Supplementary Material

Contents

A More Experimental Results for Pragmatic Identification and Reasoning

We conducted additional experiments on PIR employing various models, including RoBERTa$_{\text{large}}$ [50], DeBERTa$_{\text{base}}$ [65], and ALBERT$_{\text{base}}$ [66]. The training and testing procedures remained consistent with the aforementioned models described in the main body. All experimental results have been compiled and presented in Table 6. Analysing newly proposed result, it's obvious to observe that our conclusions mentioned in the main body still hold.

Table 6: Pragmatics Identification and Reasoning Results. The numerical results are accuracy scores in their percentage.

	C → P	**CP → R**	**C → PR**
Random	50	20	10
BERT$_{\text{base}}$	63.2 ± 1.1	91.3 ± 0.7	**50.2 ± 6.8**
RoBERTa$_{\text{base}}$	64.4 ± 1.3	92.0 ± 0.4	50.0 ± 11.28
RoBERTa$_{\text{large}}$	63.8 ± 0.0	60.8 ± 0.5	0.0 ± 0.0
GPT-2$_{\text{base}}$	64.4 ± 0.7	90.9 ± 0.9	13.06 ± 1.1
DialoGPT$_{\text{medium}}$	65.0 ± 0.6	24.5 ± 1.9	3.8 ± 1.5
DeBERTa$_{\text{base}}$	64.9 ± 0.2	**92.6 ± 0.6**	43.9 ± 1.2
ALBERT$_{\text{base}}$	**65.1 ± 0.4**	90.6 ± 0.2	34.9 ± 1.8

B Annotation Details

B.1 Details For Automatic Selection

Different methodologies are employed to address various pragmatic phenomena. To leverage prior advancements in the field, we begin by segmenting each dialogue into individual utterances. Subsequently, we employ two distinct approaches, namely string matching and pretrained model classification, to identify these phenomena within our source data. In the case of scalar implicature, which exhibits a noticeable pattern characterized by word pairs such as *(some, all)* appearing in adjacent turns of dialogues, we employ string matching to annotate instances of scalar implicature in conversations. Similarly, for popeq implicature, which often features a continuous question mark, we utilize this characteristic as a means of detection. With regards to idioms, which exhibit more evident patterns, we employ the idiom set proposed by Saxena and Paul [18] to conduct searches. For other types of phenomena that lack obvious patterns, we leverage a pretrained RoBERTa base model [50], and fine-tune it for our specific task. The sarcasm dataset by Misra [67] is used for finetuning the sarcasm model, the MOVER dataset by Zhang and Wan [42] for hyperbole and the ColBERT dataset by Annamoradnejad and Zoghi [20] for paronomasia. Several models have been proposed for metaphor detection, thus we utilize an existing model [68] specifically designed for metaphor identification.

Topic Segmentation The original dialogues employed in our study consist of lengthy and multi-turn exchanges, which are ill-suited for our research objectives. Consequently, we implement a segmentation process to break down these dialogues into shorter units. To achieve this, we employ two techniques, namely BERTScore [52] and TextTiling [69]. The segmentation procedure starts with computing the BERTScore between adjacent turns and subsequently applying the TextTiling algorithm to the generated BERTScores.

B.2 Details For Fine-grained Annotation

AMT is integral to our process. To ensure clarity and consistency, we provide explicit instructions to the workers. Additionally, to further elucidate the objectives of our study, we offer illustrative examples. The task itself is presented below the instructions and examples, with the dialogue and corresponding turn numbers provided for workers to select. Furthermore, as workers check a checkbox, we prompt them to select a confidence score and provide a rationale. In order to strike a balance between our budget, the quality of annotations, and the speed of annotation, we have determined the compensation of $0.1 per completed task. The whole view of the worker interface is presented in Figure 7. After the annotation process, we collect responses that are assigned with a confidence score of 4 or higher.

Specifically, we surveyed 10 users to accomplish our task. All users can complete a single task within 45 seconds, leading to a wage pay of around 8 dollars per hour, which is about a dollar higher than the federal minimum hourly wage of the United States.

B.3 Details on Human Refinements

Disturbing choices are chosen based on the BERTScore metric [52]. The rationale with the highest similarity, as determined by other dialogues, is selected and included in the pool of candidate options. The instructions provided to the workers align with those used for Fine-grained Annotation, wherein they are also instructed to assign a confidence score to their responses. The remuneration for workers is set at $0.05 per task. The worker interface is included in Figure 8.

AMT Workers Requirements In order to guarantee the quality of annotated data, the qualification rules for workers are strict and can be found in Table 7.

Table 7: AMT workers requirements

Country$_{\text{In}}$	United States, Canada, Great Britain, Australia, Singapore, Ireland, New Zealand
# Tasks approved$_{\text{GreaterThanOrEqualTo}}$	1300
Tasks approved Rate$_{\text{GreaterThanOrEqualTo}}$	95%

C Experimental Detail

C.1 Pragmatic Identification and Reasoning (PIR)

BERT$_{\text{base}}$ [70] BERT (Bidirectional Encoder Representations from Transformers) is a revolutionary language representation model that has had a significant impact on natural language processing (NLP) tasks. It has achieved remarkable performance across various NLP benchmarks, including question answering, sentiment analysis, named entity recognition, and many others. Its birth brings profound influence on pretrained language models.

RoBERTa$_{\text{base}}$ & RoBERTa$_{\text{large}}$ [50] RoBERTa improves upon BERT by incorporating enhancements such as larger and more diverse training data, longer pretraining duration, dynamic masking, and advanced training strategies. These improvements enable RoBERTa to achieve even better performance on a wide range of NLP benchmarks. While BERT paved the way for contextualized representations in NLP, RoBERTa further refines and pushes the boundaries of language understanding, making it a powerful and preferred choice for many researchers and practitioners in the field.

ALBERT$_{\text{base}}$ & ALBERT $_{\text{large}}$ [66] ALBERT (A Lite BERT) is a highly efficient and compact variant of the BERT model that addresses the computational limitations of the original architecture. It incorporates parameter-reduction techniques to alleviate training time constraints and achieve improved performance compared to BERT.

DeBERTa$_{\text{base}}$ [62] DeBERTa (Decoding-enhanced BERT with Disentangled Attention) is a state-of-the-art language representation model that builds upon the BERT architecture and introduces several key innovations, including disentangled attention mechanism. The performance of DeBERTa has been demonstrated to surpass that of BERT on a wide range of NLP tasks.

