Modeling the Quality of Dialogical Explanations

Milad Alshomary [†], **Felix Lange** [⋆], **Meisam Booshehri** [‡],
Meghdut Sengupta[*], **Philipp Cimiano**[‡] **and Henning Wachsmuth**[*]
[†] Columbia University, [⋆]Paderborn University, [*] Leibniz University Hannover, [⋆] University of Bielefeld
New York, USA; Paderborn, Germany; Hannover, Germany; Bielefeld, Germany
milad.alshomary@columbia.edu

Abstract

Explanations are pervasive in our lives. Mostly, they occur in dialogical form where an *explainer* discusses a concept or phenomenon of interest with an *explainee*. Leaving the explainee with a clear understanding is not straightforward due to the knowledge gap between the two participants. Previous research looked at the interaction of explanation moves, dialogue acts, and topics in successful dialogues with expert explainers. However, daily-life explanations often fail, raising the question of what makes a dialogue successful. In this work, we study explanation dialogues in terms of the interactions between the explainer and explainee and how they correlate with the quality of explanations in terms of a successful understanding on the explainee's side. In particular, we first construct a corpus of 399 dialogues from the Reddit forum *Explain Like I am Five* and annotate it for interaction flows and explanation quality. We then analyze the interaction flows, comparing them to those appearing in expert dialogues. Finally, we encode the interaction flows using two language models that can handle long inputs, and we provide empirical evidence for the effectiveness boost gained through the encoding in predicting the success of explanation dialogues.

Keywords: Corpus, Explanation, Discourse Annotation, Explainability

1. Introduction

Explanations play a significant role in our daily life. Typically, they are realized through dialogues, where one person is an *explainer* while the other takes the *explainee* position. The explainer's primary goal is to convey information about a particular concept or phenomenon to the explainee clearly and concisely. However, ensuring that the explainee understands an explanation successfully is challenging: Effective explanations require more than just information delivery. Expert explainers usually plan an explanation strategy by choosing appropriate explanation moves, dialogue acts, and topics to ensure optimal comprehension on the explainee side (Wachsmuth and Alshomary, 2022). Additionally, explainees may actively engage in dialogues by asking clarification questions and providing feedback to ensure they understand the information correctly (Madumal et al., 2019).

Most previous research has studied monological explanations (Fan et al., 2019; Situ et al., 2021), where an explainer provides a single-turn explanation, ignoring the role of the explainee in the interaction. However, Rohlfing et al. (2021) emphasized that both participants construct real-life explanations. While Madumal et al. (2019) and Wachsmuth and Alshomary (2022) found insightful interaction patterns in the dialogues between explainers and explainees, it remains unstudied what makes such dialogues successful. In daily life, explanations may fail, depending on various factors, including the level of expertise, communication style, and prior knowledge of the explainer and the explainee. Hence, building tools to assess humans in constructing successful explanation dialogues is crucial.

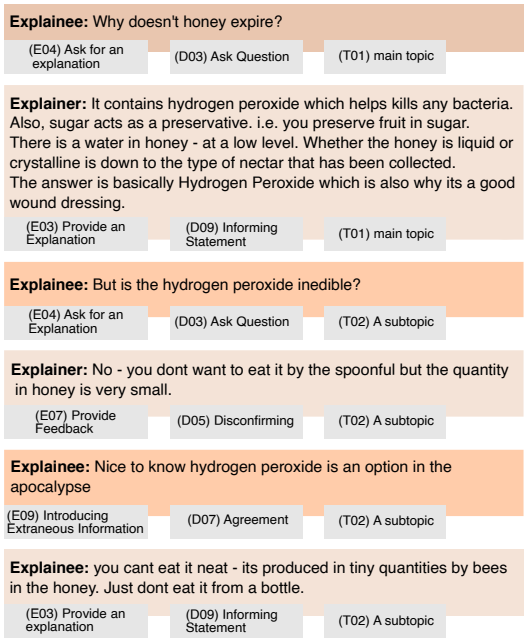

Figure 1: Example explanation dialogue from the ELI5 corpus introduced in this paper, annotated for explanation moves, dialogue acts, and topics

In this work, we take a first substantial step towards studying the quality of daily-life explanation dialogues concerning the explainee's understanding. We hypothesize that the interactions in explanation dialogues in terms of explanation moves,

LREC-COLING 2024, pages 11523–11536

dialogue acts, and topics correlate with explanation success. To study this hypothesis, we construct the first corpus of daily-life explanation dialogues. We then compare this corpus to existing expert explanation dialogues (Wachsmuth and Alshomary, 2022), and we evaluate the effectiveness of pre-trained language models in predicting the quality of explanation dialogues.

In particular, the created corpus consists of 399 daily-life explanation dialogues from the Reddit forum *"Explain Like I am Five (ELI5)"*. One example dialogue is shown in Figure 1. We annotate the corpus for the explanation quality of each dialogue as well as the interaction concepts of Wachsmuth and Alshomary (2022) (Section 3). Given the corpus, we analyze differences between daily-life and expert explanations in terms of explanation moves, dialogue acts, and topic relations. Matching intuition, we find that disagreement arises more often in daily life, reflecting the challenges an explainer faces while explaining a topic (Section 4).

To operationalize our findings, we assess whether a computational encoding of interaction flows can aid pre-trained language models in predicting the success of explanation dialogues. Specifically, we consider two popular language models on this task, namely *Longformer* (Beltagy et al., 2020) and *hierarchical attention transformers* (Chalkidis et al., 2022). As shown in Figure **??**, we augment their input with the interaction flow by prefixing each turn with its explanation move, dialogue act, and topic label. Our experiments show that adding all turn labels into the input of the *hierarchical attention transformers* results in the best error reduction on the task (Section 5).

To summarize, the main contributions of this paper are the following:

- A corpus of daily-life explanation dialogues annotated for interaction flow and quality
- Insights into the differences between daily-life and expert explanation dialogues
- First computational approaches to the assessment of explanation dialogues.

