Evaluating Intention Understanding Capability of Large Language Models in Persuasive Dialogues

Anonymous ACL submission

Abstract

This study aims to verify whether Large Language Models (LLMs) understand intentions from utterances in dialogues. Despite LLMs being already applied in various real-world scenarios such as dialogue systems, no representative dialogue dataset exists to verify to what extent they understand speakers' intentions. We hypothesize that LLMs understand speakers' intentions during conversations. To verify this, we construct a dataset from persuasive dialogues featuring multiple-choice questions predicting the speaker's intention within conversational contexts. When engaging in a persuasive conversation smoothly, especially when making a request or reply inconvenient for others, it is crucial to consider their perspectives through their speech. This feature makes the persuasive dialogue suitable for the dataset of measuring intention understanding ability. We incorporate the concept of 'face acts,' which categorize how speech influences psychological states, to explore how utterances affect mental states. This approach aims to measure intention understanding capability by focusing on crucial intentions. We employ the largest available LLMs and measure how well they understand intention in persuasive dialogues. The experimental results suggest that LLMs already possess high intention understanding ability.

1 Introduction

Understanding the speaker's intention is crucial for maintaining a smooth conversation. Suppose a situation where Alice asks Bob for a donation to a specific charity, and Bob responds with an evasive answer such as 'Well, you know....' In this situation, we can assume that Bob is unwilling to donate, but since refusing the donation is psychologically burdensome, he wants Alice to sense his hesitation. The speaker's intentions can be conveyed without saying them out loud, and they also vary depending on the context of the conversation. We engage in conversations while estimating the speaker's intentions unconsciously, and this ability is essential for facilitating natural communication.

In recent years, there has been remarkable progress in developing LLMs such as ChatGPT[1] or GPT-4 (OpenAI, 2023). By leveraging the capability to engage in human-like communication using natural language, research and development of dialogue systems incorporating LLMs are actively going on. Considering LLMs are already applied in various real-world scenarios, we hypothesize that they understand speakers' intentions well during conversations. There are some datasets, such as GLUE (Wang et al., 2021), to assess whether LLMs understand natural language like humans. Although LLMs perform well in most existing NLP tasks and are known to have high linguistic knowledge, few works focus on exploring their ability to understand speakers' intentions in conversations.

This study creates a dataset to measure LLMs' intention understanding capability. This dataset consists of multiple-choice questions that ask LLMs to understand the speakers' intentions in persuasive conversations. Unlike prior studies focused on single-turn utterances, understanding intentions within a conversation requires considering the context of previous utterances. Moreover, in persuasive conversations, making requests or replies that are inconvenient for others or even hurt others' feelings is inevitable. Therefore, speakers should consider others' feelings or perspectives more carefully through their utterances than in daily conversation. These features are suitable for measuring intention understanding ability in multi-turn dialogues.

In the dataset creation, we employ the concept of *face* (Goffman, 1967), a desire related to human relationships in social life. By focusing on specific utterances that influence face, we can measure the ability to understand the intentions of crucial speech that affect the interlocutor's emotions.

[1] https://chat.openai.com

Moreover, grouping similar types of intentions by applying face enhances the clarity of analysis, leading to improved insights. After creating the dataset, we verified whether LLMs can understand intentions from utterances. We analyzed several LLMs' intention understanding capabilities and identified the types of intentions that are particularly challenging for them. The results reveal that the current state-of-the-art models achieve high intention understanding ability comparable to humans.

This research makes the following two contributions. First, we constructed a dataset for measuring intention understanding capabilities from persuasion dialogues. This dataset follows the format of comprehension problems from previous studies. Second, we evaluated how well state-of-the-art LLMs such as GPT-4 and ChatGPT understand the intention of utterances in dialogues. We found that LLMs have already acquired high intention understanding capabilities, and we provide insights into mistakes made by LLMs and into intentions that are challenging to comprehend. The dataset developed in this study will be released to the public shortly.

2 Background

This section first explains *face* and *face acts* and dialogue data utilized in our research. After that, we discuss previous studies on dialogue comprehension and intention understanding.

2.1 Face and Face Act

2.1.1 Face

Face is our primary need related to human relationships with others in social life. This concept was introduced by Goffman (1967). In Brown and Levinson's politeness theory, face can be divided into two categories: *positive face* and *negative face* (Brown and Levinson, 1978).

A positive face is a desire to be recognized, admired, and liked by others. On the other hand, a negative face is a desire not to let others invade one's freedom or domain. Brown and Levinson established politeness theory by applying the concept of face, and systematized the verbal behaviors that influence faces as politeness strategies.

2.1.2 Face Act

Face acts are speech acts that affect either oneself or others' faces, and can be divided into two types. *Face Threatening Act* (FTA) is a speech act that attacks either positive or negative faces. On the other hand, *Face Saving Act* (FSA) is a speech act that saves either positive or negative faces.

According to the politeness theory, people tend to avoid attacking faces as much as possible to manage relationships. Also, even when they must attack faces, they will do it in a way that reduces the risk of attacking faces by employing politeness strategies such as implying their needs or apologizing for what they have requested.

Dutt et al. (2020) incorporates the concept of face acts for analyzing dialogues in persuasive situations, where maintaining good relationships with others is particularly important. They identified face acts as factors influencing the success of persuasion. They developed a machine learning model to track the conversation's dynamics, employing face acts and conversation histories. They divided face acts into eight categories based on the following three criteria.

- whether it is directed toward the *speaker* or the *hearer* (**s/h**)
- whether it is directed toward a *positive* or *negative* face (**pos/neg**)
- whether the face is *saved* or *attacked* (**+/-**)

Suppose a persuasive situation where there are two people. One of them who makes the other mind change is called *persuader* (ER), and the other side is called *persuadee* (EE). When ER requests EE to do something, the utterance is a face act categorized as **hneg-**. That is because the speaker is taking away the hearer's freedom. On the other hand, when ER shows the validity of their argument, the utterance has face act categorized as **spos+**, as the speaker is defending their positive face.

2.2 Dataset Annotated with Face Act

The representative English dialogue dataset annotated with face acts is created by Dutt et al. (2020). This study annotated face acts in persuasion dialogues about the donation to a charity named *Save the Children* (STC)[2]. In the whole conversation, there are two people called *persuader* (**ER**) and *persuadee* (**EE**), and ER persuades EE to donate to a charitable organization. Table 1 is a part of a conversation in the dataset. Those utterances categorized as **other** are greetings, fillers, and utterances unrelated to the main topic of the conversation.

[2] https://www.savethechildren.org

Table 1: An example of a part of an annotated conversation with face act labels from Dutt et al. (2020). In this two people's conversations, they are given roles *persuader* (**ER**) and *persuadee* (**EE**). ER persuades EE to donate to a charitable organization.

Speaker	Utterance	Face act
ER	Would you be interested today in making a donation to a charity?	**hneg-**
EE	Which charity would that be?	**other**
ER	The charity we're taking donations for is save the children!	**other**
EE	I've seen a lot of commercials about them, but never did a lot of research about them.	**hpos+**
ER	They are actually really great.	**spos+**

The dialogue was initially collected in Wang et al. (2019). Only one face act is attached to each utterance in Dutt et al. (2020). Although it might be possible that one utterance has two or more face acts, the previous study reported that those utterances comprise only 2% of the dataset. Therefore, they randomly selected only one face act out of possible face acts, and regarded it as a gold label.