GPT2$_{\text{base}}$ [71] Leveraging transformers decoder, Radford et al. [71] proposed GPT2. It represents a significant breakthrough in natural language processing and generation. One of the most notable

Table 8: Hyperparameters for models on **CP** $\rightarrow$ **R**

Model	learning rate	batch size	weight decay	epochs
$\text{BERT}_{\text{base}}$	5e-5	12	0.001	50
$\text{BERT}_{\text{large}}$	5e-5	12	0.001	50
$\text{ALBERT}_{\text{base}}$	5e-5	12	0.001	50
$\text{ALBERT}_{\text{large}}$	5e-5	12	0.001	50
$\text{DeBERTa}_{\text{base}}$	5e-5	12	0.001	50
$\text{RoBERTa}_{\text{base}}$	5e-5	12	0.001	50
$\text{RoBERTa}_{\text{large}}$	5e-5	12	0.001	50
$\text{GPT2}_{\text{base}}$	0.001	8	0.01	50
$\text{DialoGPT}_{\text{medium}}$	0.001	2	0.01	50

Table 9: Batch size for models on **C** $\rightarrow$ **P**

Model	Batch Size
$\text{BERT}_{\text{base}}$	80
$\text{ALBERT}_{\text{base}}$	24
$\text{ALBERT}_{\text{large}}$	24
$\text{DeBERTa}_{\text{base}}$	24
$\text{RoBERTa}_{\text{base}}$	80
$\text{RoBERTa}_{\text{large}}$	24
$\text{GPT2}_{\text{base}}$	24
$\text{DialoGPT}_{\text{medium}}$	8

features of GPT-2 is its ability to generate coherent and contextually relevant text. Through unsupervised pretraining on a large corpus of internet text, GPT-2 learns to predict the next word in a sequence of text, enabling it to generate human-like responses.

DialogGPT$_{\text{medium}}$ [35] DialogGPT is dialogue-oriented GPT. It builds upon the GPT architecture and extends it to support interactive conversations. DialogGPT is trained in a supervised manner using a dialogue dataset, which allows it to understand and generate responses in a conversational context.

The PIR task encompasses three distinct settings: **C** $\rightarrow$ **P**, **CP** $\rightarrow$ **R**, and **C** $\rightarrow$ **PR**. In the **C** $\rightarrow$ **P** setting, models are trained for 20 epochs, employing a batch size as indicated in Table 9, a learning rate of $2e-5$, and weight decay of 0.01. As for **CP** $\rightarrow$ **R** , the hyperparameters adopted are listed in Table 8. For the **C** $\rightarrow$ **PR** setting, there is no training required; instead, we simply load the best checkpoint obtained from the previous training for this task. The concrete implementation is as follows: we initially flatten the test dataset of **C** $\rightarrow$ **P**, ensuring that each instance contains both a dialogue and a pragmatic turn extracted from the same dialogue. As for the test dataset of **CP** $\rightarrow$ **R**, no modifications are made. It should be noted that, following the processing steps, both datasets own the same dialogues and corresponding pragmatic turns, resulting in identical instance numbers. For an instance to be deemed correct, the models must successfully accomplish both component tasks *i.e*. succeed in Identification and Reasoning.

C.2 Conversational Question Answering (CQA)

CQA ChatGPT was instructed to generate questions for our tasks. The prompt template that starts the questions with "Which" is depicted in Table 10. Through this methodology, we collected a total of 19,482 questions. To ensure the reliability of the answers provided to these questions, AMT is utilized. The task template is demonstrated in Figure 9. In our experiment, the hyperparameters adopted are illustrated in Table 11. To assess the performance of ChatGPT, we conducted testing using the template outlined in Table 13.

Zero-Shot Natural Language Inference Details are provided as follows. T5-XXL, and DeBERTa-v3 are tested with the pragmatic turn as premise and implied meaning as a hypothesis. The context

Table 10: ChatGPT question generation template: using "Which" to start the question.

```
You are sensitive and always view others' words as having some implied
meanings.
For the dialogue between "A" and "B" in this task, we have offered a
statement that is the implied meaning of a turn, please only offer
one reading comprehension question that can be answered with only one
word based on the dialogue and mostly focuses on the turn the statement
mentions.
The question will be tested by only by viewing the dialogue, so please
make the question hard enough that it's impossible to answer without
viewing the statement.
Use "Which" to ask the question!
Following is the dialogue:
{dialogue}
Following is the statement:
{statement}
Use "Which" to ask the question! And please make the question hard
enough that it's impossible to answer without viewing
```

Table 11: Hyperparameters for models on CQA.

Training Epoch	50
Learning Rate	$5.6e-5$
Batch Size	24
Weight Decay	0.001

Table 12: Test ChatGPT: answer questions with only one word.

```
For the dialogue between "A" and "B" in this task, please answer a
question according to the dialogue with only one word
Following is the dialogue:
{dialogue}
Following is the question: {question}
```

Table 13: ChatGPT test template of Zero-Shot CoT

```
This is a natural language inference task. Given the dialogue context:
{context} Does {pragmatic turn} entails {implied meaning}? Reply
'entails' or 'not entails'.

Think step by step.
```