To foster future research, we make our corpus and all code publicly available.[1]

2. Related Work

Explanations have long been rather neglected in NLP research, but have recently gained more attention due to the increasing importance of explainable AI, XAI in short (Danilevsky et al., 2020). For XAI in general, Confalonieri et al. (2019) discussed what make a good explanation, and Halliwell et al. (2022) pointed out that assessing the quality of explanations is as important as it is challenging due to missing ground-truth information. The impact of the audience of an explanation was noted by Barredo Arrieta et al. (2020), which matches the social science view on explanations.

In particular, Miller (2019) emphasized the social aspects of explanations, arguing that explanation success depends not only on the quality of what is being said, but also on who is involved, the social context, and what actually needs to be explained in this context. Rohlfing et al. (2021) build on this view, clarifying that explaining in an intrinsically dialogical process in which the participants co-construct an explanation. They highlight the importance of successful communication between the explainer and explainee, which is a challenge that research needs to address adequately.

Nevertheless, most NLP research so far focused on one-way explanations ignoring the role of the explainee. In early work Jordan et al. (2006) analyzed the explanations of learners, whereas Fontan and Saint-Dizier (2008) modeled the discourse structure of monological explanations. Jansen et al. (2016) modeled the required explanations on exam answers, and Son et al. (2018) looked at causality in explanations. Like us, Fan et al. (2019) constructed a corpus of question answers from the Explain Like I am Five (ELI5) subreddit, an online community that provides simple explanations for questions asked by users. However, the instances are single question-answer pairs. Situ et al. (2021) proposed an approach to explain machine learning models' behavior by highlighting essential parts of the input based on their contribution to the model's decision. Wiegreffe and Marasović (2021) gives an overview of available datasets used in literature to model explanation in the field of XAI. Our work focuses on analyzing daily-life explanation dialogues rather than single-turn explanations.

Two related works have recently studied explanation dialogues: Madumal et al. (2019) analyzed the transcripts of 398 explanation dialogues in terms of explainer and explainee interactions and proposed an interaction protocol to model these interactions. Wachsmuth and Alshomary (2022) collected a dataset of 65 explanation dialogues between an expert and explainees of different expertise levels. They proposed a taxonomy to model explanation dialogues on three dimensions, dialogue acts, explanation moves, and topics. Similar to these works, we also deal with explanation dialogues, but we aim to assess the explanation quality of the dialogues computationally.

[1] Code and data can be found here https://github.com/MiladAlshomary/explanation-quality-assessment

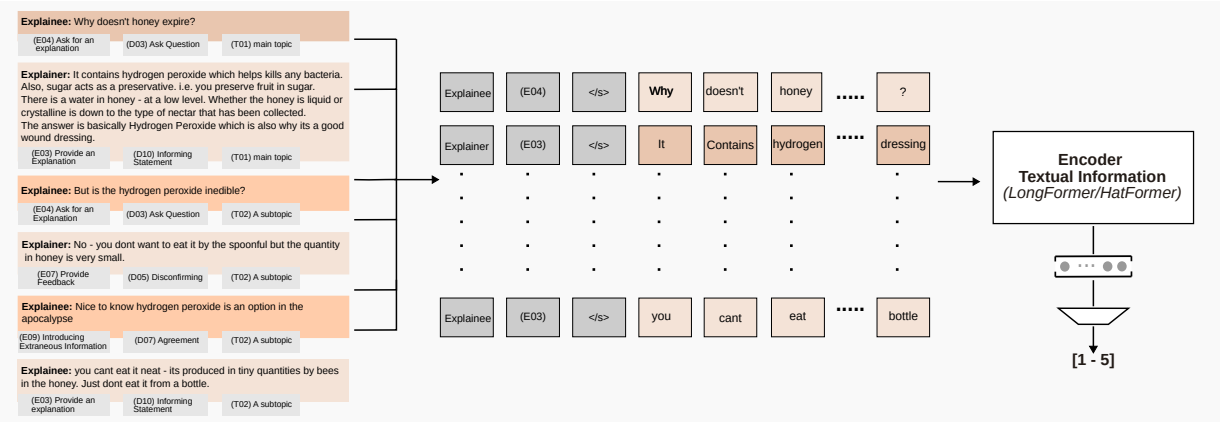

Figure 2: Our approach is to augment the input of language models with tokens reflecting the interaction flow in terms of either the explanation moves, dialogue acts, topic, or all together. Here, it is shown for the case of explanation moves.

3. Corpus Construction

This section introduces our procedure to acquire our corpus from the *"Explain Like I am Five (ELI-5)"* subreddit and to annotate it for our purposes.

3.1. Explanation Dialogue Acquisition

As mentioned in Section 2, most existing explanation datasets target single-turn explanations and are not in a dialogical form. Therefore, we curate a new explanation dialogue corpus. Similar to Fan et al. (2019), we use the *"Explain Like I am Five (ELI5)"* subreddit; however, we collect multi-turn dialogues rather than single explanations. As exemplified in Figure 1, on ELI5, a user (the *explainee*) posts a question about a particular topic requesting an easy-to-understand explanation. Others (*explainers*), in turn, interact with the explainee providing explanations. In some threads, these interactions turn into dialogues where the explainee elaborates their explanation need with questions or feedback, while explainers respond with clarifications. Such threads are in our focus.

For acquisition, we used the *Pushshift Reddit API*.[2] We collected posts over a span of three years (2019–2021), creating queries to extract the top 100 threads in terms of the number of comments in each month. For each thread, we extracted explanation dialogues as follows: We identified the thread creator as the explainee. We then extracted the first-level comments that scored high up-votes since they might consist of longer discussions, and we identified their authors as explainers. For each of these comments, we searched through the nested comments to extract the alternating interactions between the explainee and the explainer. To obtain meaningful interactions, we selected only threads with a minimum of six turns for the corpus.

This process resulted in 399 explanation dialogues, covering 204 questions that we consider to be the topics of the dialogues. The dialogues have a minimum of six turns, a maximum of 40, and an average length of 8.7 turns. The corpus consists of 3457 turns with an average of 64 tokens. We call this corpus *ELI5-dialogues*.

