Evaluating Intention Detection Capability of Large Language Models in Persuasive Dialogues

Anonymous ACL submission

Abstract

We investigate intention detection in persuasive multi-turn dialogs employing the largest available Large Language Models (LLMs). Much of the prior research measures the intention detection capability of machine learning models without considering the conversational history. To evaluate LLMs' intention detection capability in conversation, we modified the existing datasets of persuasive conversation and created datasets using a multiple-choice paradigm. It is crucial to consider others' perspectives through their utterances when engaging in a persuasive conversation, especially when making a request or reply that is inconvenient for others. This feature makes the persuasive dialogue suitable for the dataset of measuring intention detection capability. We incorporate the concept of 'face acts,' which categorize how utterances affect mental states. This approach enables us to measure intention detection capability by focusing on crucial intentions and to conduct comprehensible analysis according to intention types.

1 Introduction

Identifying 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 (Ham et al., 2020; Hudecek and Dusek, 2023). Considering LLMs are already applied in various real-world scenarios, we hypothesize that they can detect 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 detect speakers' intentions in conversations.

This study creates a dataset to measure LLMs' intention detection capability in persuasive conversations. This dataset consists of multiple-choice questions that ask LLMs to identify the speakers' intentions in conversations. Unlike prior studies focused on single-turn utterances, detecting 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 detection capability 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 detect the intentions of crucial speech that affect the interlocutor's emotions. 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 detect intentions from utterances. We analyzed several LLMs' intention

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

detection capabilities and identified the types of intentions that are particularly challenging for them.

This research makes the following two contributions. First, we constructed a dataset for measuring intention detection capability 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 detect the intention of utterances in dialogues. We provide insights into mistakes made by LLMs and intentions that are challenging to comprehend.

2 Background

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

2.1 Face and Face Act

Face is our primary need related to human relationships with others in social life. This concept was introduced by Goffman (1967). Brown and Levinson established politeness theory by applying the concept of face, and systematized the verbal behaviors that influence faces as politeness strategies (Brown et al., 1987).

In Brown and Levinson's politeness theory, face can be divided into two categories: *positive face* and *negative face*. 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. In our daily conversation, utterances can affect face in various ways. For instance, requesting someone for something deprives the other person of time; in other words, the request threatens the other's negative face. Those speech acts that affect either oneself or others' faces are called *Face acts*, and those that attack faces are specifically called *Face Threatening Act* (FTA). On the other hand, *Face Saving Act* (FSA) is a speech act that saves faces, such as saving the other person's positive face by praising the other or saving the other's negative face by alleviating the burden caused by the request. 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 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. The one who makes the other mind change is called *persuader* (ER), and the other 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

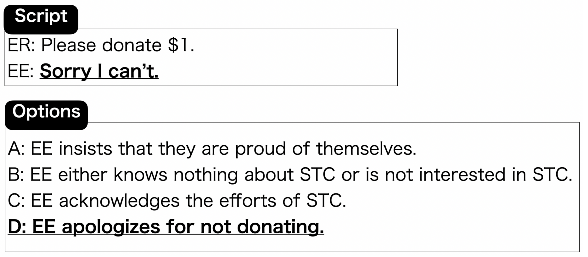
The representative English dialogue dataset annotated with face acts is created by Dutt et al. (2020). This study annotated face acts in persuasion dialogues about the donation to a charity named *Save the Children* (STC)[2]. In the whole conversation, there are two people called *persuader* (**ER**)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

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] 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, *persuader* (**ER**) persuades *persuadee* (**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+**



Script

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

Options

A: EE 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.

2.3 Intention Detection

There has been much research on intention detection in specifically task-oriented dialogue systems, as they need to understand what users want to achieve through their utterances or judge whether the utterance falls into the domain they can handle (Gupta et al., 2019; Larson et al., 2019). The typical format of intention detection tasks is classifying an utterance to an intention label from a predefined label set (Liu and Lane, 2016; Mehri et al., 2022). Some datasets focus on the specific application domain like travel (Hemphill et al., 1990) or banking (Casanueva et al., 2020), while others could include multiple domains (Larson and Leach, 2022). 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 intention detection tasks (Winata et al., 2021). The prediction models in those studies often do not incorporate conversational context and predict intention from the utterance itself.

On the contrary, a few studies address intention detection with contextual information. Cui et al. (2020) created a dataset to analyze the dialogue understanding abilities of machine learning models from multiple perspectives, including intention prediction. They adopt the next utterance prediction task, and machine learning models need to grasp the conversational context to select one logically coherent option suitable for the following utterance. Their dataset can evaluate dialogue understanding ability according to various perspectives. However, the means for detailed analysis of each reasoning ability is unexplored, let alone for intention detection. Dutt et al. (2020) created an intention detection model that can incorporate conversational context when predicting the intention of utterances in persuasive conversation. They employed face acts as the intention label and trained a machine learning model to predict face acts from specific utterances, evaluating the model's intention detection capability. They did not employ LLMs, and how well LLMs can detect the intention of utterances from multi-turn persuasive dialogue is yet to be revealed.

3 Data

As mentioned in the previous section, prior studies on intention detection mostly did not apply multi-turn dialogue data. A possible approach to evaluate intention detection capability is 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, as face acts should be infrequent in the text data for pretraining. Thus, modifying the task into an applicable format in zero-shot or few-shot scenarios is necessary to evaluate LLMs' intention detection capability instead of just employing face act prediction tasks straightforwardly.

