (or targets) the *explanandum*, the event or phenomenon requiring explanation (Table 2).
The basic explanatory why-request is recognized in argumentation theory [32, 9, 19, 24, 33, 20], but, for the most part, ignored in contemporary annotation schemes. For instance, although [5] acknowledge that dialogue acts can be used to provide justifications and explanations, they focus on "check questions," "choice questions" and "set questions." Along similar lines, [6] emphasize the importance of "check", "how" and "what" questions.
Contrastive explanation requests, on the other hand, explicitly introduce a contrastive class, highlighting the specific aspects of the explanandum that require clarification (Table 3).
This distinction, while prevalent in philosophical literature on explanation [31, 17, 18, 16] is often overlooked in dialogue annotation schemes. Incorporating contrastive requests allows for a more precise representation of the Explainee's information needs, emphasizing the specific aspects of the explanandum that should be understood. Basic and contrastive requests, as exemplified in Tables 2 and 3, introduce questions that (might) require different explanations. So, defining a contrast class sets initial normative boundaries for selecting an adequate *explanans*.

4.1.2. Motivation: Pure vs. Motivated Request

Pure explanation requests directly inquire about the explanandum (or some aspect of explanandum, if they include contrast class) without further elaboration. In contrast, **motivated explanation requests** introduce further information about the Explainee's cognitive and communicative needs (Table 4). By motivating their requests, Explainees explicitly inform the Explainer what confuses them about the explanandum, or, in other words, what exactly stands in the way of transferring understanding. Such additional information promotes effective communication, and might at times even be necessary for formulating an adequate explanans.
Such additional considerations, inspired by the works of [20] and [34] allow us to capture the broader context of

Domain	Type	Subtype	Description	Tag
(R)	**Basic**	Pure	Why E?	**R01**
		Motivated	Why E, given that M?	**R02**
	Contrastive	Pure	Why E, instead of E*?	**R03**
		Motivated	Why E, instead of E*, given that M?	**R04**
(F)	**Positive Basic**	Assert understanding	I understand H.	**F01**
	Positive Complex	Demonstrate understanding	I understand. So...	**F02**
		Qualified understanding	I understand. But...	**F03**
		Critical challenge	I understand. However... [critical question]	**F04**
	Negative Basic	Assert non-understanding	I don't think H explains E. I rather think H*.	**F05**
	Negative Complex	Request for clarification	I don't think H explains E. Can you clarify H?	**F06**
		Critical challenge	I don't think H... In fact [critical question]	**F07**

	Types of Critical Challenges	Description	Tag
(C)	Comparative plausibility	Is H the best available hypothesis?	**C01**
	Epistemic distance	To what extent is H better than the "second-best" alternative hypothesis H*?	**C02**
	Generative completeness	Is the pool of plausible hypotheses complete (big enough)?	**C03**
	Non-comparative plausibility	Is H sufficiently plausible in itself?	**C04**
	Causal accuracy	Does H accurately cause E (does H undergenerate or overgenerate)?	**C05**
	Causal responsibility	Is H a responsible (pragmatically relevant, immediate) cause of E?	**C06**
	Explanandum reliability	Is E reliable and complete (are there false positives or false negatives: undetected symptoms)?	**C07**
	Pragmatic considerations	What are the pragmatic costs or benefits of accepting H (rather than H*)?	**C08**

Table 1
The three IUBAS categories with description and tags, including critical challenge types.

Table 2
Basic explanation request.

Move	Example
Explainer: *E.*	Mark has a cough that won't go away.
Explainee: *Why E?*	Why does Mark have a cough that won't go away?

Table 3
Contrastive explanation request.

Move	Example
Explainer: *E.*	Mark has a cough that won't go away.
Explainee: *Why E, instead of E*?*	Why does Mark have a cough that won't go away, instead of a temporary cough, or no cough at all?

the explanatory request, including the Explainee's background knowledge, assumptions, and potential concerns.