3.2. Flow and Quality Annotation

In this work, we study what types of interactions emerge in daily-life explanation dialogues and whether certain interaction patterns correlate with a dialogue's quality in terms of a successful understanding by the explainee. Therefore, we annotate our corpus with turn-level labels reflecting the interaction flow and quality scores on the dialogue level. In the following, we first summarize definitions of the annotation scheme we reused from previous work, explain the new annotation of explanation quality, and then describe the process we followed to annotate our corpus accordingly.

3.2.1. Interaction Flow

We annotate the role of each turn in an explanation dialogue following the annotation scheme of Wachsmuth and Alshomary (2022). The authors of this scheme proposed three aspects of interest to model explanation dialogues: explanation moves, dialogue acts, and topic relation. In the following, we describe each dimension in detail.

Explanation Moves Wachsmuth and Alshomary (2022) devised 10 explanation-specific moves that appear in explanation dialogues: (e_1) *Test understanding*, identifying whether the explainee understood what has been explained (e_2) *Test prior*

[2] https://github.com/pushshift/api

knowledge, checking the explainee's level of expertise; (e_3) *Provide explanation*, explaining any concept or a topic; (e_4) *Request explanation*, asking for an explanation (e_5) *Signal understanding* and (e_6) *Signal non-understanding* to indicate that what has been said is understood or not; (e_7) *Provide feedback*, responding by correcting errors or similar; (e_8) *Provide assessment*, responding by rephrasing previous utterance or giving a hint; (e_9) *Provide extra information*, providing additional information to foster a complete understanding; (e_{10}) *Other*, representing any other move.

Dialogue Acts As the authors, we also considered 10 dialogue acts from a standard taxonomy[3] to represent the communicative functions of turns: (d_1) *Check question*, (d_2) *What/How question*, (d_3) *Other question*, (d_4) *Confirming answer*, (d_5) *Disconfirming answer*, (d_6) *Other answer*, (d_7) *Agreeing statement*,(d_8) *Disagreeing statement*, (d_9) *Informing statement*, and (d_{10}) *Other*.

Topic Relation Capturing the relation between the turn-level and main topics can reveal different dynamics of explanation dialogues (Garfinkel, 2009). Therefore, we follow Wachsmuth and Alshomary (2022) in annotating four types of relatedness: (t_1) *Main Topic*, when the main topic is discussed; (t_2) *Suptopic*, representing a specific aspect of the main topic (e.g., *Music* and *Musical Instruments*); (t_3) *Related topic* another topic that is related to the main topic (e.g., *Black holes* and *Gravity*); and (t_4) *No/Other topic*, representing no change in the topic from previous turns.

3.2.2. Explanation Quality

Several works have explored different quality dimensions of explanations, including trustworthiness and informativeness (Barredo Arrieta et al., 2020). However, in order to avoid imposing assumptions about what makes an explanation dialogue successful, we follow a more straightforward approach to give a holistic score for each dialogue on a 5-point Likert scale, reflecting how satisfied the explainee is with the provided explanation. A score of 1 implies a fully failed explanation (no understanding visible), whereas a score of 5 means that the explanation was fully satisfactory (understanding clearly visible). Scores in between reflect different degrees of success in between.

3.2.3. Annotation Process

Specifically, we set up the 399 dialogues using the label-studio annotation tool (Tkachenko et al., 2020-2022) and recruited annotators on the UpWork platform.[4] We hired three content editors who are native English speakers and had more than 90% job success on the platform. The annotation task was to read the whole dialogue and perform two-level annotations on the turn and dialogue levels. The annotators were asked to choose an explanation move, a dialogue act, and a topic for each turn and score the dialogue. Annotation guidelines are provided in the supplementary materials. In terms of Fleiss' κ, the annotators had an agreement of 0.73 for explanation moves, 0.49 for acts, and 0.43 for topic relation. While these values partly reflect moderate agreement, they are still notably better than those reported by Wachsmuth and Alshomary (2022). For the quality scores, the agreement was 0.61 in terms of ordinal Krippendorff's α. We consolidated the annotations using MACE (Hovy et al., 2013) and split the corpus per topic question into 154 topics for training and 50 for testing. Table 1 shows the frequency of each of the annotated labels in the training and test splits.

		Train		Test	
Turn	**Label**	**#**	**%**	**#**	**%**
(t_1)	Main topic	1411	51.7	336	46.1
(t_2)	Subtopic	517	19	94	12.9
(t_3)	Related topic	346	12.7	130	17.8
(t_4)	No/Other topic	454	16.6	169	23.2
(e_1)	Test understanding	12	0.4	5	0.7
(e_2)	Test prior knowledge	13	0.5	4	0.5
(e_3)	Provide explanation	1012	37.1	244	33.5
(e_4)	Request explanation	823	30.2	217	29.8
(e_5)	Signal understanding	36	1.3	9	1.2
(e_6)	Signal non-underst.	85	3.1	23	3.2
(e_7)	Provide feedback	711	26.1	213	29.2
(e_8)	Provide assessment	14	0.5	4	0.5
(e_9)	Provide extra. Inf.	13	0.5	3	0.4
(e_{10})	Other	9	0.3	7	1
(d_1)	Check question	113	4.1	32	4.4
(d_2)	What/How question	349	12.8	83	11.4
(d_3)	Other question	462	16.9	118	16.2
(d_4)	Confirming answer	87	3.2	29	4
(d_5)	Disconfirming answer	105	3.8	21	2.9
(d_6)	Other answer	252	9.2	70	9.6
(d_7)	Agreeing statement	192	7	79	10.8
(d_8)	Disagreeing statement	364	13.3	86	11.8
(d_9)	Informing statement	733	26.9	184	25.2
(d_{10})	Other	71	2.6	27	3.7

Table 1: Turn label distribution in our corpus for the training and testing splits.