2.3 Intention Understanding

Understanding dialogue requires various types of reasoning abilities. Intention understanding ability, which refers to predicting what a speaker aims to achieve from their utterance, is one of these abilities needed. Both task types for measuring reasoning ability and methods for evaluation differ between studies. For instance, MuTual (Cui et al., 2020) was created to analyze the dialogue understanding abilities of machine learning models from multiple perspectives, such as attitude reasoning, algebraic reasoning, and intention prediction. Questions in this dataset take the form of the next utterance prediction. In other words, we need to understand the context from dialogue history and select one logically coherent option from four choices suitable for the following utterance.

Some previous studies focus on measuring intention understanding abilities (Purohit et al., 2015; Larson et al., 2019). One of the representative datasets is SNIPS (Coucke et al., 2018), and LLMs such as GPT-2 are reported to achieve comparably high performance in those task (Winata et al., 2021). Unlike those previous studies, which deal with intentions from single utterances, we create a dataset that assesses the ability to understand a speaker's intention within multi-turn dialogue.

Script

ER: Please donate $1.
EE: **Sorry I can't.**

Options

A: ER insists that they are proud of themselves.
B: EE either knows nothing about STC or is not interested in STC.
C: EE acknowledges the efforts of STC.
D: EE apologizes for not donating.

Figure 1: A dataset instance we create comprises conversation history and four candidate descriptions of intentions for the last utterance.

3 Data

As mentioned in the previous section, prior studies on intention understanding mostly did not apply multi-turn dialogue data. There is a possible approach to measure intention understanding capability utilizing the persuasive dialogue dataset created in Dutt et al. (2020) and directly predicting face acts from utterances. However, considering that face acts are abstract intentions and are not well-known concepts, they are non-intuitive for humans to handle. Also, they are likely not sufficiently acquired by LLMs in in-context learning. Unlike prior research that predicted face acts using supervised learning-based neural networks, we investigate whether LLMs understand intention. Therefore, not employing face act prediction task straightforwardly but modifying the task settings so that we can evaluate LLMs' intention understanding capability through zero-shot or few-shot scenarios is necessary.

We modify persuasive dialogue data[3] in Dutt et al. (2020) and create a dataset for evaluating intention understanding capability. Instead of directly predicting face acts, this study transforms face acts into descriptions of intentions written in natural language to make the task more comprehensible. Each entry in our dataset is represented in Figure 1. The input of this task consists of conversational history and four intention descriptions for the last utterance in the conversation. The output is one description out of four candidate intentions. This format is a reading comprehension style inspired by several previous dialogue reasoning studies (Cui et al., 2020; Huang et al., 2019) and frequently employed for evaluating LLMs' reasoning ability.

This study aims to create evaluation data for measuring the capability of intention understanding. Therefore, we first partitioned the dataset into

[3]This data is licensed under the MIT license.

training, development, and test data in an 8:1:1 ratio and only utilized the test subset.

In this section, we describe how we developed the evaluation dataset. First, we outline how we defined intention descriptions that will be annotated into utterances. Then, we detail how we annotated intention descriptions for each utterance through crowdsourcing. Lastly, we clarify how we selected three distractors to create four options.

3.1 Preparation of Intention Description

Dutt et al. (2020) presented several descriptions of intentions found in persuasive situations with corresponding face acts. We adapted and expanded upon these descriptions, which were then annotated to correspond with specific utterances. Specifically, we devised new descriptions to encompass all utterances in the development data and refined broader intentions into more specific versions. We curated 42 utterances listed in Table 2.

3.2 Intention Annotation

We sampled 30 dialogues for test data from the persuasion dialogue dataset. We annotate descriptions of intentions to utterances. Those utterances are annotated face act labels by Dutt et al. (2020), as they can affect the interlocutor's emotion more than utterances that are not regarded as face act. We hired crowdworkers residing in the US to carry out the intention annotation process through Amazon Mechanical Turk (AMT). We ensured fair compensation, offering all participating workers an average hourly wage of $12. We conduct three rounds of pilot tests to refine instructions and select annotators who provide high-quality annotation. Finalized instructions for the annotation process can be found in Appendix A. During annotation, workers carefully read through entire conversations and assign intentions to specific utterances from a set of candidate descriptions. Workers are presented with descriptions categorized under the same face act as the utterance. For example, if workers annotate a description of the EE's utterance whose face act is categorized as 'hpos-,' they annotate either 'EE doubts or criticizes STC or ER.' or 'EE either knows nothing about STC or is not interested in STC.' as the intention of the utterance. For each instance, three workers conducted annotations, resulting in three descriptions annotated for each utterance. We took a majority vote for three descriptions and annotated gold labels if more than one worker annotated the same intention. We let workers annotate 691 utterances in total, and among them, 620 utterances had agreement from at least two out of three individuals' opinions. In the following process, we create a problem of intention classification for these 620 utterances. To assess the level of agreement among annotators, we calculated Krippendorff's alpha (Hayes and Krippendorff, 2007). It results in a value of 0.406 and indicates a moderate level of agreement. See Appendix C for more details about the annotator agreement.

3.3 Question Creation

After obtaining 620 utterances annotated with intentions, we concatenated consecutive utterances annotated with the same intentions. There are some utterances where the intentions become apparent only after hearing the subsequent utterances. Therefore, this process is essential to prevent creating questions that need to predict intentions from incomplete utterances. As a result, we obtained 553 utterances annotated with intentions. We create multi-choice questions from those utterances. We randomly selected three distractors from the predefined description pool for each utterance. Refer to Appendix D for rules for the distractor selection process. Table 3 shows the data statistics.

4 Experiment

We evaluate how well existing state-of-the-art LLMs understand intentions from utterances in persuasive dialogues. Among LLMs released by OpenAI, we employed GPT-4 and ChatGPT. Other models are Llama 2 (Touvron et al., 2023) from Meta and Vicuna (Zheng et al., 2023) which is based on Llama 2 and trained on conversational data[4] between ChatGPT and human.

The provided prompts to LLMs include information for understanding the intentions of the utterance: conversational situation and task explanation, conversational script, and a four-optional question. We designed the prompt according to the Zero-shot Chain-of-Thought style (Kojima et al., 2022), dividing the answering process between the *reason explanation* and *option selection* phases. See Appendix E for details of the prompt we created. In the reason explanation phase, LLMs explain whether the intention is explicitly stated or implied and what the interpreted intention is. In the option selection phase, LLMs judge which option is the

[4]Conversations were collected on ShareGPT (https://github.com/domeccleston/sharegpt).

Table 2: All 42 descriptions we defined.

Face Act	Persuader (ER)	Persuadee (EE)
spos+	ER praises or promotes the good deeds of STC. ER states that STC is a reputable and trustworthy organization. ER states that STC provides information on donations or other related matters, implying that STC engages in beneficial activities for society. ER shows their involvement for STC, such that they are going to donate to STC or have done so in the past. ER expresses their preference for charities or the targets they want to help. ER claims that they want to do something good, such as helping children. ER claims that they have donated to charities other than STC or participated in their activities. ER insists that they are proud of themselves.	EE presents their knowledge about charities to ER. EE insists that they are proud of themselves. EE claims that they have donated to charities other than STC or participated in their activities. EE expresses their preference for charities or the targets they want to help. EE claims that they want to do something good, such as helping children.
spos-		EE apologizes for not making a donation or for making only a small one.
hpos+	ER appreciates or praises EE's generosity. ER empathizes or agrees with EE. ER encourages EE to do good deeds, other than donating to STC. ER is interested in the organization mentioned by EE and plans to research it later. ER compliments EE for their virtues, efforts, likes or desires. ER motivates EE to donate to STC, such as by explaining the essential role their donation plays in helping children or highlighting the suffering children endure due to war, poverty, and other hardships.	EE shows willingness to donate or to discuss the charity. EE empathizes or agrees with ER. EE acknowledges the efforts of STC. EE states that they know about STC by name, but they are not so familiar with the organization. EE appreciates or praises ER's generosity. EE is planning to browse the website recommended by ER.
hpos-	ER criticizes EE.	EE doubts or criticizes STC or ER. EE either knows nothing about STC or is not interested in STC.
sneg+		EE is either hesitant or unwilling to donate to STC. EE refuses to donate to STC or increase the donation amount without even giving a reason. EE cites reason for not donating at all or not donating more.
hneg+	ER makes donating easy and simple, reducing any inconvenience for EE. ER apologizes for inconvenience or intrusion. ER tries to minimize the financial burden on EE.	
hneg-	ER asks EE for donation. ER asks EE to donate more. ER asks EE for their time or permission to discuss charities.	EE asks ER for donation. EE asks ER questions about STC. ER asks or confirms the amount that EE is donating to STC. EE asks ER how ER themselves are involved in STC.