4.2. Explainee's Feedback

Once the Explainer offers an explanation, the Explainee typically provides feedback, signaling her understanding or lack thereof. We differentiate between positive and negative feedback, further distinguishing between basic and complex variants.

4.2.1. Polarity: Positive vs. Negative Feedback

Positive feedback expresses the Explainee's comprehension of the phenomenon, i.e., offered explanation. **Negative feedback** signals a failure to understand, prompting further elaboration or clarification from the Explainer (Table 5).

Table 4
Motivated explanation request.

Move	Example
Explainer: *E.*	Mark has a cough that won't go away.
Explainee: *Why E (instead of E*), given that M?*	Why does Mark have a cough that won't go away, given that he has never smoked cigarettes?

Table 5
Positive and negative feedback.

Move	Example
Explainer: *Because of H.*	Because Mark has cancer.
Explainee: *Positive feedback*	I understand why cancer would explain Mark's cough.
Explainee: *Negative feedback*	I don't understand why cancer would explain Mark's cough.

Table 6
Complex positive feedback: demonstrating understanding.

Move	Example
Explainee: *I understand. So,...*	I understand. So, (you're saying that) Mark's life is at risk and he should immediately start with chemotherapy...

4.2.2. Complexity: Basic vs. Complex Feedback

Basic feedback provides a straightforward assessment of understanding without further elaboration. In contrast, **complex feedback** incorporates additional remarks, questions, or challenges.

4.2.3. Types of Complex Positive Feedback

Complex Positive Feedback can take several forms:

1. **Demonstration of understanding:** The Explainee may provide additional information or draw inferences to demonstrate their grasp of the explanation (Table 6).
2. **Qualified understanding:** The Explainee may signal partial understanding, acknowledging the need for further clarification on specific aspects of the explanation (Table 7).
3. **Understanding with Critical Challenge:** While understanding the nature of explanation (or conditionally understanding the phenomenon), the Explainee may challenge its plausibility, demanding further justification (Table 8).

Table 7
Complex positive feedback: qualified understanding.

Move	Example
Explainee: *I understand, but...*	I understand, but what kind and stage of cancer are we talking about?

Table 8
Complex positive feedback: understanding with critical challenge.

Move	Example
Explainee: *I understand. However, ...*	I understand that lung cancer explains this kind of cough. However, is another diagnosis still possible? Can you still run some more tests?"

This type of feedback, both positive and negative (see Section 4.2.4) often introduces **critical questions** (see Section 4.3 and Table 7).

4.2.4. Types of Complex Negative Feedback

Complex negative can also be analyzed into:

1. **Request for clarification:** The Explainee may point to specific concepts or aspects of the explanation they find unclear.
2. **Critical challenge:** The Explainee may directly challenge the plausibility of the explanation, either categorically rejecting it or requesting further justification.

As seen for their positive counterpart, critical challenges can introduce **critical questions** (Section 4.3).

4.3. Explainee's Critical Questions

Critical questions challenge the explanation and its underlying assumptions. They target various aspects of the explanation, testing its plausibility, completeness, and relevance. Inspired by existing literature on Inference to the Best Explanation, argument schemes and critical questions [35, 36, 37, 12, 15], we propose a typology of critical questions tailored to why-explanations. We categorize critical questions according to the *specific aspect of explanation* they target, as summarized in Table 1 and further exemplified in Table 7.

5. Comparative annotation

In Table 9, we present a comparative analysis of an example dialogue from the *ELI5* corpus, annotated through

the "5-levels" and our "IUBAS" scheme.

IUBAS allows for a finer-grained account of the Explainee's request (e.g. U0, where we can specify that the explanation request is based on an implicit comparison with a complementary group). Also, we can better account for shifts in the explanation move within a turn (e.g. U4-6), as well as combinations of moves within a single turn (e.g. U7). This provides a more precise account of the conversational flow and, crucially, as this example suggests, it seems that providing explanations is not limited to the Explainer's role, and neither does feedback only originate from the Explainee. This observation, once generalised over a broader set of examples, could challenge the traditional view of the Explainer/Explainees roles, a phenomenon which can be analysed in detail through our scheme.