4. Corpus Analysis

In the following, we give insights into the nature of daily-life explanations by contrasting the corpus

[3]Taxonomy of Dialogue Acts, https://dit.uvt.nl

[4]Upwork, https://www.upwork.com

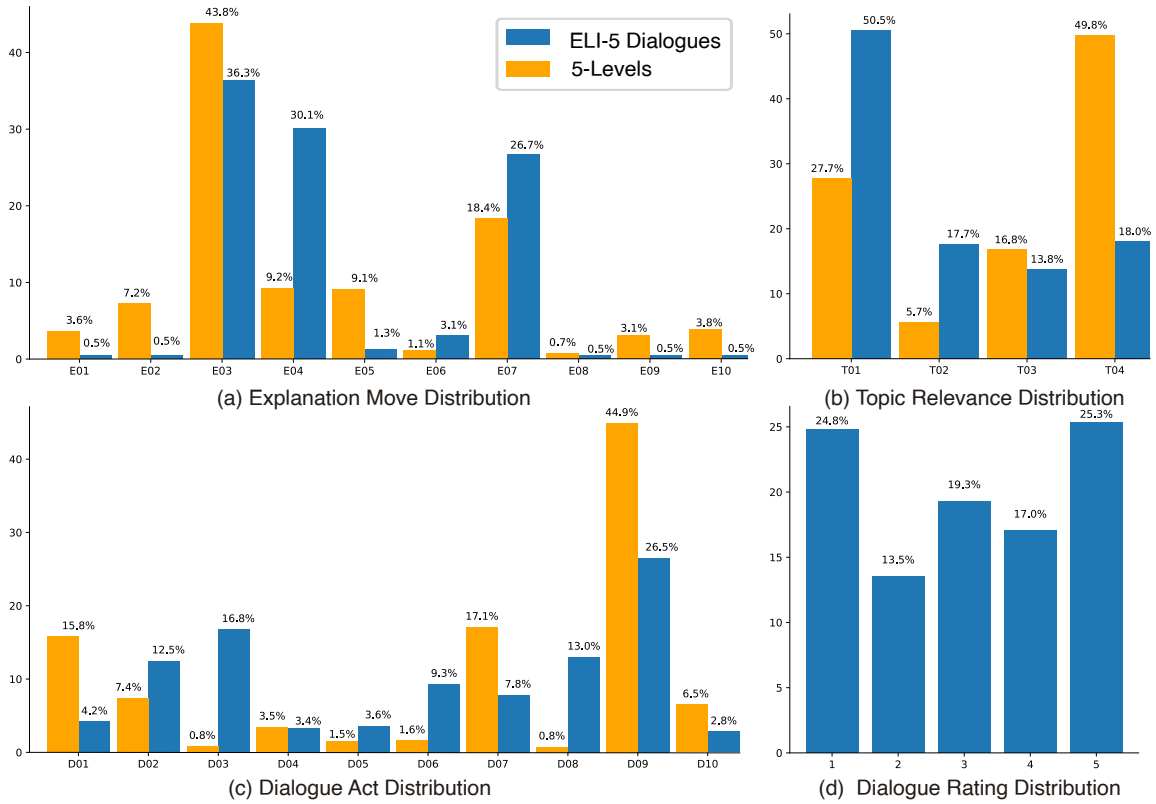

Figure 3: Explanation moves, dialogue acts, and topics distributions in our corpus and the 5-Levels the expert dialogues corpus of Wachsmuth and Alshomary (2022)

dialogues with the expert dialogues from the *5-Levels corpus* (Wachsmuth and Alshomary, 2022).

Turn Labels Figure 3a illustrates the distribution of explanation moves in the two corpora. Similar to the expert dialogues, the most frequent three explanation moves in our corpus are *provide explanation* (e_3), *request explanation* (e_4), and *provide feedback* (e_7), appearing in 36.3%, 30.1%, and 26.7% of the turns respectively. This is expected since all dialogues in our corpus start with a question that requests an explanation. In contrast, we see that our corpus contains a higher percentage of the move *signal non-understanding* (e_6, 3.1%), indicating the difficulty in achieving successful explanation dialogues. Moreover, unlike the expert dialogues, the daily-life dialogues contain few turns in which the explainer *tests the prior knowledge* of the explainee (e_2, 0.5% compared to 7.2%), or *test their understanding* (e_1, 0.5% compared to 3.6%).

As highlighted in Figure 3c, the most frequent dialogue act is *informing statement* (d_{09}), as in the 5-Levels corpus. However, we can see in our corpus fewer *check questions* (d_1) and *agreeing statements* (d_7) compared to the expert dialogues, but more of *disagreeing statements* (d_8). We attribute this to the fact that the explainer puts little effort into checking the understanding of the explainee — a move can be achieved by asking check questions. Besides, the controlled setup of the expert dialogues of Wachsmuth and Alshomary (2022) results in much agreement between the explainer and the explainee, while in daily-life dialogues, disagreement is more prominent.

As for the topic distribution Figure 3b, in our corpus, the discussed topic in each turn is primarily the *main topic* (50.5% of the turns) followed by the subtopic (17.7%) and others (18.0%). In contrast, related topics are much more discussed in expert dialogues than subtopics.

Explanation Success Figure 3d presents the quality score distribution of the dialogues. Our data is rather balanced, with scores of 1 (24.8%) and 5 (25.3%) being the most frequent. Moreover, we analyze the correlation between turn labels (explanation moves, dialogue acts, and topic relatedness) and the quality scores. Table 2 shows the frequency distribution of each turn label broken down by the quality scores. Regarding dialogue acts, labels such as *disagreement statement*, *disconfirming answer*, and *what/how questions* correlate more with low-quality dialogues, while informing and agreement statements correlate more with high-quality dialogues. Unexpectedly, looking at the explanation moves, we can see that *test-*