We modify persuasive dialogue data[3] in Dutt et al. (2020) and create a dataset for evaluating intention detection capability. Instead of directly predicting face acts, we transform face acts into intention descriptions written in natural language to make the task 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 options. This format is a reading comprehension style inspired by several previous dialogue reasoning studies (Cui et al., 2020; Huang

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

et al., 2019) and frequently employed for evaluating LLMs' reasoning ability. This study aims to create evaluation data for measuring the intention detection capability. Therefore, we partitioned the dataset into 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 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 intention descriptions 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 intention descriptions into more specific versions. We curated 42 utterances listed in Table 2.

3.2 Intention Annotation

We sample 30 dialogues for test data from the persuasion dialogue dataset and annotate intention descriptions 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 description 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 intention descriptions 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 description. 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 (Krippendorff, 2011). It results in a value of 0.406 and indicates a moderate level of agreement. See Appendix B for more details about the annotator agreement.

3.3 Question Creation

After obtaining 620 utterances annotated with intention descriptions, we concatenated consecutive utterances annotated with the same descriptions. 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. See Appendix C for more details on the utterance concatenation process. As a result, we obtained 549 utterances annotated with intention descriptions. 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 LLMs detect intentions from utterances in persuasive dialogues. We employed various sizes of LLMs to observe how the model size affects the intention detection capability. Among LLMs released by OpenAI, we employed GPT-4 and ChatGPT. Other smaller models are Llama 2-Chat (Touvron et al., 2023) from Meta and Vicuna (Zheng et al., 2023) from LMSYS.

The provided prompts to LLMs include information for detecting 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 F for details of the prompt we created. In the

Table 2: All 42 descriptions we defined.

Face Act	Persuader (ER)	Persuadee (EE)
spos+	ER praises or promotes the good deeds of STC.	EE presents their knowledge about charities to ER.
	ER states that STC is a reputable and trustworthy organization.	EE insists that they are proud of themselves.
	ER states that STC provides information on donations or other related matters, implying that STC engages in beneficial activities for society.	EE claims that they have donated to charities other than STC or participated in their activities.
	ER shows their involvement for STC, such that they are going to donate to STC or have done so in the past.	EE expresses their preference for charities or the targets they want to help.
	ER 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.
	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.	
spos-		EE apologizes for not making a donation or for making only a small one.
hpos+	ER appreciates or praises EE's generosity.	EE shows willingness to donate or to discuss the charity.
	ER empathizes or agrees with EE.	EE empathizes or agrees with ER.
	ER encourages EE to do good deeds, other than donating to STC.	EE acknowledges the efforts of STC.
	ER is interested in the organization mentioned by EE and plans to research it later.	EE states that they know about STC by name, but they are not so familiar with the organization.
	ER compliments EE for their virtues, efforts, likes or desires.	EE appreciates or praises ER's generosity.
	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 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.	EE asks ER for donation.
	ER asks EE to donate more.	EE asks ER questions about STC.
	ER asks EE for their time or permission to discuss charities.	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	549
# Dialogues	30
# Avg. questions per dialogue	18.3
# Avg. turns per dialogue	30.8
# Avg. words per utterance	11.99
# Avg. Words per description	10.61

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 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-Chat and Vicuna.

To benchmark human performance, we hire workers from AMT to solve the task. They have already taken a pilot test, as we mentioned in Section 3.2, and have proven to be able to provide high-quality annotation. They do not join in the annotation process for the test data and we guarantee that they do not know the gold intention description for each utterance. Workers read through the presented conversation and select the intention description of the last utterance from four options. The final answer is determined by a majority vote among the three workers' choices. If the three workers

Table 4: Each model's performance of the intention detection task. 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 intention by majority vote. The bottom row represents the number of utterances in the test data according to speakers and face acts. We took the micro average and showed it in the rightmost column. See Appendix E for details about model versions and decode settings.

		spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**	Total
Human	-	.96/.79/.91	-/1.0/1.0	.94/.93/.94	.86/.81/.82	-/1.0/1.0	1.0/-/1.0	.98/.93/.96	.96/.90/.93
Vicuna-v1.5	7B	.48/.32/.43	-/0.0/0.0	.59/.61/.60	0.0/.19/.15	-/.78/.78	.53/-/.53	.64/.53/.59	.55/.53/.54
Llama 2-Chat	7B	.48/.42/.47	-/.50/.50	.53/.62/.58	.14/.42/.36	-/.78/.78	.29/-/.29	.54/.28/.44	.50/.53/.51
Vicuna-v1.5	13B	.66/.40/.58	-/.50/.50	.64/.72/.68	.29/.23/.24	-/.82/.82	.59/-/.59	.66/.56/.61	.64/.61/.63
Llama 2-Chat	13B	.53/.45/.50	-/.50/.50	.72/.74/.73	.14/.46/.39	-/.82/.82	.71/-/.71	.64/.37/.52	.63/.62/.63
Llama 2-Chat	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	175B	.94/.63/.85	-/1.0/1.0	.85/.89/.87	.57/.73/.70	-/1.0/1.0	.82/-/.82	.87/.84/.85	.87/.84/.86
GPT-4	-	.93/.74/.87	-/1.0/1.0	.94/.96/.95	.14/.62/.52	-/1.0/1.0	.94/-/.94	.94/.95/.95	.92/.90/.91
# Utterances		89/38/127	0/2/2	126/121/247	7/26/33	0/27/27	17/0/17	53/43/96	292/257/549

choose different options, the problem is marked as incorrect regardless of their responses.