Also, our account of the different types of feedback request (e.g. U7, U8-9) highlight that the Explainee's reaction strongly influences the kind of explanation provided and participates in the co-construction of the explanation process. Finally, IUBAS is organized hierarchically, which makes it possible to navigate its tree-like structure and easily reconstruct the analysis of the explanatory move (Figure 1). Moreover, its structure allows for flexibility in terms of the level of granularity needed for a specific analysis.

6. Experiments

We conduct our experiments on the ELI5 dialogue quality assessment task introduced by Alshomary et al. (2024). This corpus consists of explanatory dialogues (399 in total) from the Reddit "Explain Like I'm Five" forum, each labeled with a ground-truth explanation quality score on a 1–5 Likert scale. We integrate the proposed IUBAS scheme into this task by automatically annotating the explainee turns and evaluating its impact on quality prediction.

IUBAS Annotation with GPT-4.1. To obtain IUBAS labels for the Explainee's turns, we employed the GPT-4.1[5] model to perform annotation in a zero-shot manner. We targeted only those turns where the *Explainee* explicitly participates in the dialogue, corresponding to the categories **E04** (Request Explanation) and **E07** (Request Feedback) in the original 5-level annotation scheme of Alshomary et al. These are the turns where the Explainee asks a question or provides feedback, i.e., the utterances that reflect the Explainee's reaction and understanding. For each such turn, GPT-4.1 was prompted with the dialogue context and the definition of the IUBAS categories, and it generated a IUBAS tag capturing the turn's properties, choosing among: **R** (type of explanation request, e.g., basic vs. contrastive), **F** (feedback type, e.g., positive vs. negative understanding), and **C** (presence of any critical follow-up or clarification request). This automatic labeling process produced a set of IUBAS annotations for all relevant Explainee turns in the ELI5 corpus, increasing the original labelling by approx. 20%. The resulting enriched dataset contains, for each relevant Explainee utterance, an associated label (R, F, C) indicating the Explainee's needs or feedback in that turn. We manually inspected a sample of the GPT-4.1 annotations to ensure they were coherent with the scheme's guidelines, and overall found the labels to be reasonable, providing a fine-grained view of the Explainee's role in the dialogue.

Quality Prediction Task Setup. Using the automatically annotated corpus, we replicate the dialogue quality prediction setup of Alshomary et al. (2024) to evaluate how the additional IUBAS metadata influences performance. The goal of the task is to predict the human-assigned quality score of a dialogue given the dialogue transcript (with or without annotations). We compare four input conditions:

- **No Annotation:** Each dialogue is given to the model as plain text, with no turn-level labels (baseline condition).
- **Original ELI5 Labels:** Each turn in the dialogue is followed by the original annotation tags for explanation move, dialogue act, and topic.
- **IUBAS Labels:** Each explainee turn is prefixed with its IUBAS labels (R, F, C values) as metadata, while explainer turns remain unlabeled.
- **Combined (ELI5 + IUBAS):** Both the original ELI5 turn labels and the IUBAS labels for Explainee turns are included.

We format the prompt for each dialogue by inserting the turn-level metadata (if any) immediately after each utterance, between square brakets, with a concise description of the tag itself (for example `(F01) Positive Basic Feedback - Assert understanding`). After presenting the entire dialogue, we append a final instruction asking the model to *"Rate the overall explanation quality on a 1–5 scale."* The model then outputs a single rating.

We evaluate three instruction-tuned LLMs: **Llama-3.1-8B-Instruct** [38], **Gemma-3-4b-it**[6], and **Qwen2.5-14B-Instruct-1M** [39]. We use HuggingFace's lm_eval harness [40] in the multiple choice mode, asking the model to choose between a number from 1 to 5, indicating the dialogue quality. We report RMSE and MAE against human ratings of each model's prediction, and assess significance using a paired t-test.