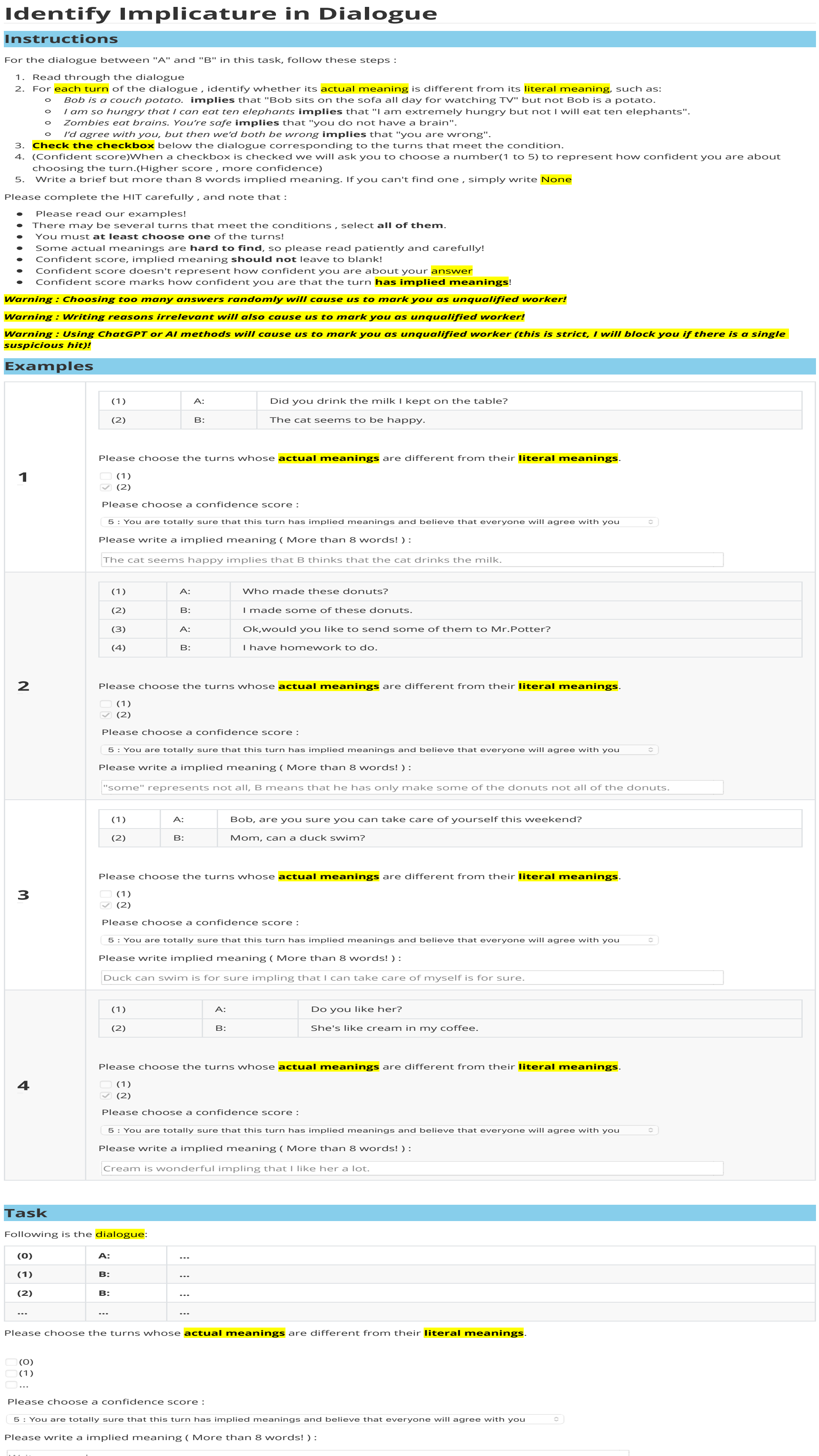
is out of reach for these models. In contrast, as shown in Table 10, ChatGPT is given the context, and the red line labeled "`Think step by step`" represents two distinct configurations: one with `step-by-step` and one without it.

D More Detail on DiPlomat

In this section, we will propose more examples of our dataset in Table 14, Table 15, Table 16, Table 17, and Table 18.

Table 14: Contextual reasoning examples of **DiPlomat**

A: Yeah. They say that he's the fastest pitcher there ever was. It's just he really couldn't find home plate. I mean, some of the stories you learn about this guy, it reads like fiction. When he was - I think this is around 1960. He's pitching in the minor leagues, and he pitched so fast he ripped the man's ear off. **B**: Oh. **A**: Yeah.	**Rationale**: The literal meaning is a simple expression of agreement, while the implied meaning is that the speaker is amazed by the story of Steve Dalkowski's feats.
B: We're talking about 2. 8 million people. Has the rise of temporary workers figured into, at least, the statistical improvement of the U. S. economy for some people? **A**: It has. Overall, about one seventh of the total job growth has been in the temp sector. The temp sector is growing nine times faster than the overall private sector as a whole. And the 2. 9 million workers represents a record number, both in the number of temp workers and in the percentage of the economy that they make up. **B**: You know in "Harvest Of Shame," Edward R. Murrow very famously said, the people we're showing you in this documentary have picked your Thanksgiving bounty with their bare hands, and this is how they live.	**Rationale**: The implied meaning of this turn is to reflect on our reliance on temporary workers in our day-to-day lives.
A: And so I got up and ran. And it wasn't too far. But I just - at that moment, I thought, I don't want to be shot in the back, and I need to find some cover. And there's really no place to hide. But there are these **B**: You found a little, like, alcove that you could duck into. **A**: There was a little alcove, yeah. And I just made myself as small as I could in that little corner.	**Rationale**: The speaker tried to protect itself from danger.
A: Well, there's a big argument in the United States about this. There's one group of folks who think that engagement policy failed. We engaged with China from 1979 until about 2013 when Xi Jinping came into power. And the idea of engagement was that coevolution was in the American interest as well as in China's interest. And you could bring China along to be a responsible player to some degree. **A**: Many hardliners in the United States government - and outside and including in the expert community - now claim that engagement was a sucker's game and that we have raised up a tiger which could now devour us. But there are different schools of thought about this, and many of us think that we still need to engage with China, albeit more strategically. **B**: That image of raising a tiger that will devour us is very dramatic.	**Rationale**: The situation is not necessarily an 'either/or' between China and the United States.



Identify Implicature in Dialogue

Instructions

For the dialogue between "A" and "B" in this task, follow these steps :

1. Read through the dialogue
2. For each turn of the dialogue , identify whether its actual meaning is different from its literal meaning, such as:
   - *Bob is a couch potato.* **implies** that "Bob sits on the sofa all day for watching TV" but not Bob is a potato.
   - *I am so hungry that I can eat ten elephants* **implies** that "I am extremely hungry but not I will eat ten elephants".
   - *Zombies eat brains. You're safe* **implies** that "you do not have a brain".
   - *I'd agree with you, but then we'd both be wrong* **implies** that "you are wrong".
3. **Check the checkbox** below the dialogue corresponding to the turns that meet the condition.
4. (Confident score)When a checkbox is checked we will ask you to choose a number(1 to 5) to represent how confident you are about choosing the turn.(Higher score , more confidence)
5. Write a brief but more than 8 words implied meaning. If you can't find one , simply write None

Please complete the HIT carefully , and note that :

- Please read our examples!
- There may be several turns that meet the conditions , select **all of them**.
- You must **at least choose one** of the turns!
- Some actual meanings are **hard to find**, so please read patiently and carefully!
- Confident score, implied meaning **should not** leave to blank!
- Confident score doesn't represent how confident you are about your answer
- Confident score marks how confident you are that the turn **has implied meanings**!