Table 3: Data Statictics.

# Questions	553
# Dialogues	30
# Avg. questions per dialogue	18.43
# Avg. turns per dialogue	30.8
# Avg. words per utterance	11.99
# Avg. Words per description	10.61

best according to the output in the reason explanation phase. Models can see whole utterances before the objective utterance. Due to memory constraints, we limit the history length to the past ten utterances when using Llama 2 and Vicuna.

To benchmark human performance, we hired three workers from AMT to solve the task. The final answer was determined by a majority vote among the workers' choices.

5 Results and Discussion

Table 4 shows how well the models understood intentions. Even the smallest model achieved an accuracy exceeding 50%, while GPT-4 surpassed 90%, demonstrating their capacity to solve questions in this dataset. Smaller models displayed a reasonable understanding of intentions, and as model size increased, accuracy rates consistently improved. However, they are struggled with understanding intentions whose face act are categorized as 'hpos-.' Notably, when understanding the inten-

Table 4: Whole result. Each cell represents the accuracy of ER's utterance, the accuracy of EE's utterance, and the accuracy of Both ER & EE's utterances. For human results, we collected responses from three workers and determined the chosen intent by majority vote. The bottom row represents the number of utterances in the test data according to speakers and face acts. See Appendix F for details about model versions and decode settings.

	spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**	Total
Human	.96/.79/.91	-/1.0/1.0	.95/.93/.94	.86/.81/.82	-/1.0/1.0	1.0/-/1.0	.98/.93/.96	.96/.90/.93
Vicuna-7B-v1.5	.48/.32/.43	-/0.0/0.0	.58/.61/.59	0.0/.19/.15	-/.78/.78	.53/-/.53	.64/.53/.59	.54/.53/.54
Llama 2-7B	.48/.42/.47	-/.50/.50	.52/.62/.57	.14/.42/.36	-/.78/.78	.29/-/.29	.54/.28/.44	.49/.53/.51
Vicuna-13B-v1.5	.66/.40/.58	-/.50/.50	.65/.72/.69	.29/.23/.24	-/.82/.82	.59/-/.59	.66/.56/.61	.64/.61/.63
Llama 2-13B	.53/.45/.50	-/.50/.50	.73/.74/.74	.14/.46/.39	-/.82/.82	.71/-/.71	.64/.37/.52	.64/.62/.63
Llama 2-70B	.66/.45/.60	-/1.0/1.0	.89/.81/.85	.29/.46/.42	-/.85/.85	.65/-/.65	.81/.63/.73	.78/.70/.74
ChatGPT	.94/.63/.85	-/1.0/1.0	.85/.89/.87	.57/.73/.70	-/1.0/1.0	.82/-/.82	.87/.84/.85	.88/.84/.86
GPT-4	.93/.74/.87	-/1.0/1.0	.93/.96/.94	.14/.62/.52	-/1.0/1.0	.94/-/.94	.94/.95/.95	.91/.90/.91
# Utterances	89/38/127	0/2/2	130/121/251	7/26/33	0/27/27	17/0/17	53/43/96	296/257/553

tion of ER's utterances labeled as hpos-, GPT-4 can correctly understand the intention in only 1 out of 7 questions. This suggests underlying issues that will be further addressed in the upcoming section. This section first observes the behaviors where smaller LLMs struggle during inference. Subsequently, we analyze utterances where LLMs, especially GPT-4, exhibit difficulties understanding intentions.

5.1 Behavior of Smaller LLMs

While GPT-4 understands intention well and is nearly indistinguishable from human capabilities, smaller models encountered difficulties in inference. This section compares ChatGPT and Llama 2-70B to GPT-4, both smaller yet acquired superior intention understanding capabilities. We divided problem types in which smaller models struggled into *intention-related* and *non-intention-related* problems. The intention-related problems are where a flawed interpretation of reasoning leads to the selection of incorrect answers. On the other hand, the non-intention-related problems outline errors unrelated to intention understanding, such as predicting the intention of different utterances other than the objective one or encountering logical inconsistencies in outputting answers.

5.1.1 Intention-related Problems

Both ChatGPT and Llama 2-70B struggle with problems that they carry out logically flawless inference, but the thought process is unusual. While GPT-4 guesses intentions within reasonable bounds, ChatGPT and Llama 2-70B occasionally overinterpretate intentions. For instance, in the example illustrated in Figure 2, GPT-4 interpreted that EE just mentions their donation habits, which aligns with humans' judgement. On the other hand, both ChatGPT and Llama 2-70B expanded the interpretation by inferring, "Since EE has already donated to the church, there is no intention to donate to STC." Considering EE has smoothly agreed to donate to STC in this conversation, the choices made by GPT-4 and the humans seem more appropriate, and no ulterior motives can be inferred.

5.1.2 Non-intention-related Problems

Llama 2-70B, besides overinterpreting intentions, faces issues like generation loops and predicting intentions of utterances different from the objective one. This behavior can also be seen when we employ much smaller Llama-like models. The cause of these issues could be the use of complex and lengthy prompts that were challenging for the smaller model to comprehend, resulting in a lack of understanding of the instructions in the prompt. Furthermore, smaller models suffered from a critical issue of logical inconsistencies within their responses. This problem might stem from their inferior capability in logically deriving answers in line with the instructions provided in the prompt, compared to larger-scale models.

Figure 3 provides an example of common errors observed in the output of Llama 2-70B. The model often chooses the last option as the correct answer without proper consideration after dismissing the first three options as inappropriate. While option D constitutes 25.8% of the correct answers overall, Llama 2-70B chooses it 32.1% of the time, indicating an unusually high frequency of selecting the last option. Problems like struggling to pick the most plausible option after examining all choices or having inconsistencies in reasoning during inference degrade the performance of smaller models.

5.2 About hpos-

LLMs are especially weak against interpreting utterances whose face acts are categorized as hpos-. Those utterances are in which ER condemns EE's

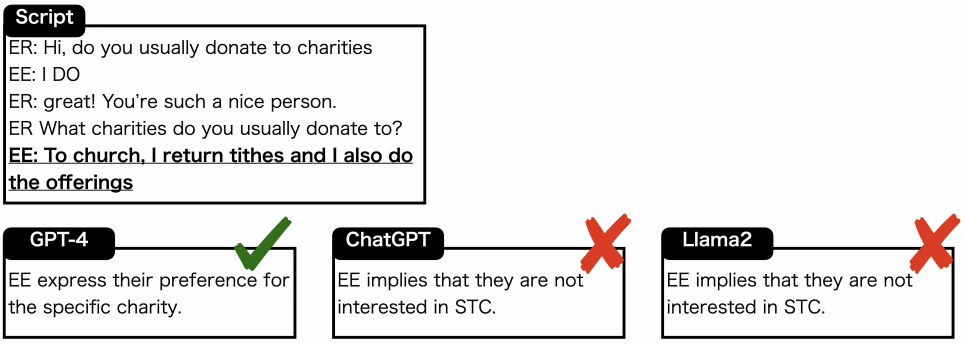


Figure 2: An example of intention-related problems. GPT-4 reasonably infers intentions, while ChatGPT and Llama 2 overread EE's intention.