[5] https://openai.com/index/gpt-4-1/

[6] https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf

Role	Uttr.	Text	5-levels	IUBAS (Ours)
EE	U0	Why are there not many "flamboyant" heterosexual males?	(**E04**) Request Explanation	(**R03**) (=**4.2.1**) Request Explanation: **Contrastive - Pure**
ER	U1	I think a lot of the flamboyance is actually an act, albeit an unintentional one.	(**E03**) Provide Explanation	(**E03**) Provide Explanation
	U2	It's a lot about fitting in with the culture.		
	U3	I know a handful of "straight" guys who were "turned" by my gay friends and in a year these previously straight-acting men are the gayest of the bunch.		
EE	U4	Thank you for not attacking my question and seeing it for the curiosity it is.	(**E07**) Provide Feedback	(**E07**) Provide Feedback
	U5	I do believe culture and fitting in does play a large role here.		(**F04** - **C03**) (=**7.1.2.3**) Feedback: **Positive - Complex - Critical challenge - Generative completeness**
	U6	But I haven't run into any flamboyant heterosexual males.		
ER	U7	I guess we'd have to look at straight males that were raised by really flamboyant parents and see how they turned out.	(**E07**) Provide Feedback	(**F02**) (=**7.1.2.1**) Feedback: **Positive - Complex - Demonstrative Understanding** (**E03**) **Provide Explanation**
EE	U8	I don't know if that would be considered cruel and unusual if done purposefully.	(**E03**) Provide Explanation	(**F07** - **C08**) (=**7.2.2.3**) Feedback: **Negative - Complex - Critical Challenge - Pragmatic considerations**
	U9	But undoubtedly there should be 2 flamboyant men that could care for a child better than at least some heterosexual couples.		(**E03**) Provide Explanation
ER	U10	Yea we'll have to do these experiments underground.	(**E07**) Provide Feedback	(**F02**) (=**7.1.2.1**) Feedback: **Positive - Complex - Demonstrative Understanding**

Table 9
Example of explanatory dialogue from the ELI5 corpus, rated high quality (4/5) and annotated using the 5-levels scheme in the original release [6]. *ER* and *EE* indicate the Explainer's and the Explainee's turn respectively. [*U0, U1, etc.*] indicate utterances (our addition). The *5-levels* column indicates the annotation of the "explanatory move" dimension according to Alshomary et al. [21]. The *IUBAS* column reports our alternative annotation using our proposed scheme. Green indicates additions of our annotation scheme, while blue indicates differences in our annotation of the dialogue for categories already present in the *5-levels* scheme.

Model	Annotation	RMSE	MAE	p-value
LLaMA	No annotation	1.43	1.00	0.010
	ELI5-only	1.42	0.96	0.770
	IUBAS-only	**1.36**	**0.96**	0.109
	IUBAS-only (C)	1.36	0.97	0.027
	IUBAS-only (F)	1.38	0.97	0.070
	IUBAS-only (R)	1.38	0.99	0.009
	ELI5 + IUBAS	1.38	0.96	0.333
Gemma	No annotation	1.61	1.17	<1e-40
	ELI5-only	1.40	1.01	<1e-21
	IUBAS-only	**1.38**	**0.99**	<1e-12
	IUBAS-only (C)	1.44	1.06	<1e-26
	IUBAS-only (F)	1.44	1.05	<1e-17
	IUBAS-only (R)	1.43	1.04	<1e-21
	ELI5 + IUBAS	1.38	1.01	<1e-4
Qwen	No annotation	1.61	1.16	<1e-28
	ELI5-only	1.41	1.01	<1e-14
	IUBAS-only	1.46	1.04	<1e-20
	IUBAS-only (C)	1.48	1.05	<1e-21
	IUBAS-only (F)	1.47	1.05	<1e-19
	IUBAS-only (R)	1.50	1.08	<1e-24
	ELI5 + IUBAS	**1.40**	**1.02**	<1e-17

Table 10
Prediction error (RMSE and MAE) and paired t-test p-values for each model and annotation strategy. Lower is better. Bold = best per model across both measures.