Explanation Moves	Freq.	Score Distribution 1	2	3	4	5
(E03) Provide Exp.	1256	22%	15%	25%	17%	21%
(E04) Ask Exp.	1040	25%	15%	22%	15%	23%
(E07) Prov. Feedback	924	**40%**	11%	13%	14%	22%
(E06) Sig. Non-Under.	108	**53%**	14%	14%	12%	7%
(E05) Sig. Under.	45	20%	13%	22%	13%	**31%**
(E08) Provide Assess.	18	**72%**	0%	22%	6%	0%
(E02) Test prior know.	17	**59%**	18%	18%	0%	6%
(E01) Test Underst.	17	**53%**	24%	6%	12%	6%
(E10) Other	16	31%	19%	31%	19%	0%
(E09) Extra. Info.	16	**62%**	25%	6%	0%	6%

Dialogue Acts	Freq.	Score Distribution 1	2	3	4	5
(D09) Info. Statement	917	19%	12%	23%	18%	**28%**
(D03) Question	580	24%	13%	23%	15%	25%
(D08) Disagreement	450	**59%**	14%	18%	6%	2%
(D02) What/how Ques.	432	**30%**	17%	20%	15%	18%
(D06) Answer	322	31%	18%	11%	18%	22%
(D07) Agreement	271	17%	7%	18%	**22%**	**35%**
(D01) Check Question	145	32%	18%	21%	14%	15%
(D05) Disconfirm.	126	**39%**	16%	25%	11%	10%
(D04) Confirm.	116	28%	13%	16%	17%	25%
(D10) Other	98	37%	13%	15%	15%	19%

Table 2: The frequency of explanation moves (left) and dialogue acts (right) in our dataset broken into each of the explanation quality levels [1-5]. Highlighted in bold values that distinguish the presence of these moves in high quality dialogues compared to low quality ones.

	Dialogue Act Flow	Freq.	Score Distribution 1	2	3	4	5
1	Ask, Inform, Ask, Inform, Ask, Inform	14	7.00%	0%	7.00%	**36.00%**	**50.00%**
2	Ask, Inform, Ask, Inform, Ask, Inform, Ask, Inform	6	0%	**33.00%**	**67.00%**	0%	0%
3	Ask, Inform, Ask, Inform, Ask, Inform, Agree	5	20.00%	20.00%	0%	**40.00%**	20.00%
4	Ask, Inform, Inform, Inform, Disagree, Inform, Disagree	2	**100.00%**	0%	0%	0%	0%
5	Ask, Inform, Agree, Inform, Answer, Inform, Answer	2	0%	0%	0%	**100.00%**	0%

Table 3: The most frequent dialogue act flows in our dataset broken into their frequency in each of the explanation quality levels [1-5]. Highlighted in bold values that distinguish the presence of these flows in high quality dialogues compared to low quality ones.

ing prior knowledge, *providing assessment*, and *testing understanding* appear most in low-quality dialogues. This could be because explainers only use these moves after failing to provide a good explanation. However, as expected, moves like signaling understanding and non-understanding correlate with high and low-quality dialogue, respectively. We further look at different turn-label sequences in our dialogues and their frequencies concerning different quality levels. In terms of dialogue acts, as shown in Table 3, successful dialogues are those of three rounds of asking questions and providing informing statements (flows #1 and #3) and ending with an agreeing statement, while longer interactions indicate lower quality, especially if ended with disagreeing statement (flows #2 and #4).

5. Automatic Assessment of Explanation Dialogue Quality

This section presents our study of the automatic assessment of explanation dialogue quality. We investigate whether augmenting the input of language models with interaction flow (encoded via special tokens) can boost their effectiveness.

5.1. Experiment Setup

Task Definition Give an explanation dialogue of n turns between an explainer and explainee, $d = [\tau_1, \tau_2, \cdots, \tau_n]$, the task is to predict a score $S \in [1 \cdots 5]$ reflecting the dialogue quality (in our data, $n \geq 6$). Each turn τ_i is composed of a sequence of m tokens, $\tau_i = (w_1, w_2, \cdots w_m)$, and has a set of three labels; the explanation act e_i, the dialogue act d_i, and the topic label t_i.

Models and Baselines We evaluate two recent pre-trained language models that allow processing long sequences of texts, the *LongFormer* (Beltagy et al., 2020) and the *Hierarchical Attention Transformer (HAT)* (Nawrot et al., 2022). To model the interactive nature of dialogues as shown in Figure 2, we add prefix tokens for each turn τ_i, representing the speaker's role (*explainer* or *explainee*) and the turn's interactive role represented as e_i, d_i, or t_i or all together. We then concatenate all the turns' tokens as a single sequence representing the final input to the models. We compare the effectiveness of the two language models for every setting with and without the different turn labels. We also consider the average baseline that always predicts the average score from the training set.

#	Model	Training Data	Explanation Moves			Dialogue Acts			Topics		
			ELI-5	5-Levels	Overall	ELI-5	5-Levels	Overall	ELI-5	5-Levels	Overall
1	BERT	ELI-5	0.30	0.25	0.27	0.34	0.32	0.38	0.35	0.41	0.41
2		5-Levels	0.16	**0.39**	0.32	0.14	0.44	0.30	0.22	0.47	0.40
3		Both	0.29	0.38	**0.39**	0.37	0.46	0.47	0.35	0.48	0.46
4	BERT-Seq	ELI-5	0.33	0.21	0.23	0.35	0.30	0.36	0.33	0.36	0.37
5		5-Levels	0.16	0.38	0.31	0.13	0.43	0.30	0.32	0.48	0.44
6		Both	0.36	0.37	0.37	0.37	0.46	0.47	0.35	0.49	0.47
7	RoBERTa	ELI-5	0.35	0.21	0.26	0.39	0.28	0.39	0.38	0.40	0.42
8		5-Levels	0.18	**0.39**	0.33	0.16	0.44	0.32	0.29	**0.54**	0.44
9		Both	0.35	0.35	**0.39**	0.39	**0.48**	0.48	**0.40**	0.53	**0.50**
10	RoBERTa-Seq	ELI-5	**0.39**	0.20	0.24	0.38	0.27	0.36	0.34	0.31	0.36
11		5-Levels	0.17	0.34	0.27	0.16	0.43	0.31	0.29	0.53	0.42
12		Both	0.34	0.38	0.38	**0.40**	0.47	**0.49**	0.35	**0.54**	0.49

Table 4: The macro F_1-score of the four evaluated models on the turn-level prediction of explanation moves, dialogue acts, and topics in 5-fold cross validation. The best score overall for each dataset is highlighted in **bold**.