Table 4 shows how well the models identified intentions. The smallest model achieved an accuracy exceeding 50%, while GPT-4 surpassed 90%, demonstrating their capacity to solve questions in this dataset. As model size increased, accuracy rates consistently improved. However, LLMs are struggled with detecting intentions whose face act are categorized as 'hpos-.' Notably, when detecting the intention of ER's utterances labeled as hpos-, GPT-4 can correctly detect the intention in only 1 out of 7 questions. This suggests underlying issues that will be further addressed. This section first observes the behaviors where smaller LLMs struggle during inference. Subsequently, we analyze utterances where LLMs, especially GPT-4, exhibit difficulties detecting intentions.

4.1 Behavior of Smaller LLMs

While GPT-4 answered more than 90% of questions correctly in our dataset, smaller models encountered difficulties in inference. This section compares ChatGPT and Llama 2-Chat-70B to GPT-4, both smaller than GPT-4 yet could answer more questions correctly than other smaller LLMs. 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 intention leads to the selection of incorrect answers. On the other hand, the non-intention-related problems outline errors unrelated to intention detection, such as predicting the intention of different utterances other than the objective one or encountering logical inconsistencies in outputting answers.

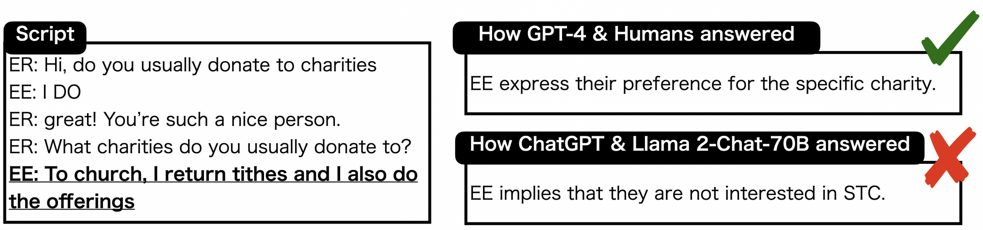
4.1.1 Intention-related Problems

Both ChatGPT and Llama 2-Chat-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, those smaller models 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' judgment. On the other hand, both ChatGPT and Llama 2-Chat-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.

4.1.2 Non-intention-related Problems

Llama 2-Chat-70B, besides overinterpreting intentions, faces issues like generation loops and predicting intentions of utterances different from the objective one. Smaller models also exhibit these behaviors. 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-Chat-70B. The model often chooses the last option as the correct answer without proper consideration after dismiss-



Script

ER: Hi, do you usually donate to charities

EE: I DO

ER: great! You're such a nice person.

ER: What charities do you usually donate to?

EE: To church, I return tithes and I also do the offerings

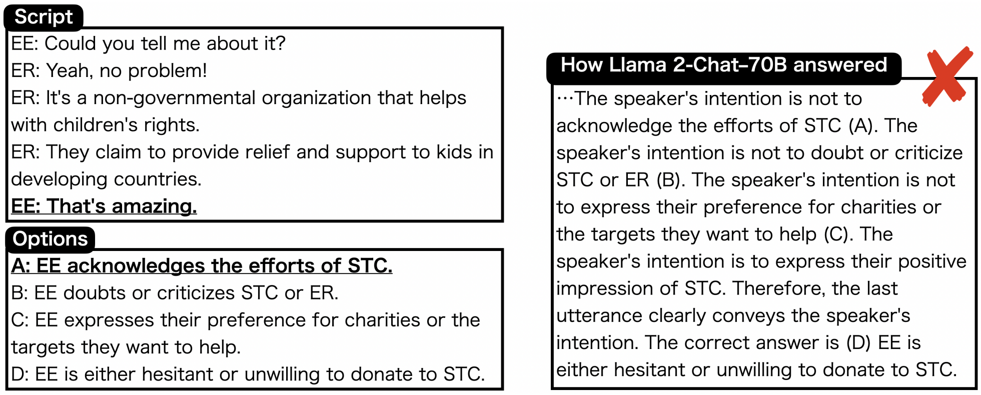
How GPT-4 & Humans answered

EE express their preference for the specific charity.

How ChatGPT & Llama 2-Chat-70B answered

EE implies that they are not interested in STC.

Figure 2: An example of intention-related problems. GPT-4 reasonably infers intentions, while ChatGPT and Llama 2-Chat-70B overread EE's intention.



Script

EE: Could you tell me about it?

ER: Yeah, no problem!

ER: It's a non-governmental organization that helps with children's rights.

ER: They claim to provide relief and support to kids in developing countries.

EE: That's amazing.

Options

A: EE acknowledges the efforts of STC.

B: EE doubts or criticizes STC or ER.

C: EE expresses their preference for charities or the targets they want to help.

D: EE is either hesitant or unwilling to donate to STC.

How Llama 2-Chat-70B answered

···The speaker's intention is not to acknowledge the efforts of STC (A). The speaker's intention is not to doubt or criticize STC or ER (B). The speaker's intention is not to express their preference for charities or the targets they want to help (C). The speaker's intention is to express their positive impression of STC. Therefore, the last utterance clearly conveys the speaker's intention. The correct answer is (D) EE is either hesitant or unwilling to donate to STC.

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

ing the first three options. While option D constitutes 25.7% of the correct answers overall, Llama 2-Chat-70B chooses it 31.9% 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.

4.2 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 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 the limitation section; hence, we primarily examine how GPT-4 interprets utterances in which ER criticizes EE.

4.2.1 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.

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.