6.1. Results and Analysis

Table 10 summarizes performance. Across all models, incorporating IUBAS annotations improves predictive accuracy over the unannotated baseline. Notably, the *IUBAS-only* condition consistently outperforms the *ELI5-only* setup for LLaMA and Gemma models (e.g., RMSE 1.36 vs. 1.42 for LLaMA). The best overall performance is typically achieved by the *combined* condition (ELI5+IUBAS), confirming the complementarity of the two annotation types.

Ablation experiments on IUBAS dimensions show that the *F-only* and *C-only* variants perform nearly as well as the full IUBAS scheme, while *R-only* annotations provide slightly smaller gains. The strongest single dimension was *F-only* for Gemma, while *C-only* was best for LLaMA. All annotation-enhanced variants significantly outperform the no-label baseline ($p < 0.05$).

7. Conclusion

In this paper, we introduced IUBAS, a framework that contributes to a richer understanding of the Explainee's role within explanatory dialogues. We incorporate contrastivity and motivation alongside a categorization of feedback and critical questions, providing a more comprehensive account for analyzing and modeling such interactions. By adopting this scheme, we can move towards developing more sophisticated conversational AI systems capable of engaging in truly human-like explanatory dialogues, ultimately enhancing communication effectiveness and fostering deeper understanding.

References

[1] T. Lombrozo, Explanation and abductive inference, The Oxford Handbook of Thinking and Reasoning (2012).

[2] M. T. Chi, M. Bassok, M. W. Lewis, P. Reimann, R. Glaser, Self-explanations: How students study and use examples in learning to solve problems, Cognitive science 13 (1989) 145–182.

[3] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in ai, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, ACM, 2019. URL: http://dx.doi.org/10.1145/3287560.3287574. doi:10.1145/3287560.3287574.

[4] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988. doi:https://doi.org/10.1016/j.artint.2018.07.007.

[5] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, D. Traum, Towards an ISO standard for dialogue act annotation, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/560_Paper.pdf.

[6] H. Wachsmuth, M. Alshomary, "mama always had a way of explaining things so i could understand": A dialogue corpus for learning to construct explanations, 2022. arXiv:2209.02508.

[7] N. Feldhus, A. Anagnostopoulou, Q. Wang, M. Alshomary, H. Wachsmuth, D. Sonntag, S. Möller, Towards modeling and evaluating instructional explanations in teacher-student dialogues, in: Proceedings of the 2024 International Conference on Information Technology for Social Good, GoodIT '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 225–230. URL: https://doi.org/10.1145/3677525.3678665. doi:10.1145/3677525.3678665.

[8] M. Di Maro, M. Di Bratto, S. Mennella, A. Origlia, F. Cutugno, et al., Argumentation in recommender

dialogue agents (arda): An unexpected journey from pragmatics to conversational agents, OPEN LINGUISTICS 11 (2025).

[9] D. Walton, A new dialectical theory of explanation, Philosophical Explorations 7 (2004) 71–89. doi:10.1080/1386979032000186863.

[10] G. R. Mayes, Argument explanation complementarity and the structure of informal reasoning, Informal Logic 30 (2010) 92–111. doi:10.22329/il.v30i1.419.

[11] T. Govier, Problems in Argument Analysis and Evaluation, Windsor Studies in Argumentation, University of Windsor, 2018. URL: https://books.google.hr/books?id=pulfDwAAQBAJ.

[12] D. Walton, C. Reed, F. Macagno, Argumentation Schemes, Cambridge University Press, New York, 2008.

[13] J. H. M. Wagemans, Argumentative patterns for justifying scientific explanations, Argumentation 30 (2015) 97 – 108. URL: https://api.semanticscholar.org/CorpusID:56085286.

[14] S. Yu, F. Zenker, Peirce knew why abduction isn?t ibe–a scheme and critical questions for abductive argument, Argumentation 32 (2017) 569–587. doi:10.1007/s10503-017-9443-9.