Warning : Choosing too many answers randomly will cause us to mark you as unqualified worker!

Warning : Writing reasons irrelevant will also cause us to mark you as unqualified worker!

Warning : Using ChatGPT or AI methods will cause us to mark you as unqualified worker (this is strict, I will block you if there is a single suspicious hit)!

Examples

1

(1)	A:	Did you drink the milk I kept on the table?
(2)	B:	The cat seems to be happy.

Please choose the turns whose **actual meanings** are different from their **literal meanings**.

- [ ] (1)
- [x] (2)

Please choose a confidence score :

5 : You are totally sure that this turn has implied meanings and believe that everyone will agree with you

Please write a implied meaning (More than 8 words!) :

The cat seems happy implies that B thinks that the cat drinks the milk.

2

(1)	A:	Who made these donuts?
(2)	B:	I made some of these donuts.
(3)	A:	Ok,would you like to send some of them to Mr.Potter?
(4)	B:	I have homework to do.

Please choose the turns whose **actual meanings** are different from their **literal meanings**.

- [ ] (1)
- [x] (2)

Please choose a confidence score :

5 : You are totally sure that this turn has implied meanings and believe that everyone will agree with you

Please write a implied meaning (More than 8 words!) :

"some" represents not all, B means that he has only make some of the donuts not all of the donuts.

3

(1)	A:	Bob, are you sure you can take care of yourself this weekend?
(2)	B:	Mom, can a duck swim?

Please choose the turns whose **actual meanings** are different from their **literal meanings**.

- [ ] (1)
- [x] (2)

Please choose a confidence score :

5 : You are totally sure that this turn has implied meanings and believe that everyone will agree with you

Please write implied meaning (More than 8 words!) :

Duck can swim is for sure impling that I can take care of myself is for sure.

4

(1)	A:	Do you like her?
(2)	B:	She's like cream in my coffee.

Please choose the turns whose **actual meanings** are different from their **literal meanings**.

- [ ] (1)
- [x] (2)

Please choose a confidence score :

5 : You are totally sure that this turn has implied meanings and believe that everyone will agree with you

Please write a implied meaning (More than 8 words!) :

Cream is wonderful impling that I like her a lot.

Task

Following is the dialogue:

(0)	**A:**	...
(1)	**B:**	...
(2)	**B:**	...
...	...	...

Please choose the turns whose **actual meanings** are different from their **literal meanings**.

- [ ] (0)
- [ ] (1)
- [ ] ...

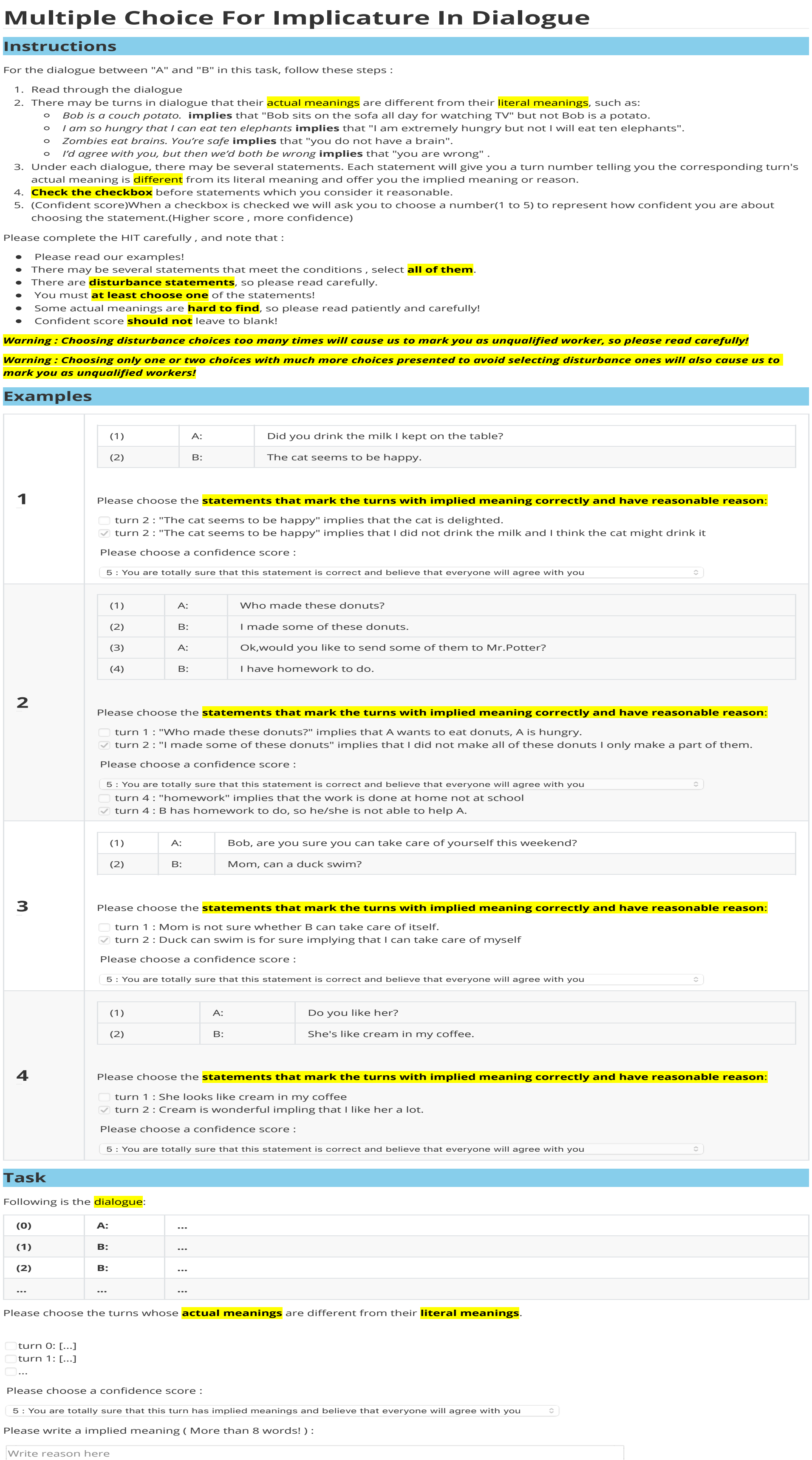
Please choose a confidence score :

5 : You are totally sure that this turn has implied meanings and believe that everyone will agree with you

Please write a implied meaning (More than 8 words!) :