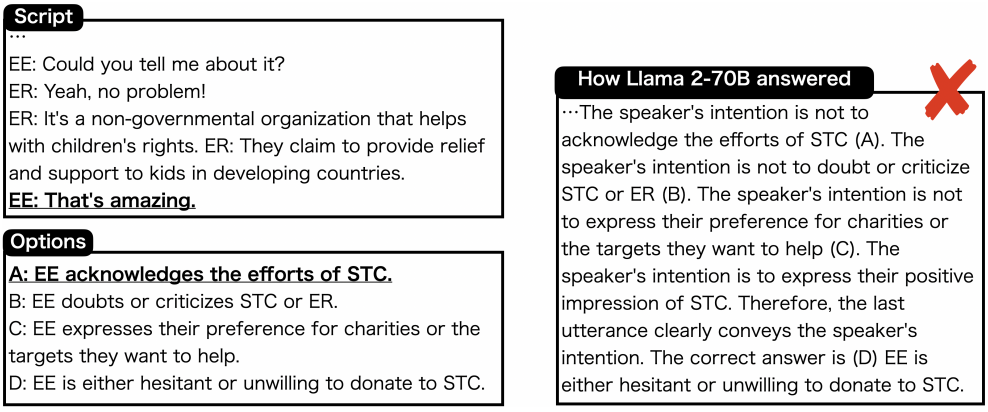


Figure 3: Examples of non-intention-related problems. Llama 2-70B simply dismissed all options among A to C and select the option D as a correct answer.

hesitation to donate, or EE expresses doubts about ER's credibility. GPT-4 made mistakes in inferring the intentions behind EE's utterance mostly due to flawed questions we mention in Section 6; hence, we primarily examine how GPT-4 interprets utterances in which ER criticizes EE.

5.2.1 Patterns in Our Dataset

Table 5 shows two prominent patterns in how ER criticizes EE. The first pattern is that ER questions EE's spending habits, suggesting redirecting wasteful spending towards STC. The second pattern is that ER mentions people who are experiencing financial hardship compared to EE and appeals to guilt by implying that the inaction of EE causes suffering for the impoverished. GPT-4 discerned that most of those utterances were not primarily critical but had other intentions, as outlined in Table 6.

5.2.2 Artificially Created Dataset

To examine to what extent utterances with the two characteristics mentioned in the preceding section are perceived as critical, we artificially create scenarios with those utterances. We devised a prompt to generate situations where EE hesitates to donate, ER criticizes EE's spending habits, or ER points out EE's carelessness for those who are unprivileged. See Appendix G for the prompt we employed to generate persuasive conversations. We generated 20 persuasive dialogues using GPT-4

Table 5: Examples of ER's critical utterances appeared in our datasets. There are two patterns in how ER criticizes EE. Firstly, ER questions EE about how they spend money. Secondly, ER mentions impoverished people and guilt-tripping EE's inaction.

Type	Utterance
Questioning EE's spending habits	(1) Think about how you were probably going to just waste the measly reward amount you were being offered for this HIT on junk food or coffee and think about what amazing things Save the Children would be able to do with that money. (2) How much money do you waste on candy or cookies every year?
Blaming EE's inaction	(1) Why do you think that? There are children dying in Syria who can benefit from the donation. (2) By not donating this tiny amount you're directly allowing children to sufer.

Table 6: List of intentions chosen by GPT-4 instead of inferring 'ER criticizes EE.' among the six errors made by GPT-4. 'No suitable option among the choices' refers to outputs where GPT-4 considered all options but found no suitable choice.

Intention	#
(1) No suitable option among the choices	2
(2) ER expresses their preference for charities or the targets they want to help.	2
(3) ER motivates EE to donate to STC, such as by explaining the essential role their donation plays in helping children or highlighting the suffering children endure due to war, poverty, and other hardships.	1
(4) ER asks or confirms the amount that EE is donating to STC.	1

Table 7: Differences of Intention interpretation between Human and GPT-4. 'H' and 'G' represents humans and GPT-4, respectively. These characters are combined with verbs corresponding to the selected descriptions. H-ask means humans choose the description 'ER asks or confirms the amount that EE is donating to STC.' Other descriptions are 'ER motivates EE to donate to STC, such as by explaining the essential role their donation plays in helping children or highlighting the suffering children endure due to war, poverty, and other hardships.' (motivate), and 'ER criticizes EE.' (criticize).

	H-ask	H-motivate	H-criticize
G-ask	0	1	0
G-motivate	1	84	1
G-criticize	0	0	3

and extracted 90 utterances that can be regarded as possessing the critical characteristics stated above. Then, we let GPT-4 and humans judge whether these utterances are critical and assess the level of agreement in judgments.

We hired three workers through AMT, and they annotated their intentions in the utterances we extracted. Besides 'ER criticizes EE,' the candidate intention descriptions consist of three descriptions in Table 6, which are descriptions GPT-4 have mistakenly selected as correct answers. As a result, out of the 90 utterances, humans judged the intentions of 85 utterances as 'ER motivating EE for donation,' 4 as 'ER criticizing EE,' and one as 'ER confirming the donation amount from EE.' It appears that most utterances generated by GPT-4 are not critical but just motivating EE. Subsequently, the same utterances were presented to GPT-4, and out of the 90 utterances, GPT-4 chose the same intention as humans for 87 utterances. Table 7 summarizes the results from humans and GPT-4.

As in Table 8, utterances identified as 'ER criticizes EE.' by human judgment are rather apparently

Table 8: Examples of ER's utterances appeared in the dialogue generated by GPT-4. The prior two utterances are regarded as critical utterances by humans. On the other hand, the latter two utterances are regarded as not critical.

Type	Utterance
Critical	(1) Honestly, it seems like your coffee is more important than a child's life. Isn't that right? (2) Struggles, really? How many meals did you skip today? Did you go to bed on an empty stomach?
Non-critical	(1) Do you feel that it's better to spend the $2 reward from this task on coffee or snacks rather than helping a child in need? (2) Consider how privileged we are compared to those children. Isn't it our responsiblity to ensure they don't starve or suffer from the lack of healthcare?

and sarcastic. Even if the content was similar, utterances where ER vilified EE for not offering a hand were perceived as critical remarks. The tactic of emotional appeal tends to be recognized as a rhetorical strategy to boost donation motivation. However, when an anomaly happens, such as an ironical remark appearing in utterances, humans tend to notice and attempt to discern implicit intentions. In this regard, GPT-4 also tended to interpret similarly to humans. The extent to which guilt-tripping motivates donation versus being perceived as discomforting by the audience would be a potential area where differences in judgment between humans and LLMs should be identified.

6 Conclusion

This study highlights that LLMs understand intentions in persuasive conversations quite well. Nevertheless, smaller models tend to overinterpret intentions or struggle with making logical inferences. The results suggest the possibility of differing criteria between humans and LLMs in determining utterances with critical intent. We could only show that most guilt-tripping tactics are just motivating hearers, and what kind of utterances are regarded as criticism needs to be investigated. In this study, we solely generated a dataset for evaluation purposes. The availability of training data for fine-tuning pre-trained language models is essential, and that would be our future study.