[15] P. Olmos, Metaphilosophy and argument: The case of the justification of abduction, Informal Logic 41 (2021) 131–164. doi:10.22329/il.v41i2.6249.

[16] G. Gaszczyk, Helping others to understand: A normative account of the speech act of explanation, Topoi 42 (2023) 385–396. doi:10.1007/s11245-022-09878-y.

[17] P. Lipton, Inference to the Best Explanation, International library of philosophy and scientific method, Routledge/Taylor and Francis Group, 2004. URL: https://books.google.hr/books?id=WIfYNExpSC0C.

[18] S. R. Grimm, The goal of explanation, Studies in History and Philosophy of Science Part A 41 (2010) 337–344. doi:10.1016/j.shpsa.2010.10.006.

[19] D. Walton, A dialogue system specification for explanation, Synthese 182 (2011) 349–374. doi:10.1007/s11229-010-9745-z.

[20] J. A. van Laar, E. C. W. Krabbe, The burden of criticism: Consequences of taking a critical stance, Argumentation 27 (2013) 201–224. doi:10.1007/s10503-012-9272-9.

[21] M. Alshomary, F. Lange, M. Booshehri, M. Sengupta, P. Cimiano, H. Wachsmuth, Modeling the quality of dialogical explanations, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italy, 2024, pp. 11523–11536. URL: https://aclanthology.org/2024.lrec-main.1007.

[22] A. Cawsey, Explanation and Interaction: The Computer Generation of Explanatory Dialogues, ACL-MIT Press series in natural-language processing, Bradford Book, 1992. URL: https://books.google.hr/books?id=hQt1-7gA334C.

[23] J. Moore, Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context, A Bradford book, CogNet, 1995. URL: https://books.google.hr/books?id=nRx0QgAACAAJ.

[24] D. Walton, Abductive Reasoning, University of Alabama Press, 2014. URL: https://books.google.hr/books?id=DNqKAwAAQBAJ.

[25] D. Walton, The speech act of clarification in a dialogue model, Studies in communication sciences 7 (2007). URL: https://api.semanticscholar.org/CorpusID:149373911.

[26] A. Arioua, M. Croitoru, Formalizing explanatory dialogues, in: Scalable Uncertainty Management, 2015. URL: https://api.semanticscholar.org/CorpusID:7365540.

[27] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Häb-Umbach, I. Horwath, E. Hüllermeier, F. Kern, S. Kopp, K. Thommes, A.-C. Ngonga Ngomo, C. Schulte, H. Wachsmuth, P. Wagner, B. Wrede, Explanation as a social practice: Toward a conceptual framework for the social design of ai systems, IEEE Transactions on Cognitive and Developmental Systems 13 (2021) 717–728. doi:10.1109/TCDS.2020.3044366.

[28] A. Zaninello, B. Magnini, Medexpdial: Machine-to-machine generation of explanatory dialogues for medical qa, in: Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, 2024.

[29] L. Fichtel, M. Spliethöver, E. Hüllermeier, P. Jimenez, N. Klowait, S. Kopp, A.-C. N. Ngomo, A. Robrecht, I. Scharlau, L. Terfloth, A.-L. Vollmer, H. Wachsmuth, Investigating co-constructive behavior of large language models in explanation dialogues, 2025. URL: https://arxiv.org/abs/2504.18483. arXiv:2504.18483.

[30] H. Bunt, D. K. J. Heylen, C. Pelachaud, R. Catizone, D. R. Traum, The dit++ taxanomy for functional dialogue markup, 2009. URL: https://api.semanticscholar.org/CorpusID:60074224.

[31] F. I. Dretske, Contrastive statements, Philosophical Review 81 (1972) 411–437. doi:10.2307/2183886.

[32] C. Hamblin, Fallacies, University paperbacks, Methuen, 1970. URL: https://books.google.hr/books?id=bYYIAQAAIAAJ.