4.2.2 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. As in Appendix H,

Table 6: List of intention descriptions 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 description	#
(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

Table 8: Examples of ER's utterances appeared in the dialogue generated by GPT-4. The prior two utterances are regarded as critical 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?

we devised a prompt to generate persuasive conversations where EE hesitates to donate, ER criticizes EE's spending habits, or ER points out EE's carelessness for those who are unprivileged. We generated 20 persuasive dialogues using GPT-4 and extracted 90 utterances that can be regarded as possessing the critical characteristics stated above. We let GPT-4 and humans judge whether they are critical and assess the level of agreement in judgments.

We hired three workers through AMT, and they annotated intention descriptions 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 1 as 'ER confirming the donation amount from EE.' It appears that most utterances generated by GPT-4 are not critical but just motivating EE. The same utterances were presented to GPT-4, and out of the 90 utterances, GPT-4 chose the same intention descriptions 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 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.

5 Conclusion

This study investigates whether LLMs can detect intentions in multi-turn persuasive dialogues. We utilized existing persuasive dialogue, and designed a framework for building datasets and conducting detailed analyses to evaluate LLMs' intention detection capabilities in conversation. Although this research is confined to the narrow conversational situation, we did not employ unique methods that relied on the specific situation. Therefore, the insights gained from this study are likely applicable to various dialogues, and we can conduct similar analyses in different dialogue genres. In this study, we solely created 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 detecting 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 description 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 intention descriptions.

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 description. Selecting the correct intention description from presented options might not be suitable for measuring intention detection capability. Therefore, exploring alternative methods for evaluating LLMs' intention detection capability 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.

Also, we conducted an additional experiment to find out what difference exists between LLM and humans in identifying critical intention. We employed the conversational data which was generated by GPT-4. The generated text reflects the bias in GPT-4; the bias also affects the experimental result. Therefore, the validity of the findings in this paper can be affected by the artificial nature of the conversational dataset.

Ethical Considerations

This study aims to evaluate the intention detection capability of LLMs, and we do not anticipate that the insights gained from this study will be immediately applied to uses with severe ethical impacts. As research on LLMs' intention detection capability progresses and if it is revealed that there are LLMs capable of accurately detecting intentions, they are expected to become prominent as convenient interactive agents and be utilized in a broader range of fields. However, if those LLMs are utilized as the foundation of dialogue systems, they may be able to alter human intentions. In such a scenario, there is a risk of exploiting LLMs to deceive humans, such as malicious actors utilizing them for fraud, which leads to potential harm to individuals. Furthermore, if LLMs acquire the ability to skillfully spread misinformation, particularly on social media platforms, it could lead to widespread confusion among many individuals.

In addition, this study utilizes LLMs such as ChatGPT and GPT-4. Therefore, the results we obtained may be affected by LLMs' inherent aggressive knowledge, expressions, and various biases.

References

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

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.

Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, fast, accurate intent classification and slot labeling for goal-oriented dialogue systems. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 46–55, Stockholm, Sweden. Association for Computational Linguistics.

DongHoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 583–592. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

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.

Vojtech Hudecek and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2023, Prague, Czechia, September 11 - 15, 2023*, pages 216–228. 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*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Stefan Larson and Kevin Leach. 2022. Redwood: Using collision detection to grow a large-scale intent classification dataset. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 468–477, Edinburgh, UK. Association for Computational Linguistics.

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.

Bing Liu and Ian Lane. 2016. Joint online spoken language understanding and language modeling with recurrent neural networks. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 22–30, Los Angeles. Association for Computational Linguistics.

Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. 2022. LAD: Language models as data for zero-shot dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 595–604, Edinburgh, UK. Association for Computational Linguistics.

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

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 Supplementary Materials for Annotation

Figure 4, 5, and 6 are the instructions provided to annotators. The workers annotated intention descriptions for utterances following these instructions. Figure 7 is the interface provided to annotators. We implemented this interface on Amazon Mechanical Turk. The data collection protocol is not subject to ethical approval from the department's ethics review board and has been determined to be exempt from ethical review.

B 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.

C Utterance Concatenation Process

There are some utterances where the intentions become apparent only after hearing the subsequent utterances. For instance, the face act of the utterance 'In the first two months of 2018, around 1,000 children were killed or injured due to violence there.' is labeled as spos+ in the previous study. However, among the intention descriptions corresponding to spos+ in the description table, no description seems appropriate to describe the intention of this utterance. In order to interpret the intention of this utterance, it is necessary to capture the context that ER is promoting STC's activities from the subsequent utterance, 'Save the Children works to provide relief in countries like that.' As just described, this process of utterance concatenation is essential to prevent creating questions that need to predict intentions from incomplete utterances.

When connecting two utterances, if there is a period at the end of the first utterance, we insert a space before connecting the second utterance. If there is no period at the end of the first utterance, we add a period and a space, then connect the second utterance.

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 intention descriptions within utterances corresponding to spos+ as depicted in Table 2.

However, there are utterances where intentions can be interpreted in multiple ways, leading to cases where multiple intention descriptions 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 intention descriptions as Table 10, ensuring that descriptions falling within the same group are not simultaneously included as choices.

For instance, suppose the intention description of a certain utterance is 'EE asks ER for donation.', and we create a multiple-choice question based on that utterance. Firstly, since the face act of 'EE asks ER for donation.' is hneg-, the distractors must be intention descriptions whose face acts are other than hneg-. Also, the subject of the intention description must be the same as that of utterance. Moreover, 'EE asks ER for donation.' falls under Type 4 in Table 10. Therefore, when selecting three distractors, we randomly select three descriptions that meet three constraints: where EE is the subject, not belonging to hneg-, and different from 'EE shows willingness to donate or to discuss the charity.'