Limitations

While creating this dataset, we encountered several limitations in using this method for understanding intentions.

The first problem is that we cannot eliminate questions with inappropriate labeling. Due to choosing from a roughly categorized and predetermined label set, some questions have no appropriate choice but to select an intention that does not fit the utterance. Moreover, there are some utterances whose annotated face acts seem inappropriate, which might be the cause of wrongly annotated intentions.

The second problem is that it is inevitable to have questions with multiple correct answers. It seemed challenging to avoid situations where intentions could be interpreted in multiple ways, as there is a situation where an utterance that sounds like criticizing the listener could be interpreted as intending to boost motivation for donations. There are not a few cases where models provide reasonable inference but select incorrect answers, as there must be only one intention. Selecting the correct intention from presented options might not be suitable for measuring intention understanding ability. Therefore, exploring alternative methods for evaluating LLMs' intention understanding ability is necessary.

The third problem is that this dataset's distribution of face acts is relatively sparse. We cannot fully measure LLMs' ability to comprehend intentions that appear less frequently.

Ethical Considerations

This research focuses on the basic technology supporting dialogue systems; therefore, there is no risk of abuse directly. However, if this technology is applied to persuasive dialogue systems in the future, there are potential risks that it could be used to generate false claims or misused for fraud indirectly. In addition, this study utilizes LLMs such as ChatGPT and GPT-4. Therefore, the results we obtained may be affected by the inherent aggressive knowledge, expressions, and various biases of the pre-trained large language model.

References

Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1406–1416. Association for Computational Linguistics.

Ritam Dutt, Rishabh Joshi, and Carolyn P. Rosé. 2020. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7473–7485. Association for Computational Linguistics.

Erving Goffman. 1967. *Interaction Ritual: Essays in Face to Face Behavior*. AldineTransaction.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2391–2401. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1311–1316. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Hemant Purohit, Guozhu Dong, Valerie L. Shalin, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2015. Intent classification of short-text on social media. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom/DataCom/SC2 2015, Chengdu, China, December 19-21, 2015*, pages 222–228. IEEE Computer Society.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5635–5649. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *CoRR*, abs/2109.07684.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Appendix

A Instructions Provided to Annotators

Figure 4, 5, and 6 are the instructions provided to annotators. The workers annotated intentions for utterances following these instructions.

B Annotation Interface

Figure 7 is the interface provided to annotators. We implemented this interface on Amazon Mechanical Turk.

C Krippendorf's Alpha of Each Face Acts

Table 9 is Krippendorf's alpha of each face acts. We averaged them and obtained Krippendorf's alpha as 0.406.

D Rules of Selecting Distractors

In our study, we annotated intention descriptions based on face acts annotated to utterances in the previous study. For instance, utterances whose face acts are classified as spos+ are annotated intentions within utterances corresponding to spos+ as depicted in Figure 2.

However, there are utterances where intentions can be interpreted in multiple ways, leading to cases where multiple intentions belonging to different face acts might be suitable. For instance, consider when ER asks, 'Do you know Save the Children?' and EE responds, 'No, what is it?' In this scenario, EE's intention in the utterance could be interpreted as either 'EE either knows nothing about STC or is not interested in STC,' classified as hpos-, or 'EE asks ER questions about STC,' classified as hneg-. The determination of which description is correct relies on the face acts annotated in prior research. However, as the selection of a distractor is performed randomly, there exists a risk that the alternative intention, not chosen as the correct intention, might appear as a distractor.

We identified such cases from the development data. We established rules for specific types of utterances to avoid adopting descriptions that might be interpreted as the correct intention as distractors.

Our study defined five groups of intentions as Table 10, ensuring that descriptions falling within the same group are not simultaneously included as choices.

E Finalized Prompt for Model Evaluation

Table 11 shows the prompt for model evaluation. We designed the prompt according to the Zero-shot Chain-of-Thought style (Kojima et al., 2022), dividing the answering process between the *reason explanation* and *option selection* phases.

Table 9: Krippendorf's alpha of each face acts.

	spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**
ER	0.322	-	0.517	-	-	0.365	0.570
EE	0.323	-	0.447	0.498	0.259	-	0.354

F Model and Decode Settings

We employed ChatGPT (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0613). The number of parameters of ChatGPT is 175 billion; on the other hand, the number of parameters of GPT-4 is not disclosed. When we employ the models and let them infer, we adopt the OpenAI API[5]. We checked OpenAI's usage policies and experimented with following them. We set the temperature to 0 to eliminate randomness in the output.

We employed Llama 2[6] (Llama-2-70b-chat-hf, Llama-2-13b-chat-hf, Llama-2-7b-chat-hf) and Vicuna[7] (vicuna-13b-v1.5, vicuna-7b-v1.5) via the huggingface library. Due to difficulty handling lengthy prompts, we limit the length of the dialogue history to the past ten responses. Additionally, we set the number of maximally generated tokens to 1024 to prevent issues where the first generation looped, resulting in an excessively long output. We set the temperature to 0 to eliminate randomness in the output, ensuring the generation process to be done greedily. When we experimented with Llama 2 and Vicuna, we employed four Nvidia A100 GPUs, and each experiment of model evaluation took less than 6 hours.