[33] J. Blair, C. Tindale, Groundwork in the Theory of Argumentation: Selected Papers of J. Anthony Blair, Argumentation Library, Springer Netherlands, 2011. URL: https://books.google.hr/books?

id=IM9p6GgnJAcC.
[34] M. Rescorla, Shifting the Burden of Proof?, The Philosophical Quarterly 59 (2008) 86–109. URL: https://doi.org/10.1111/j.1467-9213.2008.555.x. doi:10.1111/j.1467-9213.2008.555.x.
[35] G. H. Harman, The inference to the best explanation, Philosophical Review 74 (1965) 88–95. doi:10.2307/2183532.
[36] J. R. Josephson, S. G. Josephson (Eds.), Abductive Inference: Computation, Philosophy, Technology, Cambridge University Press, New York, 1994.
[37] D. Walton, Abductive, presumptive and plausible arguments, Informal Logic 21 (2001). doi:10.22329/il.v21i2.2241.
[38] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[39] A. Yang, B. Yu, C. Li, D. Liu, F. Huang, H. Huang, J. Jiang, J. Tu, J. Zhang, J. Zhou, et al., Qwen2. 5-1m technical report, arXiv preprint arXiv:2501.15383 (2025).
[40] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2024. URL: https://zenodo.org/records/12608602. doi:10.5281/zenodo.12608602.

Appendix

Limitations

While the manual annotation of a full dataset falls outside the scope of our current proposal, we believe that future work should involve testing the agreement between the automated annotation and human-annotation. Additionally, the proposed typology could be expanded to account for the different kinds of explanations and reasoning patterns on the Explainer's side, too.

Ethical Considerations

This research focuses on analyzing explanatory dialogue, and it is crucial to acknowledge the potential ethical implications of applying such schemes to real-world situations, especially in sensitive domains like healthcare or by covering topics such as ethnicity, physical ability, gender and sexual orientation (as in the case of the reported example in Table 9). Careful consideration should also be given to data privacy, informed consent, and potential biases in the annotation process.

Table 11
The IUBAS scheme (green) represented as an extension of the *explanatory move* (E) dimension of the *5-levels* scheme (blue).

[E] tag	Value	Description
1	Test understanding	Checking whether the listener understood the explanation.
2	Test prior knowledge	Checking the listener's prior knowledge of the topic.
3	Provide explanation	Explaining a concept or topic to the listener.
4	Request explanation	Requesting an explanation from the listener.
	Contrastivity	***Is a contrastive class introduced?***
4.1	**Basic**	Directly inquiring about *E*, the event or phenomenon requiring explanation.
4.2	**Contrastive**	Introducing a contrastive class, high-lighting specific aspects of *E*.
	Motivation	***Is additional information provided?***
4.1.1 (R01)	Basic - **Pure**	Why *E*?
4.1.2 (R02)	Basic - **Motivated**	Why *E*, given that *M*?
4.2.1 (R03)	Contrastive - **Pure**	Why *E*, instead of E^*?
4.2.2 (R04)	Contrastive - **Motivated**	Why *E*, instead of E^*, given that *M*?
5	Signal understanding	Informing the listener that their last utterance was understood.
6	Signal non-understanding.	Informing the listener that the utterance was not understood.
7	Provide feedback	Responding qualitatively to an utterance by correcting errors or similar.
	Polarity	***Does the feedback confirm or disconfirm* H?**
	Complexity	***Is the feedback simple or complex?***
7.1	**Positive feedback**	Agreeing with *H*.
7.1.1 (F01)	Positive - **Basic**	Agreeing with *H* without further elaboration.
7.1.2	Positive - **Complex**	Agreeing with *H* with further elaboration.
7.1.2.1 (F02)	Positive - Complex - **DU**	**Demonstrative understanding**: I understand. So...
7.1.2.2 (F03)	Positive - Complex - **QU**	**Qualified understanding**: I understand. But...
7.1.2.3 (F04)	Positive - Complex - **CC**	**Critical challenge**: I understand. However... [*critical question*] (see Table 7)
7.2	**Negative feedback**	Disagreeing with *H*.
7.2.1 (F05)	Negative - **Basic**	Disagreeing with *H* without further elaboration.
7.2.2	Negative - **Complex**	Disagreeing with *H* with further elaboration.
7.2.2.1	Negative - Complex - **P**	**Pure**: I don't think *H* explains *E*. I rather think H^*.
7.2.2.2 (F06)	Negative - Complex - **CR**	**Clarification request**: I don't think *H* explains *E*. Can you clarify $h \in H$?
7.2.2.3 (F07)	Negative - Complex - **CC**	**Critical challenge**: I don't think *H*... In fact [*critical question*] (see Table 7)
8	Provide assessment	Assessing the listener by rephrasing their utterance or giving a hint
9	Provide extra info	Giving additional information to foster a complete understanding
10	Other	Making any other explanation move