Measures For evaluation, we compute the root mean squared error and the mean absolute error.

Predicting Turn Labels In practice, turn labels are not automatically available at inference time but must be predicted. Therefore, we also study the task of turn label prediction and test the performance of the trained quality models when the input contains ground-truth labels versus predicted labels. Since predicting turn labels is not the main focus of this paper, we retrain models available in previous work and focus on studying the generalization of these models across domains.

In particular, we experiment with the models *BERT* (Devlin et al., 2019) and *BERT-seq* of Wachsmuth and Alshomary (2022). The former simply uses BERT to predict the label from the text of a single turn, while the latter utilizes a CRF layer to model dependencies between the turn labels. Moreover, we also tested RoBERTa (Liu et al., 2019) as a backbone model, resulting in another two models, *RoBERTa* and *RoBERTa-seq*. We perform 5-fold cross-validation for each of the four models on each of the three data sources: *ELI-5*, *5-Levels*, and *overall*. This results in 12 models that we report their results in terms of average F1-score across all labels of the respective task.

5.2. Results

Predicting Turn Labels Table 4 presents the results of predicting explanation moves, dialogue acts, and topics, broken into the corresponding performance on the two datasets and overall. In all cases, models trained on both datasets performed best (**Overall** column), indicating the benefit of collecting heterogeneous datasets that cover multiple domains for the task. As for domain generalization across the two datasets (training on one dataset and evaluating on the other), BERT model generalized best from the ELI-5 dialogues to the 5-Levels dataset in all cases. For example, when predicting dialogue acts, among all models trained on ELI-5 Dialogues and evaluated on the 5-Levels dataset (#1, #4, #7, and #10 rows), it achieves the best F1-score of 0.32. Moreover, we can also notice that all models were better in generalizing from the ELI-5 Dialogues dataset to the 5-Levels dataset compared to the other way around (comparing #1, #4, #7, and #10 rows to #2, #5, #8, and #11 respectively). These results are not comparable to the results of Wachsmuth and Alshomary (2022) because we perform five-fold validation in our experiments compared to their 13-fold validation.

On the *ELI-5 Dialogues* dataset, *Roberta-seq* achieved the best results in predicting the explanation moves (when trained on the same dataset) and dialogue acts (when trained on both datasets), resulting in an F1-score of 0.39 and 0.40 respectively. In predicting topic relation, *RoBERTa* achieved the highest F_1-score of 0.4 when trained on both datasets. Therefore, we use these models to predict the turn labels in the following experiment.

Predicting Explanation Quality To obtain reliable results for each of the evaluated models (LongFormer and HatFormer with different augmented inputs), we performed 10-fold cross-validation on the training split and evaluated an ensemble of the ten trained models on the test split. We started fine-tuning from the *allenai/longformer-base-4096* and *kiddothe2b/adhoc-hierarchical-transformer-base-4096* checkpoints and selected the best checkpoint after training for 20 epochs.

Dialogue #1 **Rating:** 4

Explainee: Why are there not many "flamboyant" heterosexual males?

Request Explanation *Other Question* *Main Topic*

Explainer: I think a lot of the flamboyance is actually an act, albeit an unintentional one. It's a lot about fitting in with the culture. 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.

Provide Explanation *Informing Statement* *Main Topic*

Explainee: Thank you for not attacking my question and seeing it for the curiosity it is. I do believe culture and fitting in does play a large role here. But I haven't run into any flamboyant heterosexual males.

Provide Feedback *Agreeing Statement* *Main Topic*

Explainer: I guess we'd have to look at straight males that were raised by really flamboyant parents and see how they turned out.

Provide Feedback *Agreeing Statement* *Main Topic*

Explainee: I dont know if that would be considered cruel and unusual if done purposefully. But undoubtedly there should be 2 flamboyant men that could care for a child better than at least some heterosexual couples.

Provide Explanation *Informing Statement* *Main Topic*

Explainer: Yea we'll have to do these experiments underground

Provide Feedback *Agreeing Statement* *Main Topic*

Dialogue #2 **Rating:** 2

Explainee: how we extract meaning from the language we read? Do we link words to pictures in our mind?

Request Explanation *How Question* *Main Topic*

Explainer: Your brain stores knowledge more as abstract concepts. The word links to that concept, and the image links to that concept. See a cat and your brain identifies it as a cat then gives you the thought "that's a cat". See the word cat and your brain identifies the meaning of a cat and then gives you the image of a cat in your imagination. These concepts are stored without the use of either word or image, but are linked to the separate storages of the words and images, so those systems usually fire together. This is also how you can know what you want to say but not quite recall the word for it - your brain has accessed the abstract concept, but has misplaced the link between it and the word for it.

Provide Explanation *Informing Statement* *Main Topic*

Explainee: surely the image is the concept? I am sure for certain, attributes of an object are linked to the image rather than its name. Like you can know what to do with a pair of scissors even if you don't know their name in english.

Provide Explanation *Confirming Answer* *Subtopic*

Explainer: Nope. In fact, there are people who have literally no ability to form a mental image at all, and yet who still have normal ability to understand what things mean.

Provide Explanation *Disagreeing Statement* *Suptopic*

Explainee: how so, please explain?

Request Explanation *How Question* *Other*

Explainer: It's not yet known exactly what causes this, we only know that people like this exist.

Provide Feedback *Disconfirming Answer* *Other*

Table 5: Two example dialogues from our dataset that were rated with a high score of 4 (#1) and a low score of 2 (#2) by the annotators.