Write reason here

Figure 7: Fine-grained annotation worker interface



Multiple Choice For Implicature In Dialogue

Instructions

For the dialogue between "A" and "B" in this task, follow these steps :

1. Read through the dialogue
2. There may be turns in dialogue that their actual meanings are different from their literal meanings, such as:
   - *Bob is a couch potato.* **implies** that "Bob sits on the sofa all day for watching TV" but not Bob is a potato.
   - *I am so hungry that I can eat ten elephants* **implies** that "I am extremely hungry but not I will eat ten elephants".
   - *Zombies eat brains. You're safe* **implies** that "you do not have a brain".
   - *I'd agree with you, but then we'd both be wrong* **implies** that "you are wrong" .
3. Under each dialogue, there may be several statements. Each statement will give you a turn number telling you the corresponding turn's actual meaning is different from its literal meaning and offer you the implied meaning or reason.
4. **Check the checkbox** before statements which you consider it reasonable.
5. (Confident score)When a checkbox is checked we will ask you to choose a number(1 to 5) to represent how confident you are about choosing the statement.(Higher score , more confidence)

Please complete the HIT carefully , and note that :

- Please read our examples!
- There may be several statements that meet the conditions , select **all of them**.
- There are **disturbance statements**, so please read carefully.
- You must **at least choose one** of the statements!
- Some actual meanings are **hard to find**, so please read patiently and carefully!
- Confident score **should not** leave to blank!

Warning : Choosing disturbance choices too many times will cause us to mark you as unqualified worker, so please read carefully!

Warning : Choosing only one or two choices with much more choices presented to avoid selecting disturbance ones will also cause us to mark you as unqualified workers!

Examples

1

(1)	A:	Did you drink the milk I kept on the table?
(2)	B:	The cat seems to be happy.

Please choose the **statements that mark the turns with implied meaning correctly and have reasonable reason:**

- [ ] turn 2 : "The cat seems to be happy" implies that the cat is delighted.
- [x] turn 2 : "The cat seems to be happy" implies that I did not drink the milk and I think the cat might drink it

Please choose a confidence score :

5 : You are totally sure that this statement is correct and believe that everyone will agree with you

2

(1)	A:	Who made these donuts?
(2)	B:	I made some of these donuts.
(3)	A:	Ok,would you like to send some of them to Mr.Potter?
(4)	B:	I have homework to do.

Please choose the **statements that mark the turns with implied meaning correctly and have reasonable reason:**

- [ ] turn 1 : "Who made these donuts?" implies that A wants to eat donuts, A is hungry.
- [x] turn 2 : "I made some of these donuts" implies that I did not make all of these donuts I only make a part of them.

Please choose a confidence score :

5 : You are totally sure that this statement is correct and believe that everyone will agree with you

- [ ] turn 4 : "homework" implies that the work is done at home not at school
- [x] turn 4 : B has homework to do, so he/she is not able to help A.

3

(1)	A:	Bob, are you sure you can take care of yourself this weekend?
(2)	B:	Mom, can a duck swim?

Please choose the **statements that mark the turns with implied meaning correctly and have reasonable reason:**

- [ ] turn 1 : Mom is not sure whether B can take care of itself.
- [x] turn 2 : Duck can swim is for sure implying that I can take care of myself

Please choose a confidence score :

5 : You are totally sure that this statement is correct and believe that everyone will agree with you

4

(1)	A:	Do you like her?
(2)	B:	She's like cream in my coffee.

Please choose the **statements that mark the turns with implied meaning correctly and have reasonable reason:**

- [ ] turn 1 : She looks like cream in my coffee
- [x] turn 2 : Cream is wonderful impling that I like her a lot.

Please choose a confidence score :

5 : You are totally sure that this statement is correct and believe that everyone will agree with you

Task

Following is the dialogue:

(0)	**A:**	...
(1)	**B:**	...
(2)	**B:**	...
...	...	...

Please choose the turns whose **actual meanings** are different from their **literal meanings**.

- [ ] turn 0: [...]
- [ ] turn 1: [...]
- [ ] ...

Please choose a confidence score :

5 : You are totally sure that this turn has implied meanings and believe that everyone will agree with you

Please write a implied meaning (More than 8 words!) :

Write reason here

Figure 8: Human refinements worker interface

Dialogue Question Answering

Instructions

For the dialogue between "A" and "B" in this task, follow these steps:

1. Read our example!
2. Read through the dialogue
3. Under each dialogue, there may be several questions.
4. There is a useful reference statement that may help you answer the corresponding question, please read it carefully
5. Answer each question with only one word!

Please complete the HIT carefully , and note that :

i. Reference Statements are useful, but don't just rely on it, read the dialogue as well!
ii. Our turn number start from 0
iii. There are incorrect reference statements.

Warning : If your answer is close to incorrect reference statement, but far away from the dialogue, we will mark you as unqualified workers!

Examples

Following is the dialogue:

(0)	B:	How did you feel when you found out, this summer, about the abuse of children that was going on in Pennsylvania not far from places you knew?
(1)	A:	Yeah, it hurt reading the report because I was reading about these parishes that I went to growing up.I was born in '96, so growing up, you would hear kids joke about, oh, you know, priests molesting kids and whatnot.But I never knew that there was actually a - anything behind that.I just sort of thought it was people making fun of a religion.And then I didn't learn, honestly, until recently that the abuse scandal is something that was real and something that the world has known about since - what?- I think it was the early 2000s, whenever the Boston Globe or whatever that newspaper is broke the story.I didn't know that that was a thing.It wasn't something I'd ever been exposed to, so I wasn't really aware of the fact that this is a problem that was going on in my church.You know, it's unthinkable.
(2)	B:	And it's hurt your faith?
(3)	A:	I think that's a difficult thing to answer.I will say that it has hurt my faith in the Catholic Church.I don't think it has actually hurt my personal, religious faith.I'm just starting to see less of a connection between what I believe and the teachings of the Catholic Church.

Reference Statement	turn 3 : The speaker's faith in the Catholic Church has been damaged by the abuse scandal, but their personal faith remains intact.
Question	How has the abuse scandal affected the speaker's faith in the Catholic Church?
Answer	Damage.