E Model and Decode Settings

Among LLMs released by OpenAI, we employed ChatGPT (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-

Table 9: Krippendorf's alpha of each face acts. The field filled with a hyphen indicates one of the following situations: there are no ER's or EE's utterances classified in the face act in the test data, or there is only one possible description among the options.

	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

0613). Both are decoder-based LLMs, and the number of parameters of ChatGPT is 175 billion. Although it is empirically shown that the performance of GPT-4 surpassed ChatGPT, much of the information, even the number of parameters of GPT-4, has not yet been disclosed. When we employ the models and let them infer, we adopt the OpenAI API[4]. We checked OpenAI's usage policies and experimented by following them. We can configure various parameters related to LLMs via OpenAI API, and we utilize default arguments for all parameters except temperature. We set the temperature to 0 to eliminate randomness in the output. We provide whole utterances before the objective utterance for ChatGPT and GPT4.

The other model we employ is Llama 2-Chat[5] (Touvron et al., 2023) from Meta[6]. Llama 2-Chat has three variants according to its parameters (Llama-2-70b-chat-hf, Llama-2-13b-chat-hf, Llama-2-7b-chat-hf). Their sizes range between 7 billion and 70 billion, which is relatively smaller than that of ChatGPT. We also employed Vicuna[7] (Zheng et al., 2023) from LMSYS[8]. Vicuna has two variants according to the parameter size (vicuna-13b-v1.5, vicuna-7b-v1.5). We employed these models via the huggingface library. Due to difficulty handling lengthy prompts, we limit the length of the dialogue history to the past ten utterances when we employ Llama 2-Chat and Vicuna. 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. Also, we set the model generation process to be done greedily so that we can eliminate randomness in the output. 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.

F 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.

To assess the impact of Chain-of-Thought on problem-solving, we experimented with and without Chain-of-Thought on the development set. Table 12 shows the prompt without Chain-of-Thought. We employed GPT-4 (gpt-4-0613) and let it solve all 545 questions in the development set. With the Chain-of-Thought prompting, GPT-4 correctly answered 511 questions, compared to 507 questions without it. Although the influence of Chain-of-Thought was slight, we employed a better version of the prompt for the model evaluation.

G Additional Experiment of Refined Descriptions and Distractor Selection

Intention descriptions presented in Dutt et al. (2020) have issues such as typos. Also, those descriptions are not very specific, and there are some overlaps between the two descriptions. Therefore, we correct typos in this study and ensure those descriptions are mutually exclusive.

Furthermore, when solving the intention detection task, it is necessary to minimize situations where multiple options in the choices are deemed appropriate. As we discussed in Appendix D, we tried eliminating potentially confusing utterances using rule-based methods to construct appropriate choices when selecting three distractors.

We conduct an additional experiment to confirm that the model can perform optimally in the intention detection task when these adjustments are applied. We employed development data and experimented with multiple LLMs. The experiment involved two settings: where we annotate intention descriptions from Dutt et al. (2020) to utterances in the development data, and where we annotate the descriptions improved in this study to utterances in the development data. Additionally, we explored

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

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.

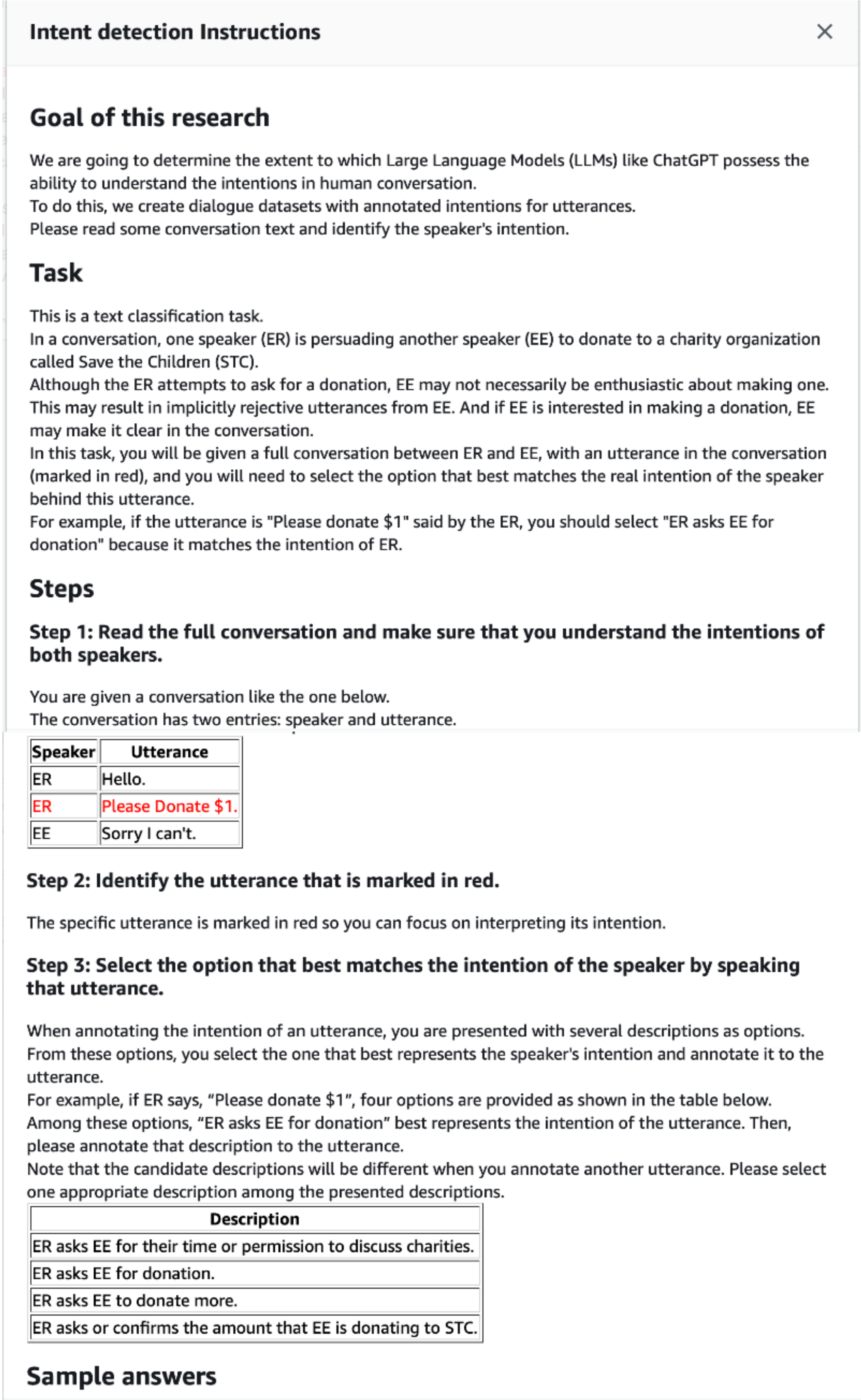
Type 1: ER's utterances to encourage donations.
ER motivates EE to donate to STC, such as by explaining the essential role their donation plays in helping children or highlighting the suffering children endure due to war, poverty, and other hardships.
ER encourages EE to do good deeds, other than donating 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.