Table 6 shows the quality assessment results for the baseline and the two models with different augmented inputs, given either ground truth or predicted turn labels. Compared to the average baseline, the *plain* models reduce the RMSE by 0.18 and 0.26, respectively (1.60 as opposed to 1.42 and 1.34), indicating the applicability of modeling this task automatically. Encoding ground-truth interaction flows in terms of dialogue acts resulted in a further reduction of 0.13 and 0.03 RMSE for the HatFormer and Longformer, respectively. When evaluating the models on predicted turn labels, the turn label prediction errors propagate to the effectiveness in predicting quality scores. Nevertheless, we are still able to maintain error reduction better than the baseline. In the case of the HatFormer, the lowest RMSE and MAE are 1.28 and 1.05, resulting from encoding all turn labels into the input. For the LongFormer, encoding only the dialogue act into the input maintains the best results. In practice, Using the HatFormer with all turn labels encoded in the input (HatFormer w/ALL) gives the best error reduction on the task with an RMSE of 1.28 and MAE of 1.05.

Early Prediction As a follow-up analysis, we study the automatic quality prediction at an early stage (first few turns). From a practical viewpoint, this could be used to give the explainer insights into how the dialogue might end up so they can

Approach	Ground Truth RMSE	Ground Truth MAE	Predictions RMSE	Predictions MAE
Average Baseline	1.60	1.42	1.60	1.42
HatFormer	1.42	1.17	1.42	1.17
w/ Dialogue Act	**1.29**	***1.05**	1.31	1.09
w/ Expl. Move	1.41	1.21	1.43	1.22
w/ Topic	1.41	1.20	1.41	1.20
w/ ALL	1.30	**1.05**	**1.28**	**1.05**
LongFormer	1.34	1.13	1.34	1.13
w/ Dialogue Act	**1.31**	**1.05**	**1.32**	**1.06**
w/ Expl. Move	**1.31**	**1.05**	**1.32**	1.09
w/ Topic	1.35	1.15	1.34	1.14
w/ ALL	1.32	1.08	1.34	1.10

Table 6: Explanation quality results: Root mean squared error (RMSE) and mean absolute error (MAE) of the two models with different augmented inputs, and of the average baseline for two scenarios: In *ground truth*, quality is prediced based on ground-truth turn labels. In predictions, the turn labels are predicted with the developed methods. Highlighted in **bold** are best scores. * indicates statistical significance with 95% confidence.

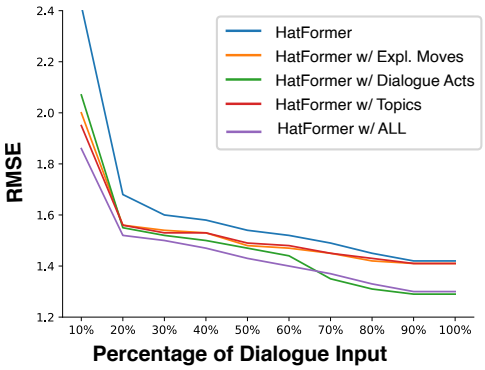

Figure 4: The root mean squared error (RSME) of all models for early predictions of explanation quality, that is, when the input to the models is only a defined initial percentage (10%, . . ., 100%) of the full explanation dialogue.

adjust their strategies accordingly and successfully deliver an explanation. Therefore, we evaluate the effectiveness of the HatFormer variations in terms of RMSE, when the input is only the first k turns that form 10%, 20%, $\cdots$, 100% of the entire dialogue (rounded). Figure 4 illustrates the results. Expectedly, the RMSE of all the models decreases with increasing the dialogue portion taken as an input. Encoding turn labels into the HatFormer results in a reduction of the RMSE for all tested dialogue proportions. Overall, we observe that at the 70% mark, the HatFormer w/ Dialogue Acts and HatFormer w/ ALL already achieved good results.

Qualitative Analysis Table 5 show two example dialogues. Dialogue #1 was rated four as a high-quality explanation dialogue, while Dialogue #2 was rated with a score of two, reflecting low-quality dialogue. We can observe that a successful explanation dialogue contains a pattern of requesting and providing explanations and feedback with agreeing statements. However, a low-quality dialogue consists of a sequence of providing explanations without feedback along with disagreeing statements. Moreover, we predict the quality scores of these dialogues using the *HatFormer* baseline and *HatFormer w/ Dialogue Act*. For Dialogue #1, the baseline predicted a score of 0.57, while our model generated a score of 4.02, which is closer to the ground truth. We think focusing only on the dialogue content is insufficient to infer its quality. Explanation-level interactions can give the model better signals that help predict accurate scores.

6. Conclusion

We studied real-life explanation dialogues and how to assess their success. To this end, we constructed a dataset of real-life explanation dialogues from the *Explain Like I am Five* Subreddit. We annotated it according to the explanation taxonomy of Wachsmuth and Alshomary (2022) and rated the quality of these dialogues in terms of the explainee's understanding. Our analysis provides insights into the difference between these dialogues and expert explanation dialogues. We then assessed the performance of pre-trained language models in predicting the quality of explanation dialogues and found that encoding specific interaction flows into their input boosts effectiveness.

In quantifying the explanation dialogue quality, we relied on the annotators' intuition of guessing the explainee's understanding. Although this might be a good proxy, it does not sometimes reflect the real understanding of the explainee. Moreover, in encoding dialogue interactions, better methods could be explored in the future, such as LSTMs, Transformer-based models, or via prompting large language models LLMs.

Acknowledgments

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/1 2021 – 438445824. We thank the anonymous freelancers on Upwork for their annotations of our corpus.

7. Ethics

While constructing our corpus, we did not crawl any personal information that revealed the authors' identity. We ensured that the workers who annotated our corpus got paid more than the minimum wage in the U.S., namely 250$ for a workload of 20 hours.

As for the ethical implications of our work, we would like to emphasize that our research only aims to give insights into the nature of interactions in explanation dialogues. These insights and the constructed corpus can enable research in the field of explainable AI to design systems to explain model decisions to humans optimally.

8. Bibliographical References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sarah Bourse and Patrick Saint-Dizier. 2012. A repository of rules and lexical resources for discourse structure analysis: the case of explanation structures. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2778–2785, Istanbul, Turkey. European Languages Resources Association (ELRA).

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).