Task

Following is the dialogue:

(0)	**A:**	**...**
(1)	**B:**	**...**
(2)	**B:**	**...**
...	**...**	**...**

Reference Statement	**...**
Question	**...**

Answer	**[...] Write one word answer here.**

Figure 9: Answer collecting worker interface

Table 15: Figurative language reasoning examples of **DiPlomat**

A: Thank you. How are you? B: I'm pretty good. Thank you. You must be stuck like glue on this, but, you know, you've played in three World Cups, including one of the wins for the U. S. team in 1999. How would you describe what it's like to be out there on that field in that final game?	**Rationale**: Stuck like glue means to be attached to something, which is a particular issue or a person.
B: So in terms of what to do about it, we've said Twitter and Facebook have shut down these accounts, which prompts me to wonder - does shutting down a fake account do that much? Can't the Chinese government, if it's determined to go down this path, just open up two new ones in place of the one that was closed? **A**: It is a cat-and-mouse game, and the companies are constantly trying to get ahead of it. [$\cdots$] As you said, they can always set up new accounts.	**Rationale**: Mice are constantly trying to get away from cats and cats are constantly trying to catch mice. In the same way, the Chinese government will always be trying to escape restrictions on social media accounts and media companies will always be trying to find fake accounts.
A: I really didn't feel safe because the Turkish government is very famous for hunting down those who oppose Erdogan. So, I mean, I just didn't want to really risk my life by going to Europe. But, you know, I talked to my team. I told them all, like, how many times I want to come because I want to be with you guys there, and I want to get a win with you guys. And then, later on, they came back with the news and said, you know what?I think the best decision is if you don't come. Let's just not risk it for one game. **B**: Do you feel safe in New York and elsewhere in the U. S. ? **A**: I have been getting last two, three days hundreds death threats, but I think I feel safe in America. But anywhere else in the world, I wouldn't really feel safe.	**Rationale**: He is implying that he is still not safe.

Table 16: Commonsense reasoning examples of **DiPlomat**

B: Yeah - African-American mayor from Tallahassee. **A**: Yes. So this is sort of a test of whether real progressive candidates can win in these sort of purplish states. [...]	**Rationale**: "Purplish" states are not really colored. They refer to US states that are neither clearly Republican (red) nor Democrat (blue) in their voting.
B: He wrote a lot of letters by hand, didn't he? **A**: He wrote tons of letters. I bet there are a hundred thousand - hundreds out there[$\cdots$]	**Rationale**: tons of letters implies a very large number and not to full a ton.
B: Well, Pluto's official designation is a dwarf planet. And I have to tell you the people who sent this probe all the way out to Pluto are a little angry about that because when they launched it a decade ago, Pluto was still a planet. **A**: (Laughter) **B**: It got downgraded in the intervening years. **A**: That seems so unfair.	**Rationale**: A is expressing sympathy for the people who sent the probe, showing that they understand why they feel so disappointed.

Table 17: External knowledge reasoning examples of **DiPlomat**

B: Inside of his house, family pictures decorate the walls and the fridge. Les has 15 great grandchildren. He grew up in an orphanage, and he couldn't wait to leave to join the military. And so in early 1944, he boarded a ship and crossed the Atlantic Ocean to go to the frontline. **A**: I loved that sailing on, of course. It was so dramatic. You could see all these ships bobbing up and down on the ocean. And destroyers were weaving in and out of them to make sure they uncovered any mines or anything.	**Rationale**: Sailing across the ocean during wartime was a perilous experience.
A: . . . equivalent to a nuclear bomb? **B**: Well, it's about - its equivalent - the energy in that explosion is about 10 times the energy in the first atomic bomb. . .	**Rationale**: The energy released in the explosion is incredibly powerful.
B: So in your polling, in your research, do you find that it's going to come down to maybe a couple thousand votes from these unaffiliated voters and on what issues?Or will they vote? **A**: It is likely at the moment to be a very narrow victory. President Bush won in 2004 with five percent. That was 100,000 votes. In other words, if it is one percent, that would be 20,000 votes, and right now, the polls are moving around in just single percentage points. So it could be that narrow. **B**: Now, I have read that Colorado is going to be this year's Florida and Ohio, that this is going to be the state that decides the election. **A**: I think it could be, and the interesting thing is that Obama and Palin were both in Jefferson County a couple of days ago, indicating that there may be actually even a county that could be looked at to be beyond an entire state.	**Rationale**: The turn is suggesting that the county of Jefferson in Colorado could be a key factor in deciding the election, despite the fact that it is only one of many counties in the state and there are other swing states in the election.

Table 18: Others examples of **DiPlomat**

A: There's that feeling - I mean, so many of us have parents in the industry. I mean, that's what this region is about, especially around Detroit, and Wayne State's in Detroit, the heart of Detroit. So, it's nerve-racking. Everyone is nervous. Everyone doesn't know what's going to happen next. We're all watching the news very closely. But at the same time, it's interesting, because with my generation, we almost seem to, kind of, not be as directly impacted. I mean, our family is, it puts stress on us, but the day to day of the university and the day to day at school doesn't seem to have changed that much. **B**: I understand you have friends there who are engineering majors. Do they have any sense of what their future looks like, and will be it there in Michigan? **A**: Everybody is secure in their choices and secure in their decision. Everybody thinks that the industry will come around, especially now with the news that GM is getting money from the government. And everybody is more hopeful, and I mean, the auto industry has always been one of the largest industries and a staple in America, and to think that that industry is just going to vanish, nobody is willing to concede that.	**Rationale**: A believes that the auto industry will not vanish despite the current situation
B: In the meantime, what more have you learned in your reporting about the death of Carlos Hernandez Vasquez? **A**: Well, a couple of things. One thing that really stands out is that Carlos Hernandez Vasquez died in a Border Patrol station. The previous migrant children who died were taken to the hospital first; Hernandez Vasquez was not even though immigration authorities clearly knew that he was sick. He was diagnosed with the flu by a nurse practitioner.	**Rationale**: The death of Carlos Hernandez Vasquez could have been prevented if he had been taken to the hospital.
B: So, how do you and the retired general, James Jones, know each other? **A**: My gosh, I think - I can't even remember when I first met him. It's been so long ago. I'm sure I met him when he was head of the legislative liaison over the Senate. But I really became acquainted with him when he became a brigadier general, and, of course, I followed his career. Of course, he served very ably as a commandant in the marine corps and then as the European commander, just been with him from time to time. And I just consider him a very good friend.	**Rationale**: A has a high opinion of James Jones' character and career.