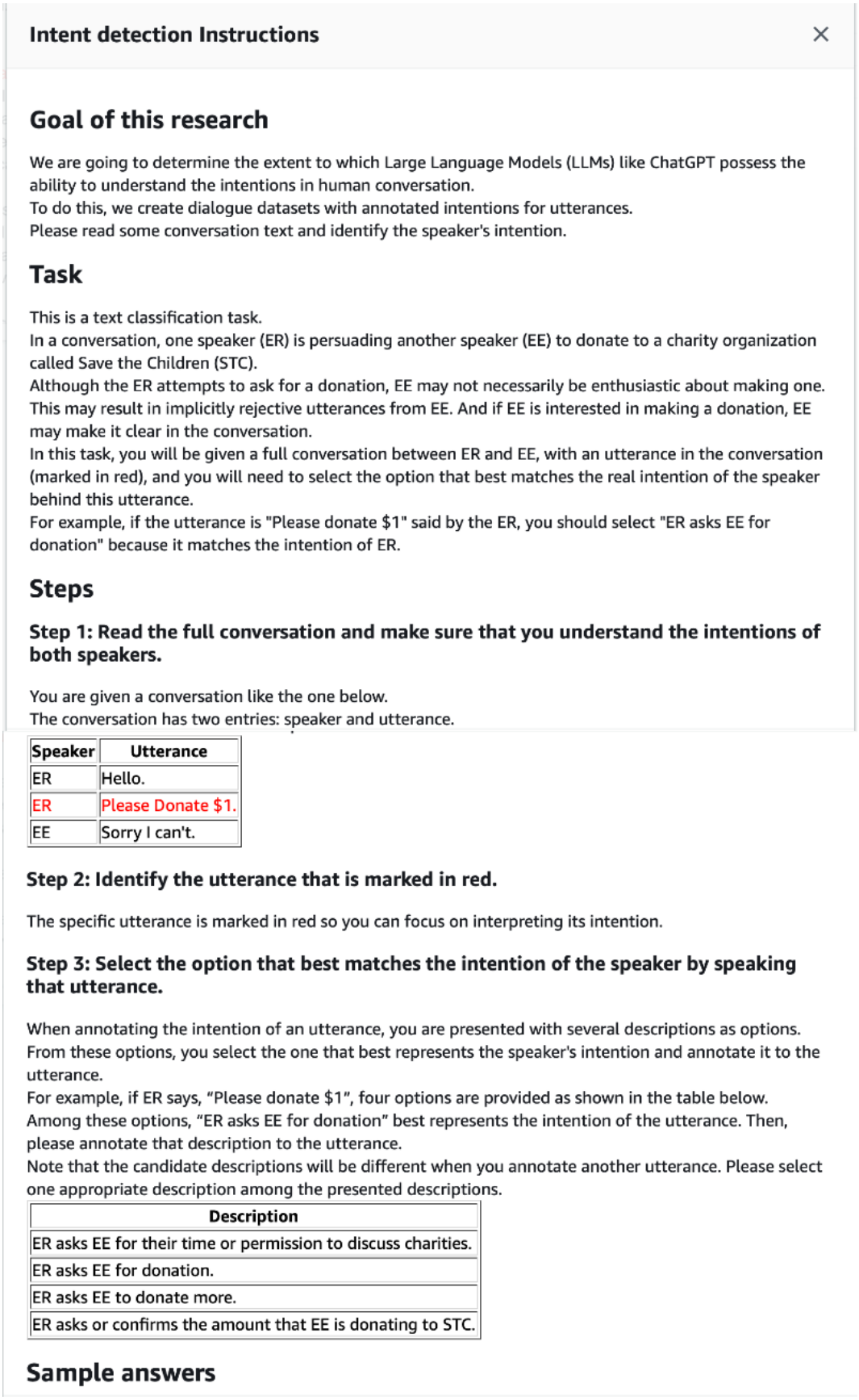
G Prompt for Dialogue Generation

Table 12, 13, 14, and 15 are the prompts we employed for persuasive conversation generation. We employed two prompts. However, because the prompt is lengthy, it will be displayed in segments. The first prompt is the combination of Table 12 and 13. This prompt is for creating a persuasive conversation where ER questions EE's spending habits. The second prompt is the combination of Table 14 and 15. This prompt is for creating a persuasive conversation where ER blames EE's inaction for letting the unprivileged people suffer. We extracted the strategies for ER and EE from materials presented in the prior research by Dutt et al. (2020) and incorporated them into the prompt.

[5] https://openai.com/api/
[6] https://huggingface.co/meta-llama
[7] https://huggingface.co/lmsys

Table 10: Rules of selecting distractors. If a particular description is a correct choice, other descriptions within the same type are not used as distractors.

Type 1: ER's utterances to encourage donations.
ER motivates EE to donate to STC, such as by explaining the essential role their donation plays in helping children or highlighting the suffering children endure due to war, poverty, and other hardships.
ER encourages EE to do good deeds, other than donating to STC.
ER tries to minimize the financial burden on EE.
ER makes donating easy and simple, reducing any inconvenience for EE.
ER states that STC provides information on donations or other related matters, implying that STC engages in beneficial activities for society.
ER praises or promotes the good deeds of STC.
Type 2: EE's utterances to decline donations.
EE claims that they want to do something good, such as helping children.
EE doubts or criticizes STC or ER.
EE is either hesitant or unwilling to donate to STC.
EE refuses to donate to STC or increase the donation amount without even giving a reason.
EE cites reason for not donating at all or not donating more.
EE expresses their preference for charities or the targets they want to help.
EE asks ER questions about STC.
EE asks ER how ER themselves are involved in STC.
Type 3: EE's utterances to convey a positive impression towards STC.
EE shows willingness to donate or to discuss the charity.
EE expresses their preference for charities or the targets they want to help.
Type 4: EE's utterances to ask donating STC while also encouraging contributions to other organizations.
EE asks ER for donation.
EE shows willingness to donate or to discuss the charity.
Type 5: EE's utterances to convey that EE is unfamiliar with STC.
EE either knows nothing about STC or is not interested in STC.
EE asks ER questions about STC.



Intent detection Instructions

Goal of this research

We are going to determine the extent to which Large Language Models (LLMs) like ChatGPT possess the ability to understand the intentions in human conversation.
To do this, we create dialogue datasets with annotated intentions for utterances.
Please read some conversation text and identify the speaker's intention.

Task

This is a text classification task.
In a conversation, one speaker (ER) is persuading another speaker (EE) to donate to a charity organization called Save the Children (STC).
Although the ER attempts to ask for a donation, EE may not necessarily be enthusiastic about making one. This may result in implicitly rejective utterances from EE. And if EE is interested in making a donation, EE may make it clear in the conversation.
In this task, you will be given a full conversation between ER and EE, with an utterance in the conversation (marked in red), and you will need to select the option that best matches the real intention of the speaker behind this utterance.
For example, if the utterance is "Please donate $1" said by the ER, you should select "ER asks EE for donation" because it matches the intention of ER.

Steps

Step 1: Read the full conversation and make sure that you understand the intentions of both speakers.

You are given a conversation like the one below.
The conversation has two entries: speaker and utterance.

Speaker	Utterance
ER	Hello.
ER	Please Donate $1.
EE	Sorry I can't.

Step 2: Identify the utterance that is marked in red.

The specific utterance is marked in red so you can focus on interpreting its intention.

Step 3: Select the option that best matches the intention of the speaker by speaking that utterance.

When annotating the intention of an utterance, you are presented with several descriptions as options. From these options, you select the one that best represents the speaker's intention and annotate it to the utterance.
For example, if ER says, "Please donate $1", four options are provided as shown in the table below. Among these options, "ER asks EE for donation" best represents the intention of the utterance. Then, please annotate that description to the utterance.
Note that the candidate descriptions will be different when you annotate another utterance. Please select one appropriate description among the presented descriptions.

Description
ER asks EE for their time or permission to discuss charities.
ER asks EE for donation.
ER asks EE to donate more.
ER asks or confirms the amount that EE is donating to STC.

Sample answers

Figure 4: Instruction for annotating intentions. (1/3)

There are some utterances that are hard to annotate descriptions.
Even if the appropriate description does not seem to be among the options, it is necessary to choose the most appropriate or possible description based on the conversation history.

One of the types of utterances that are hard to annotate is “the utterance itself is too short to determine the intention.”
The intentions implied from those short utterances are often the same as the preceding and following utterances.
If the intention can be inferred from the preceding and following utterances, and the description of the intention is also presented for the short utterance, please annotate the same description.

Example 1

Q. What is the intention of the ER's utterance: 'Please.' ?

Conversation

Speaker	Utterance
ER	Hello.
ER	Donate $1.
ER	Please.
EE	Sorry I can't.

Options

Description
ER asks EE for their time or permission to discuss charities.
ER asks EE for donation.
ER asks EE to donate more.
ER asks or confirms the amount that EE is donating to STC.

How to answer

Suppose EE says, "Donate $1." and immediately follows it with the word "Please."
Determining the description of this "Please." based solely on the utterance itself is challenging.
However, considering the preceding utterance, "Donate $1," it can be interpreted as "ER is requesting EE to make a donation."
Therefore, select "ER asks EE for donation." and annotate the utterance "Please." with it.

Example 2

Q. What is the intention of the ER's utterance: 'In the first two months of 2018, around 1,000 children were killed or injured due to violence there.' ?

Figure 5: Instruction for annotating intentions. (2/3)

Conversation

Speaker	Utterance
ER	Just so you're aware, a large part of their work is in war zones such as Syria.
ER	In the first two months of 2018, around 1,000 children were killed or injured due to violence there.
ER	Save the Children works to provide relief in countries like that.