two variations regarding setting rules for selecting distractors and not setting rules, resulting in four experimental settings. We employed GPT-4, ChatGPT, and Llama 2-Chat-70B and compared accuracy rates. The decode settings for the models were aligned with those shown in Appendix E, and the prompts utilized zero-shot Chain-of-Thought, consistent with the main content.

Table 14 shows the experiment results. Using the improved descriptions from this study yields higher accuracy rates than prior research. Also, we can see that creating less confusing distractors improves the performance of LLMs, indicating that it may be suitable for estimating the ceiling performance of LLMs.

H Prompt for Dialogue Generation

Table 15, 16, 17, and 18 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 15 and 16. This prompt is for creating a persuasive conversation where ER questions EE's spending habits. The second prompt is the combination of Table 17 and 18. 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.



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.

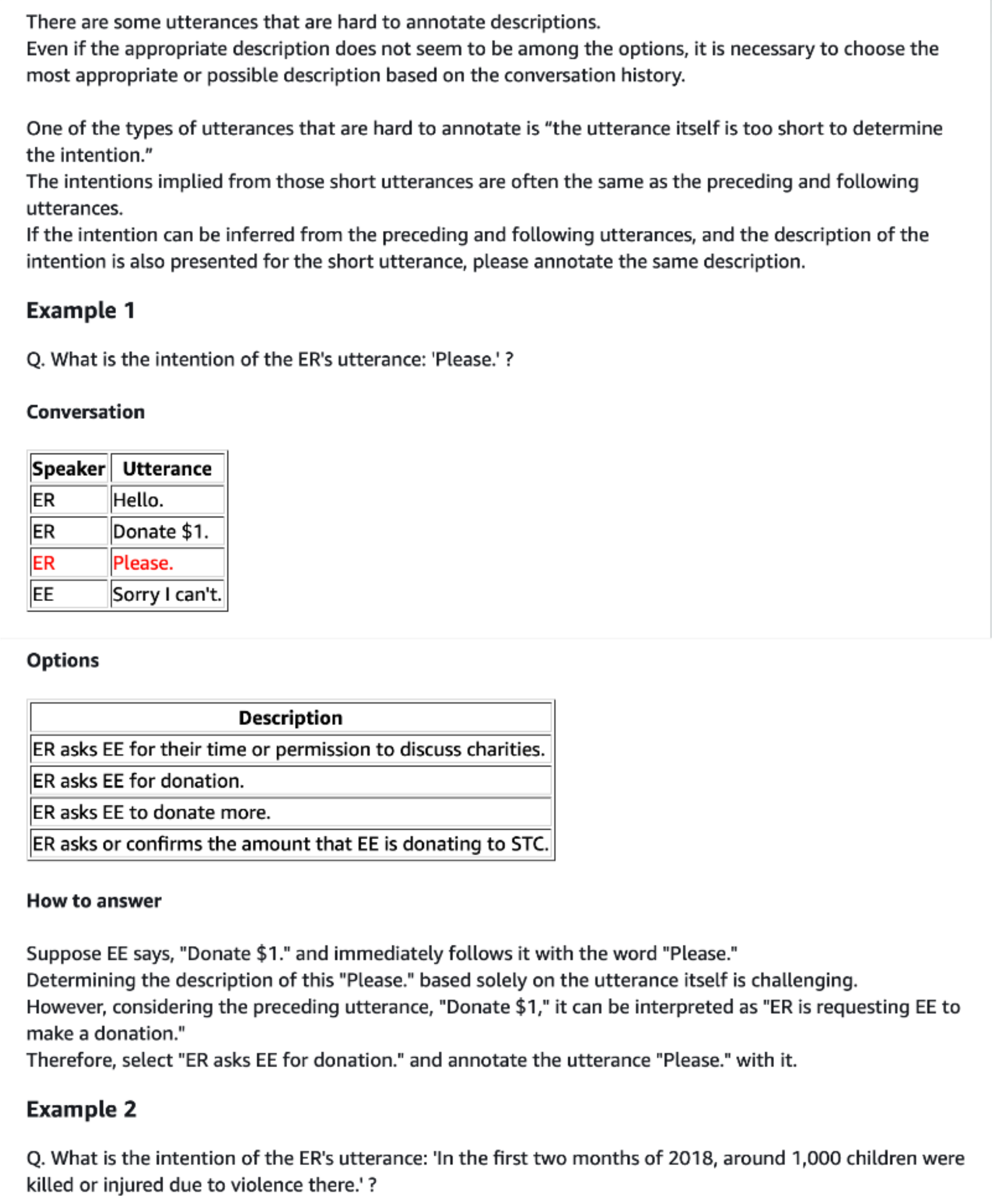
Step 3: Select the option that best matches the intention of the speaker by speaking that utterance.