E Grice Maxims and Pragmatic Reasoning

The Gricean maxims have garnered substantial attention as a foundational theory within the domain of pragmatics. This theoretical framework comprises four distinct maxims: (1) The Maxim of Quality, (2) The Maxim of Quantity, (3) The Maxim of Relevance, and (4) The Maxim of Manner [16, 15]. In contrast to rigid rules or theorems, the Gricean maxims, which capture the prevalent dynamics of conversations, are susceptible to frequent breaches in the context of human communication. These breaches, stemming from the intricacies of real-world interaction, notably manifest in the violation of one or more of these maxims. Such breaches, aligned with the cooperative principle, give rise to pragmatic phenomena that necessitate the engagement of pragmatic reasoning by recipients of the communication [15].

F Computational Resources

For our experiment, we utilized two A100s and one 3090. The majority of our experiments were conducted on the A100s, while for practical reasons, only Unified-QA-base, BART-base, and T5-small were tested on the 3090. It is important to mention that each experiment was run on a single GPU. We record the training time of models in Appendix F.

Model	**C → P**	**CP → R**	**Device**
$\text{BERT}_{\text{base}}$	0.8min/epoch	0.9min/epoch	A100
$\text{RoBERTa}_{\text{base}}$	0.8min/epoch	0.9min/epoch	A100
$\text{RoBERTa}_{\text{large}}$	2.5min/epoch	2.8min/epoch	A100
$\text{GPT-2}_{\text{base}}$	5.8min/epoch	6.2min/epoch	A100
$\text{DialoGPT}_{\text{medium}}$	2.4min/epoch	4.2min/epoch	A100
$\text{DeBERTa}_{\text{base}}$	0.9min/epoch	0.9min/epoch	A100
$\text{ALBERT}_{\text{base}}$	0.5min/epoch	0.8min/epoch	A100

Table 19: Training Time of Models

G Limitations & Negative Societal Impacts

We acknowledge two limitations in our study: bias and subjectivity. Since our dialogues primarily stem from an interview dataset, a considerable focus is placed on political topics. This is reasonable, as pragmatic phenomena frequently emerge in the statements of politicians to advance their specific goals. However, this focus introduces a certain degree of bias into our dataset. The second limitation relates to the absence of subjectivity. In our methodology, the data undergoes two stages of human annotation, ensuring higher quality and objectivity. However, pragmatic reasoning is inherently subjective, and prioritizing objectivity compromises the preservation of subjectivity, resulting in a limitation in terms of subjectivity coverage. Our dataset exhibits minimal negative societal impacts. This is primarily due to the fact that our dialogues are transcriptions of publicly available TV shows, which inherently limits the potential for negative effects.

H Ethics Concern

Were any ethical review processes conducted (e.g., by an institutional review board)? No official processes were done, as our research is not on human subjects, but our data comes from published dataset.

Does the dataset contain data that might be considered confidential? No, our data comes from an existing public interview dataset.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. Few of the dialogues may talk about offensive topics. **Does the dataset identify subpopulations (e.g., by age or gender)?** Not explicitly.

Is it possible to identify individuals (i.e., one or more natural persons) directly or indirectly (i.e., in combination with other data) from the dataset? Yes, our data contains names of famous people.

I Responsibility & Dataset Liscence

We bear all responsibility in case of violation of rights and our dataset is under the license of CC BY-NC-SA (Attribution-NonCommercial-ShareAlike).

J Datasheets for Our Dataset

J.1 Motivation

1. For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)

   This dataset was created to study pragmatic reasoning in dialogues, a specific gap is mentioned above in Appendix G.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

   This dataset was created by the authors of this paper.

3. Who funded the creation of the dataset? (If there is an associated grant, please provide the name of the grantor and the grant name and number.)

   The institute of the authors funded the creation of the dataset.

4. Any other comments?

   None.

J.2 Composition

5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)

   An instance of our dataset represent a piece of dialogue. Description is provided in our paper.

6. How many instances are there in total (of each type, if appropriate)?

   We answer the question in our paper. Our datasets owns 4,177 dialogues.

7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).)

   It is a sample of all possible cases. As pragmatic phenomena aren't proved to be limited, we can't guarantee a full sampling of them.

8. What data does each instance consist of?

   We mention it in our paper.

9. Is there a label or target associated with each instance? If so, please provide a description.

   Yes. The description is in our paper.

10. Is any information missing from individual instances? (If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)

    No. We leverage the original dialogues.

11. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? (If so, please describe how these relationships are made explicit.)

    No. Instances are weakly related, but focus on the same phenomenon.

12. Are there recommended data splits (e.g., training, development/validation, testing)? (If so, please provide a description of these splits, explaining the rationale behind them.)

    Yes. We provide it.

13. Are there any errors, sources of noise, or redundancies in the dataset? (If so, please provide a description.)

    Yes. Some workers try to finish the work as quickly as possible, therefore when we ask them to offer a rationale for choosing a certain turn as a pragmatic turn, they simply type an "a" in the box. However, the situation is rare, and we blocked the workers and clean the data out of our dataset.

14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)

    It's self-contained.

15. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? (If so, please provide a description.)

    No.

16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? (If so, please describe why.)

    Yes. Some of the topic are big events, they may be offensive for some people. However, we consider our dataset's offensiveness to be limited, for the source dataset is a TV show transcript.

17. Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)

    Yes.

18. Does the dataset identify any subpopulations (e.g., by age, gender)? (If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)

    No. This is not explicitly identified

19. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? (If so, please describe how.)

    Yes; their names are given in running text.

20. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? (If so, please provide a description.)

Yes. Our dataset may have dialogues talking about religious, politics and so on.

21. Any other comments?

    None.

J.3 Collection Process

1. How was the data associated with each instance acquired? (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)

   The data all comes from an interview dataset already published. (See our paper)

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? (How were these mechanisms or procedures validated?)

   Software program and manual human curation (2 times). See our paper for details.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

   Randomly.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

   Crowdworkers. They are paid nicely. See Appendix for detail.

5. Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)

   The dataset was collected in the early Spring of 2023, which does not necessarily reflect the timeframe of the data collected.

6. Were any ethical review processes conducted (e.g., by an institutional review board)? (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)

   No review processes were conducted with respect to the collection and annotation of this data (though review was done for other aspects of this work; see the paper linked at the top of the datasheet).

7. Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)

   Yes.

8. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

   Other sources. By curating a published dataset.

9. Were the individuals in question notified about the data collection? (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)

   No.

10. Did the individuals in question consent to the collection and use of their data? (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)

    No. All data are public.

11. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)

    N/A.

12. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)

    No. We consider our dataset having a limited negative effect, for all of our data has been published for more than a year.

13. Any other comments? None.

J.4 Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)

   No.