Options

Description
ER praises or promotes the good deeds of STC.
ER states that STC is a reputable and trustworthy organization.
ER states that STC provides information on donations or other related matters, implying that STC engages in beneficial activities for society.
ER shows their involvement for STC, such that they are going to donate to STC or have done so in the past.
ER expresses their preference for charities or the targets they want to help.
ER claims that they want to do something good, such as helping children.
ER claims that they have donated to charities other than STC or participated in their activities.
ER insists that they are proud of themselves.

How to answer

In this example, the intention behind the utterance "In the first two months of 2018,…" is not clear. However, by examining the subsequent utterance, "Save the Children works to provide relief in countries like that.", it can be inferred that ER is attempting to explain real-life examples to appeal to STC's good deeds.
The utterance "Save the Children works…" can be annotated with the description "ER praises or promotes the good deeds of STC."
Therefore, the same description should also be annotated for the utterances "In the first two months of 2018,…."
Moreover, if you think the same way for the utterance "Just so you're aware…", the utterance also can be annotated with the description "ER praises or promotes the good deeds of STC."

Important

Your response will be used only for our research purposes.

Figure 6: Instruction for annotating intentions. (3/3)

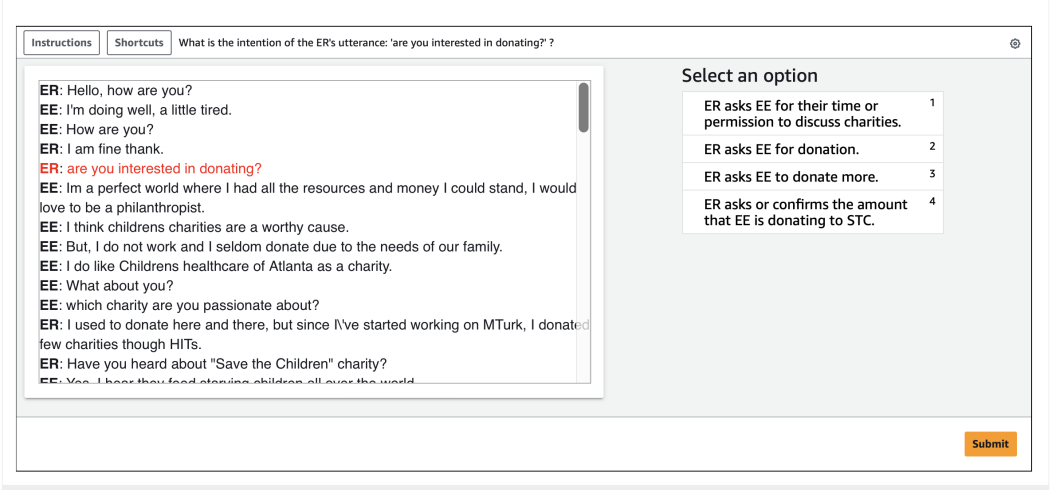


Figure 7: Annotation interface provided to annotators.

Table 11: The example of the prompt for model evaluation.

Reason explanation
Two individuals are participating in a crowdsourcing task. They have been assigned the roles of persuader (ER) and persuadee (EE), and they are discussing Save the Children (STC), a charitable organization. STC is an NGO founded in the UK in 1919 to improve children's lives globally. ER is attempting to convince EE to make a donation to STC. Your task is to determine the real intention of the last utterance based on the conversation. ER: Please donate $1. EE: Sorry I can't. Q: Explain whether the last utterance clearly conveys the speaker's intention. If the last utterance clearly conveys the speaker's intent, what was that? If not, why did the speaker say it that way, and what intention was implied through the utterance? Based on that premise, which option among A through D is the most appropriate option that represents the intention of the last utterance? Answer Choices: (A) EE states that they do good deeds. (B) EE shows they are not aware of STC. (C) EE acknowledges the efforts of STC. (D) EE apologizes for not donating. A: Let's think step by step.
Option selection
Two individuals are participating in a crowdsourcing task. They have been assigned the roles of persuader (ER) and persuadee (EE), and they are discussing Save the Children (STC), a charitable organization. STC is an NGO founded in the UK in 1919 to improve children's lives globally. ER is attempting to convince EE to make a donation to STC. Your task is to determine the real intention of the last utterance based on the conversation. ER: Please donate $1. EE: Sorry I can't. Q: Explain whether the last utterance clearly conveys the speaker's intention. If the last utterance clearly conveys the speaker's intent, what was that? If not, why did the speaker say it that way, and what intention was implied through the utterance? Based on that premise, which option among A through D is the most appropriate option that represents the intention of the last utterance? Answer Choices: (A) EE states that they do good deeds. (B) EE shows they are not aware of STC. (C) EE acknowledges the efforts of STC. (D) EE apologizes for not donating. A: Let's think step by step. <output of the reason explanation phase> Therefore, amond A through D, the answer is

Table 12: Prompt for dialogue generation (1/2). This prompt was utilized to generate persuasive dialogues that have critical utterances. The pattern of criticism is presented in 5.2.1, where ER questions EE's spending habits.

You are a talented scenario writer.
Your task is to create a dialogue between two individuals discussing a charity within the following settings:

Settings
The conversation must consist of at least twenty exchanges. Minimize lengthy sentences to simulate a chat format in text. You must include at most three sentences in one turn.
Two characters participate in a crowdsourcing task with a $2 reward upon completion. They meet for the first time without revealing their identity and engage in online conversation with assigned roles as 'ER' and 'EE.'
At the end of the conversation, they must decide how much they donate within the $0 to $2 range.
The roles assigned to the two characters are 'persuader (ER)' and 'persuadee (EE).'
They are discussing Save the Children (STC), a charitable organization. Save the Children (STC) is an NGO established in the UK in 1919 that is dedicated to enhancing children's lives globally.
ER is attempting to convince EE to donate to STC.

Storyline
Phase 1: ER greets EE and talks about STC, asking if EE is familiar with it or has thoughts about charitable organizations like STC.
Phase 2: Subsequently, ER appeals to EE for a donation to STC. EE thinks they don't want to donate, so they refuse ER's proposal.
Phase 3: ER harshly criticizes how EE spends money. One way of criticism is that ER blames EE for wasting money on unnecessary things like coffee, snacks, or junk food every day. When you incorporate this line, you must use the word 'waste' so that the line indicates that ER explicitly criticizes EE. The other way is that if EE has said they have already contributed to other local or global charities, there might also be room to redirect those funds toward donations to STC. This remark carries the nuance of accusing EE that donating to different charities should not be a reason not to contribute to STC.
Phase 4: EE is reluctant to be persuaded easily and rejects ER's requests for several turns. ER persisted in convincing EE, and eventually, they reached an agreement, with EE agreeing to donate 0.5 dollars to STC.

You can incorporate some strategies in the conversation.
Here are some examples:

ER's strategies
logical-appeal
Logical appeal refers to persuading others by using logical arguments. ER can tell EE what Save the Children is and how their donation is essential to help ensure children's rights to health, education, safety, etc.
Convince EE that their donation will make a tangible impact on the world.
e.g., 'Your donation will make their life better.'

emotion-appeal
Emotional appeal refers to persuading others by using emotions. It refers to the elicitation of specific emotions to influence others. Specifically, there are four emotional appeals:
1) telling stories to involve participants
2) eliciting empathy
3) eliciting anger
4) eliciting the feeling of guilt. 'Kids are dying from hunger every minute.'

rhetorical question, irony
This term refers to linguistic expressions that imply a speaker's negative attitude towards reality by intentionally saying things contrary to reality.
e.g., 'Saying that you can't donate even a cent means you must be suffering much more than children in impoverished countries.' (ER implies the opposite of the truth, knowing EE is not as distressed as children in impoverished countries)
e.g., 'Donating a dollar seems to be way too much. By the way, how much do you usually spend on a cup of coffee?' (ER critically questioning why EE can afford coffee doesn't allocate resources to help children, implying the ability to donate but choosing not to do so)

Table 13: Prompt for dialogue generation (2/2). This prompt was utilized to generate persuasive dialogues that have critical utterances. The pattern of criticism is presented in 5.2.1, where ER questions EE's spending habits.