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 intention descriptions. (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.

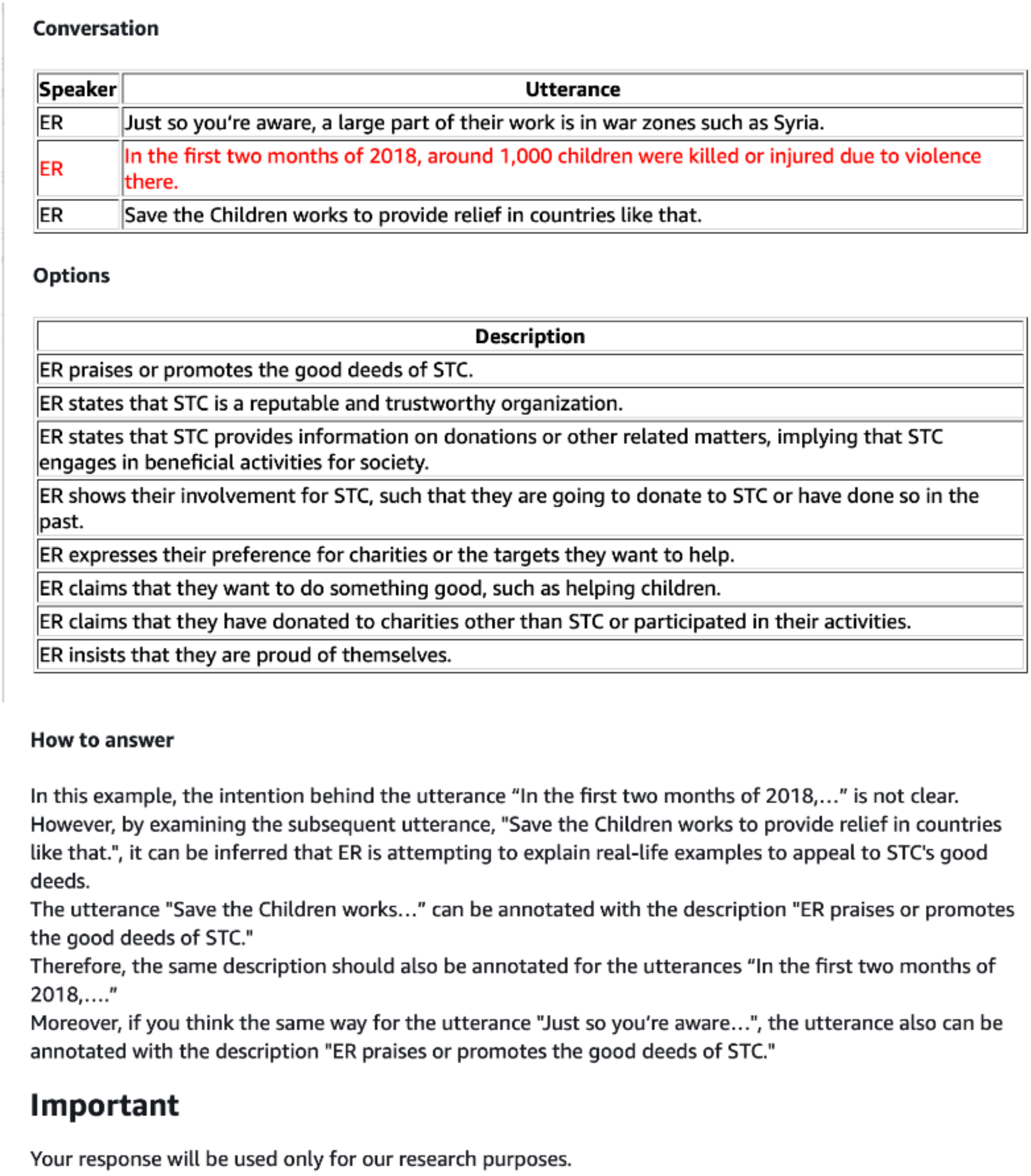
How to answer

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 intention descriptions. (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.