J.5 Uses

1. Has the dataset been used for any tasks already? (If so, please provide a description.)

   Yes. See our paper for details.

2. Is there a repository that links to any or all papers or systems that use the dataset? (If so, please provide a link or other access point.)

   No.

3. What (other) tasks could the dataset be used for?

   Many more. Such as generation of implied meanings.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)

   No.

5. Are there tasks for which the dataset should not be used? (If so, please provide a description.)

   No.

6. Any other comments?

   None.

J.6 Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? (If so, please provide a description.)

   Yes, the dataset is freely available.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? (Does the dataset have a digital object identifier (DOI)?)

   On our website.

3. When will the dataset be distributed?

   It's already been distributed.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)

   The dataset is licensed under a CC license.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)

   Not to our knowledge.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)

   Not to our knowledge.

7. Any other comments?

   None.

J.7 Maintenance

1. Who is supporting/hosting/maintaining the dataset?

   The authors.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

   We will post our email address.

3. Is there an erratum? (If so, please provide a link or other access point.)

   Currently, no. As errors are encountered, future versions of the dataset may be released (but will be versioned). They will all be provided in the same location.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')? (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)

   Yes.However, the frequency isn't determined, and we'll publish the updated dataset on the same website if an renewal occurs, and we'll anounce it on the website.

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.)

   No.

6. Will older versions of the dataset continue to be supported/hosted/maintained? (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)

   Yes. The older versions of the dataset will be available on the website.

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)

   Yes. They can email us.

8. Any other comments?

   None.

References

[1] Jens Ambrasat, Christian von Scheve, Markus Conrad, Gesche Schauenburg, and Tobias Schröder. Consensus and stratification in the affective meaning of human sociality. *Proceedings of the National Academy of Sciences*, 111(22):8001–8006, 2014. 2

[2] Alan P Fiske. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, 99(4):689, 1992. 2

[3] John A Bargh and Tanya L Chartrand. The unbearable automaticity of being. *American psychologist*, 54 (7):462, 1999. 2

[4] Edward Finegan. *Language: Its structure and use*. Cengage Learning, 2014. 2

[5] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022. 2

[6] Silke Anders, Roos de Jong, Christian Beck, John-Dylan Haynes, and Thomas Ethofer. A neural link between affective understanding and interpersonal attraction. *Proceedings of the National Academy of Sciences*, 113(16):E2248–E2257, 2016. 2

[7] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 2

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 2

[10] Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*, 2022. 2

[11] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. We're afraid language models aren't modeling ambiguity. 2023. 2

[12] Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2074–2085, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.182. URL https://aclanthology.org/2021.findings-acl.182. 2, 3, 4

[13] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 9

[14] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022. 2

[15] Alan Cruse. Meaning in language: An introduction to semantics and pragmatics. 2, 4, 22

[16] Herbert P Grice. Logic and conversation. 1975. 2, 4, 22

[17] Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. Mermaid: Metaphor generation with symbolism and discriminative decoding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. 2, 4

[18] Prateek Saxena and Soma Paul. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, 2020. 2, 4, 5, 12

[19] Tosin P. Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Liwicki, and Marcus Liwicki. Potential idiomatic expression (pie)-english: Corpus for classes of idioms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022. 2, 4

[20] Issa Annamoradnejad and Gohar Zoghi. Colbert: Using bert sentence embedding in parallel neural networks for computational humor, 2022. 2, 4, 12

[21] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.653. URL `https://aclanthology.org/2020.emnlp-main.653`. 2, 5

[22] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4640. URL `https://aclanthology.org/W15-4640`. 3

[23] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 3

[24] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. 3

[25] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018. 3, 10

[26] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 3

[27] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 3

[28] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 3

[29] Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge. In *Transactions of the Association for Computational Linguistics (TACL)*, 2019. 3

[30] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. Dream: A challenge dataset and models for dialogue-based reading comprehension. In *Transactions of the Association for Computational Linguistics (TACL)*, 2019. 3

[31] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 3

[32] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 3

[33] Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. Mutual: A dataset for multi-turn dialogue reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 3, 4

[34] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models imppressive? learning implicature and presupposition. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 3

[35] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 3, 14

[36] Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*, 2022. 3

[37] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 3

[38] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020. 3

[39] Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. Overview of the ninth dialog system technology challenge: Dstc9, 2020. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[41] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023. 4

[42] Yunxiang Zhang and Xiaojun Wan. MOVER: Mask, over-generate and rank for hyperbole generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6018–6030, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.440. URL https://aclanthology.org/2022.naacl-main.440. 4, 12

[43] Orion Weller and Kevin Seppi. The rJokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.753. 4

[44] Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336 (6084):998–998, 2012. doi: 10.1126/science.1218633. URL https://www.science.org/doi/abs/10.1126/science.1218633. 4

[45] Kevin Stowe, Prasetya Utama, and Iryna Gurevych. IMPLI: Investigating NLI models' performance on figurative language. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5375–5388, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.369. URL https://aclanthology.org/2022.acl-long.369. 4

[46] Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. It's not rocket science : Interpreting figurative language in narratives. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 4

[47] Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Cicero: A dataset for contextualized commonsense inference in dialogues. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 4

[48] Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Cider: Commonsense inference for dialogue explanation and reasoning. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021.

[49] Pei Zhou, Pegah Jandaghi, Bill Yuchen Lin, Justin Cho, Jay Pujara, and Xiang Ren. Probing commonsense explanation in dialogue response generation. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 4

[50] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019. URL http://arxiv.org/abs/1907.11692. 5, 12, 13

[51] Yunxiang Zhang and Xiaojun Wan. Mover: Mask, over-generate and rank for hyperbole generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 5

[52] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 5, 7, 12, 13

[53] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[54] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084. 7

[55] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 9

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research (JMLR)*, 2020. 9, 10

[57] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 9

[58] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41. 9

[59] Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In *International Conference on Computational Linguistics (COLING)*, pages 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-1066. 9

[60] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287. 10

[61] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015. 10

[62] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations (ICLR)*, 2021. 10, 13

[63] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 10

[64] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 582–601, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.38. 10

[65] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. 2020. 12

[66] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*, 2020. 12, 13

[67] Rishabh Misra. News headlines dataset for sarcasm detection, 2022. 12

[68] Lennart Wachowiak, Dagmar Gromann, and Chao Xu. Drum up SUPPORT: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.flp-1.7. 12

[69] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, mar 1997. ISSN 0891-2017. 12

[70] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 13

[71] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 13