Figure 1: The hierarchical structure our IUBAS annotation scheme.

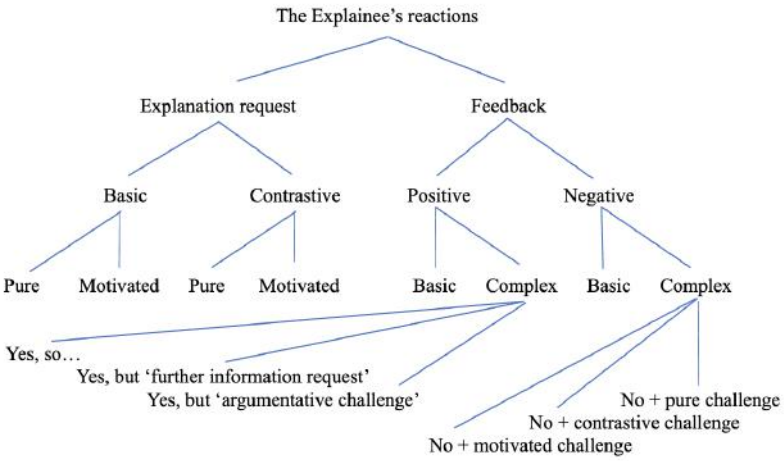

Table 12
Typology of critical questions for the Complex Negative Feedback's Critical challenges.

Tag	Question Type	Description (question)	Example
C01	**Comparative plausibility**	*Is H the best available hypothesis?*	Is 'lung cancer' the best explanation of Mark's symptoms among available explanations?
C02	**Epistemic distance**	*To what extent is H better than the "second-best" alternative hypothesis H*?*	If 'lung cancer' is the best hypothesis, is it significantly or only slightly better than the most plausible alternative hypothesis (e.g., asthma)?
C03	**Generative completeness**	*Is the pool of plausible hypotheses complete (big enough)?*	Did doctors overlook some promising hypotheses, to begin with (e.g., sinusitis)?
C04	**Non-comparative plausibility**	*Is H sufficiently plausible in itself?*	Even if 'lung cancer' is the best available explanation, is it likely?
C05	**Causal accuracy**	*Does H accurately cause E (does H undergenerate or overgenerate)?*	Does 'lung cancer' cause Mark's condition? Perhaps this diagnosis does not explain all the symptoms, or entails symptoms that were not detected.
C06	**Causal responsibility**	*Is H a responsible (pragmatically relevant, immediate) cause of E?*	Is 'lung cancer' the cause we are looking for? Perhaps we are dealing with multiple causes: the patient coughs because of lung cancer, but also because of contracting COVID-19.
C07	**Explanandum reliability**	*Is E reliable and complete (are there false positives or false negatives: undetected symptoms)?*	*Is cough the only symptom that needs to be explained?* Is it a real symptom (or is the patient faking it)?
C08	**Pragmatic considerations**	*What are the pragmatic costs or benefits of accepting H (rather than H*)?*	What is the cost of being mistaken if one proceeds as if the patient has cancer, or as if she has asthma?