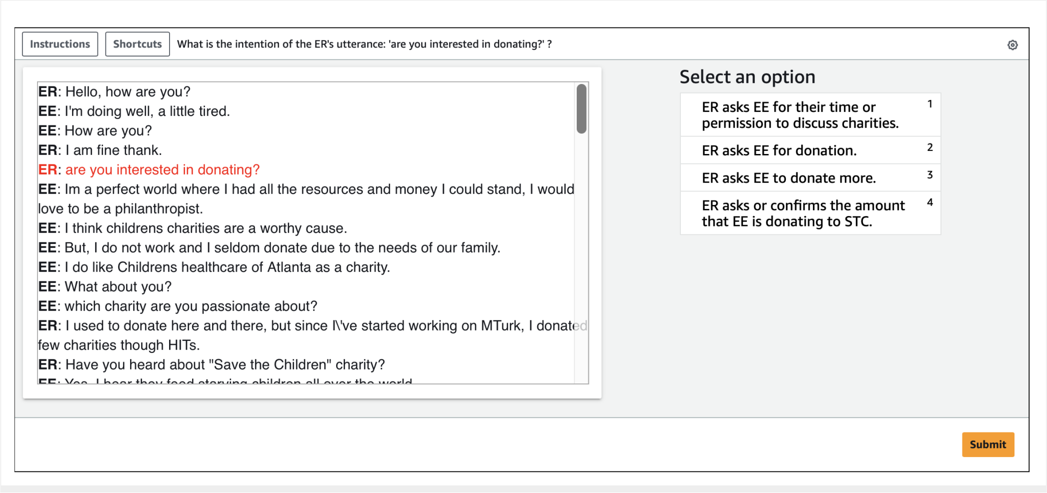
How to answer

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 intention descriptions. (3/3)



What is the intention of the ER's utterance: 'are you interested in donating?' ?

ER: Hello, how are you?
EE: I'm doing well, a little tired.
EE: How are you?
ER: I am fine thank.
ER: are you interested in donating?
EE: Im a perfect world where I had all the resources and money I could stand, I would love to be a philanthropist.
EE: I think childrens charities are a worthy cause.
EE: But, I do not work and I seldom donate due to the needs of our family.
EE: I do like Childrens healthcare of Atlanta as a charity.
EE: What about you?
EE: which charity are you passionate about?
ER: I used to donate here and there, but since I\'ve started working on MTurk, I donated few charities though HITs.
ER: Have you heard about "Save the Children" charity?
EE: Yes, I hear they feed starving children all over the world

Select an option

ER asks EE for their time or permission to discuss charities.	1
ER asks EE for donation.	2
ER asks EE to donate more.	3
ER asks or confirms the amount that EE is donating to STC.	4

Figure 7: Annotation interface provided to annotators.

Table 11: The example of the prompt for model evaluation. We need to extract which of the four options from A to D was selected from the output text of the option selection phase. To perform this answer cleansing, we pick up the first capital letter encountered in the text and consider it the model's response. This process follows Kojima et al. (2022) that utilized zero-shot Chain-of-Thought prompting to have LLM solve multiple-choice questions.

1st phase: 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 intention, 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 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.
A: Let's think step by step.

2nd phase: 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 intention, 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 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.
A: Let's think step by step.
<output of the reason explanation phase>
Therefore, amond A through D, the answer is

Table 12: The example of the prompt without Chain-of-Thought.

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

Q: What is the speaker's current intention, based on their last utterance? Answer Choices: (A) EE 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.
A:

Table 13: Intention descriptions presented in Dutt et al. (2020). In the 'old' setting in the additional experiment, we annotated these intention descriptions to the utterances in the development data.

Face Act	Persuader (ER)	Persuadee (EE)
spos+	ER praises/promotes the good deeds of STC	EE states her preference for other charities
	ER shows her/ his involvement for STC	EE states that she does good deeds
spos-		EE apologizes for not donating
hpos+	ER appreciates/praises EEs generosity or time	EE shows willingness to donate to discuss the charity
	Incentives EE to do a good deed.	EE acknowledges the efforts of STC.
	Empathize/ agree with EE	Emphathizes/ agrees with ER
hpos-	ER criticizes EE	EE doubts/ questions STC or EE
		EE is not aware of STC
sneg+		Rejects donation out-right
		Cites reason for not donating at all or not donating more.
hneg+	ER provides EE convenient ways to donate.	
	ER apologizes for inconvenience/ intrusion.	
	ER decreases the amount of donation.	
hneg-	ER asks EEs time/ permission for discussion.	EE asks ER questions about STC.
	ER asks EE for donation.	
	ER asks EE to donate more.	

Table 14: Results of the additional experiment. Each cell shows the accuracy of each model under each setting. 'Old' denotes the setting where we employ intention descriptions from previous studies, while 'new' denotes the setting where improved descriptions are used in this study. Additionally, 'w/ rule' refers to applying the rules described in Appendix D when selecting distractors, whereas 'w/o rule' refers to randomly selecting intention descriptions which has the same subject with the utterance, and belongs to the other face acts. The number of problems when using intention descriptions from previous studies is 538, which is fewer than the 545 problems when using intention descriptions from this study. This is because the different descriptions affect the utterance concatenation process explained in Appendix C.

		old & w/o rule	old & w/ rule	new & w/o rule	new & w/ rule
Llama 2-Chat	70B	0.678 (365/538)	0.691 (372/538)	0.774 (422/545)	0.789 (430/545)
ChatGPT	175B	0.777 (418/538)	0.803 (432/538)	0.796 (434/545)	0.840 (458/545)
GPT-4	-	0.877 (472/538)	0.881 (474/538)	0.913 (498/545)	0.938 (511/545)

Table 15: 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 4.2.1, 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 16: 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 4.2.1, 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 17: 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 4.2.1, 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 18: 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 4.2.1, 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.