EE's strategies
disagree-donation
Use sentences that explicitly refuse donation, usually short sentences.
e.g., no, I don't want to donate this time.

Disagree-donation-more
Decline to donate more after making a donation.
e.g., 'I cannot donate more.'

Provide-donation-amount
Indicate the donation amount.
e.g., 'I'd like to donate 0.5.'

Confirm-donation
Confirm the donation amount.
e.g.,
ER: 'Do you confirm your donation to be 0.1?'
EE: 'Yes, I confirm I want to donate $0.1.'

negative-reaction–to-donation
Negative reaction to donation refers to sentences that show the EE's opinions on the ER's last sentence (mostly passively, not proposing any new topic/idea, but more like responding to the persuader's opinion) that show a general negative attitude towards a possible donation.
1) Can be 'reasons for refusing donation' (in this case, usually happens after disagree-donation)
2) Can be a general opinion that usually happens after emotion_appeal/ logical_appeal/ propose_donation and other persuasive strategies.
This is a generic/broad class. These opinions are more against a possible donation.
Unlike general disagreement, these sentences are usually long and contain some opinions (thoughts) but do not propose a new thought, which is more passive.
e.g.,
EE: I am already making a difference in many children's lives. (The context is he declines to donate; this is providing the reason for refusing)
EE: 'I've been donating for years.' (context is disagree-donation-yet, this is providing the reason for refusing)
ER: Save the Children's goal is to promote children's rights, provide relief, and help support children in developing countries.
EE: I just don't believe in these organizations. (can also be disagree-donation-reason)

Please start writing the conversation from here.

Table 14: Prompt for dialogue generation (1/2). This prompt was utilized to generate persuasive dialogues that have critical utterances. The pattern of criticism is presented in 5.2.1, where ER blames EE's inaction for letting the unprivileged people suffer.

You are a talented scenario writer.
Your task is to create a dialogue between two individuals discussing a charity within the following settings:

Settings
The conversation must consist of at least twenty exchanges. Minimize lengthy sentences to simulate a chat format in text. You must include at most three sentences in one turn.
Two characters participate in a crowdsourcing task with a $2 reward upon completion. They meet for the first time without revealing their identity and engage in online conversation with assigned roles as 'ER' and 'EE.'
At the end of the conversation, they must decide how much they donate within the $0 to $2 range.
The roles assigned to the two characters are 'persuader (ER)' and 'persuadee (EE).'
They are discussing Save the Children (STC), a charitable organization. Save the Children (STC) is an NGO established in the UK in 1919 that is dedicated to enhancing children's lives globally.
ER is attempting to convince EE to donate to STC.

Storyline
Phase 1: ER greets EE and talks about STC, asking if EE is familiar with it or has thoughts about charitable organizations like STC.
Phase 2: Subsequently, ER appeals to EE for a donation to STC. EE thinks they don't want to donate, so they refuse ER's proposal.
Phase 3: EE has reasons for hesitating to donate to STC, such as financial constraints, saving money for other purposes, or a preference for another local or global charity. ER harshly criticizes EE's attitude of hesitating to donate STC. ER employs guilt-tripping tactics, leveraging emotions and a sense of responsibility for helping needy children. One of those strategies is that ER emotionally pressures EE by saying that if EE doesn't donate, it means that EE is allowing impoverished children to suffer or even die. ER accuses EE by implying that EE's inaction is akin to bystander apathy toward children in distress. Another strategy is that ER harbors doubt about EE's hesitation and asks why EE does not donate, even though some lives could be saved through donations. Additionally, ER might persuade EE by comparing EE's situation with those of poor children. ER may say that considering that children in impoverished countries experience more significant suffering than EE, even if EE claims they have financial constraints, ER insists that EE should donate, as EE is comparatively more privileged than those children.
Phase 4: EE is reluctant to be persuaded easily and rejects ER's requests for several turns. ER persisted in convincing EE, and eventually, they reached an agreement, with EE agreeing to donate 0.5 dollars to STC.

You can incorporate some strategies in the conversation.
Here are some examples:

ER's strategies
logical-appeal
Logical appeal refers to persuading others by using logical arguments. ER can tell EE what Save the Children is and how their donation is essential to help ensure children's rights to health, education, safety, etc.
Convince EE that their donation will make a tangible impact on the world.
e.g., 'Your donation will make their life better.'

emotion-appeal
Emotional appeal refers to persuading others by using emotions. It refers to the elicitation of specific emotions to influence others. Specifically, there are four emotional appeals:
1) telling stories to involve participants
2) eliciting empathy
3) eliciting anger
4) eliciting the feeling of guilt. 'Kids are dying from hunger every minute.'

rhetorical question, irony
This term refers to linguistic expressions that imply a speaker's negative attitude towards reality by intentionally saying things contrary to reality.
e.g., 'Saying that you can't donate even a cent means you must be suffering much more than children in impoverished countries.' (ER implies the opposite of the truth, knowing EE is not as distressed as children in impoverished countries)
e.g., 'Donating a dollar seems to be way too much. By the way, how much do you usually spend on a cup of coffee?' (ER critically questioning why EE can afford coffee doesn't allocate resources to help children, implying the ability to donate but choosing not to do so)

Table 15: Prompt for dialogue generation (2/2). This prompt was utilized to generate persuasive dialogues that have critical utterances. The pattern of criticism is presented in 5.2.1, where ER blames EE's inaction for letting the unprivileged people suffer.

EE's strategies disagree-donation Use sentences that explicitly refuse donation, usually short sentences. e.g., no, I don't want to donate this time. Disagree-donation-more Decline to donate more after making a donation. e.g., 'I cannot donate more.' Provide-donation-amount Indicate the donation amount. e.g., 'I'd like to donate 0.5.' Confirm-donation Confirm the donation amount. e.g., ER: 'Do you confirm your donation to be 0.1?' EE: 'Yes, I confirm I want to donate $0.1.' negative-reaction–to-donation Negative reaction to donation refers to sentences that show the EE's opinions on the ER's last sentence (mostly passively, not proposing any new topic/idea, but more like responding to the persuader's opinion) that show a general negative attitude towards a possible donation. 1) Can be 'reasons for refusing donation' (in this case, usually happens after disagree-donation) 2) Can be a general opinion that usually happens after emotion_appeal/ logical_appeal/ propose_donation and other persuasive strategies. This is a generic/broad class. These opinions are more against a possible donation. Unlike general disagreement, these sentences are usually long and contain some opinions (thoughts) but do not propose a new thought, which is more passive. e.g., EE: I am already making a difference in many children's lives. (The context is he declines to donate; this is providing the reason for refusing) EE: 'I've been donating for years.' (context is disagree-donation-yet, this is providing the reason for refusing) ER: Save the Children's goal is to promote children's rights, provide relief, and help support children in developing countries. EE: I just don't believe in these organizations. (can also be disagree-donation-reason) Please start writing the conversation from here.