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022. An exploration of hierarchical attention transformers for efficient long document classification.

Shruthi Chari, Oshani Seneviratne, Daniel M Gruen, Morgan A Foreman, Amar K Das, and Deborah L McGuinness. 2020. Explanation ontology: a model of explanations for user-centered ai. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II*, pages 228–243. Springer.

Roberto Confalonieri, Tarek R. Besold, Tillman Weyde, Kathleen Creel, Tania Lombrozo, Shane T. Mueller, and Patrick Shafto. 2019. What makes a good explanation? Cognitive dimensions of explaining intelligent machines. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pages 25–26.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu, and Seokhwan Kim. 2018. PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210, Montréal, Canada. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Josefine Finke, Ilona Horwath, Tobias Matzner, and Christian Schulz. 2022. (de)coding social practice in the field of xai: Towards a co-constructive framework of explanations and understanding between lay users and algorithmic systems. In *Artificial Intelligence in HCI*, pages 149–160, Cham. Springer International Publishing.

Lionel Fontan and Patrick Saint-Dizier. 2008. Analyzing the explanation structure of procedural texts: Dealing with advice and warnings. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 115–127. College Publications.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Alan Garfinkel. 2009. *Forms of Explanation: Rethinking the Questions in Social Theory*, revised edition. Yale University Press, New Haven & London, New Haven; London.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. ArXiv: 1806.00069.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940. Association for Computational Linguistics.

Nicholas Halliwell, Fabien Gandon, Freddy Lecue, and Serena Villata. 2022. The Need for Empirical Evaluation of Explanation Quality. In *AAAI 2022 - Workshop on Explainable Agency in Artificial Intelligence*, Vancouver, Canada.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia. Association for Computational Linguistics.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th*

International Conference on Computational Linguistics: Technical Papers, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

Pamela W. Jordan, Maxim Makatchev, and Umarani Pappuswamy. 2006. Understanding complex natural language explanations in tutorial applications. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 17–24, New York City, New York. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1033?1041, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada. Association for Computational Linguistics.

Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Łukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. 2022. Hierarchical transformers are more efficient language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1559–1571.

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.

Xuelin Situ, Ingrid Zukerman, Cecile Paris, Sameen Maruf, and Gholamreza Haffari. 2021. Learning to explain: Generating stable explanations fast. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5340–5355, Online. Association for Computational Linguistics.

Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. 2018. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3350–3359, Brussels, Belgium. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

John M. Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Keith Vander Linden. 1992. The expression of local rhetorical relations in instructional text. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 318–320.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

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.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Alberdingk Tim Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Henning Wachsmuth and Benno Stein. 2017. A universal model for discourse-level argumentation analysis. *Special Section of the ACM Transactions on Internet Technology: Argumentation in Social Media*, 17(3):28:1–28:24.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint arXiv:2102.12060*.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Ziqi Zhang, Philip Webster, Victoria Uren, Andrea Varga, and Fabio Ciravegna. 2012. Automatically extracting procedural knowledge from instructional texts using natural language processing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 520–527, Istanbul, Turkey. European Language Resources Association (ELRA).

A. Annotation Analysis

Table 9 shows the distribution of topic relation over different quality scores. In terms of explanation moves sequences, as shown in Table 7, prominent flows that occur in high-quality explanation dialogues are those that contain two rounds of requesting and providing explanations followed by some feedback (flows #2 and #3). Not surprisingly, dialogues that do barely provide explanations (flow #7 and #8) are ranked to be of lower quality. Regarding topic sequences, Table 8 shows that diverging from the main topic mostly correlate with low-quality dialogues.

	Explanation Move Flow	Freq.	Score Distribution 1	2	3	4	5
1	Req., Explain, Req., Explain, Req., Explain,	17	6%	0%	12%	**41%**	**41%**
2	Req., Explain, Req., Explain, Feedback, Feedback	9	11%	0%	33%	0%	**56%**
3	Req., Explain, Req., Explain, Req., Explain, Req., Explain,	8	0%	**25%**	**62%**	0%	12%
4	Req., Explain, Feedback, Feedback, Feedback, Feedback	5	**20%**	**20%**	20%	0%	40%
5	Req., Explain, Req., Explain, Feedback, Explain, Feedback	4	0%	**50%**	0%	25%	25%
6	Req., Feedback, Feedback, Feedback, Feedback, Feedback	3	**67%**	0%	0%	0%	33%

Table 7: The most frequent explanation move flows in our dataset broken into their frequency in each of the explanation quality levels [1-5]. Highlighted in bold values that distinguish the presence of these flows in high quality dialogues compared to low quality ones.

	Topic Relation Flow	Freq.	Score Distribution 1	2	3	4	5
1	Main, Main, Main, Main, Main, Main	62	11.00%	13.00%	13.00%	**18.00%**	**45.00%**
2	Main, Main, Main, Main, Main, Main, Main	12	42.00%	8.00%	8.00%	42.00%	0%
3	Main, Other, Other, Other, Other, Other, Other	9	**56.00%**	11.00%	22.00%	11.00%	0%
4	Main, Other, Other, Other, Other, Other	7	**100.00%**	0%	0%	0%	0%
5	Main, Suptopic, Suptopic, Suptopic, Suptopic, Suptopic	7	14.00%	14.00%	29.00%	29.00%	14.00%

Table 8: The most frequent topic relation flows in our dataset broken into their frequency in each of the explanation quality levels [1-5]. Highlighted in bold values that distinguish the presence of these flows in high quality dialogues compared to low quality ones.

Topic Relation	Freq.	Score Distribution 1	2	3	4	5
(T01) Main Topic	1747	26%	14%	18%	16%	**26%**
(T04) No topic	623	**47%**	11%	12%	11%	19%
(T02) Subtopic	611	25%	18%	31%	16%	10%
(T03) Related Topic	476	25%	12%	24%	19%	20%

Table 9: The frequency of topic relation labels in our dataset broken into each of the explanation quality levels [1-5]. Highlighted in bold values that distinguish the presence of these moves in high quality dialogues compared to low quality ones.