PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities

Anonymous ACL submission

Abstract

LLMs have demonstrated remarkable capability for understanding semantics, but their understanding of pragmatics is not well studied. To this end, we release a Pragmatics Understanding Benchmark (PUB)[1] dataset consisting of *fourteen* tasks in *four* pragmatics phenomena, namely, *Implicature, Presupposition, Reference, and Deixis*. We curate high-quality test sets for each task, consisting of Multiple Choice Question Answers (MCQA). PUB includes a total of $28k$ data points, $6.1k$ are newly annotated. We evaluate *nine* models varying in the number of parameters and type of training. Our study reveals several key observations about the pragmatic capabilities of LLMs: 1. chat-fine-tuning strongly benefits smaller models, 2. large base models are competitive with their chat-fine-tuned counterparts, 3. there is a huge variance in performance across different pragmatics phenomena, and 4. a noticeable performance gap between human capabilities and model capabilities. We hope that PUB will enable comprehensive evaluation of LLM's pragmatic reasoning capabilities.

1 Introduction

Pragmatics, within linguistics, examines how context shapes language understanding in communication (Grice, 1975). It centers on real-life language use, considering context, speaker intentions, presuppositions, and implied meanings to derive interpretations beyond literal words. Human's proficiency in pragmatics stems from their inherent cognitive skills and social awareness. Our minds adeptly process not only spoken words but also context and implied messages. In Natural Language Processing (NLP), Large Language Models (LLMs) (Brown et al., 2020; Scao et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023) have emerged as a transformative force in recent years. LLMs have shown remarkable abilities on many downstream tasks like Natural Language Understanding (Wang et al., 2019b; Williams et al., 2018), text generation (Paperno et al., 2016; Merity et al., 2016), code synthesis (Chen et al., 2021; Hendrycks et al., 2021), question answering (Mihaylov et al., 2018; Kwiatkowski et al., 2019; Rajpurkar et al., 2018) and reasoning (Wang et al., 2019a; Cobbe et al., 2021; Geva et al., 2021; Clark et al., 2018), etc. While semantics involves the study of words and their meanings in a language, *pragmatics* extends this inquiry by considering word's meanings within the context in which they are used.

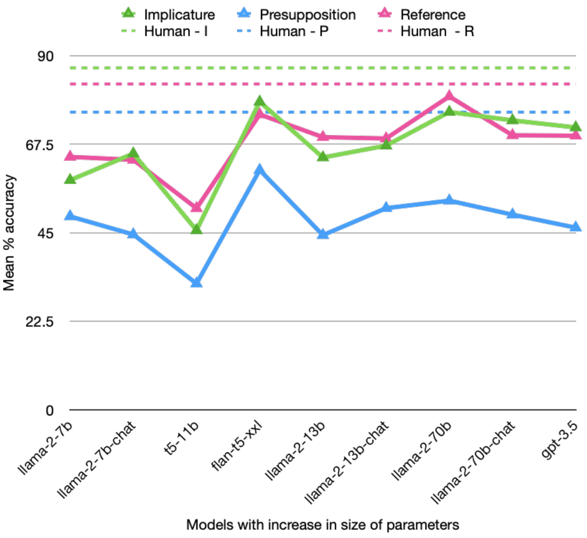


Figure 1: Average performance of models on three different pragmatics phenomena. Average accuracy for reference and deixis are merged and plotted as *Reference* as they are closely related phenomena. Human - I, P, R represent the performance of human evaluators on Implicature, Presupposition, and Reference respectively

Given LLMs' increased interaction with humans via practical, real-world applications like chatbots, search engines, and web browsers, the following

[1] The benchmark is available at `github.com/pub`

research question arises: *Do LLMs understand pragmatics in conversations?*

Most benchmarks until now deal only with abilities like problem-solving (Cobbe et al., 2021) or semantic understanding (Wang et al., 2019b; Srivastava et al., 2022) where LLMs have started approaching human performance. However, due to lack of reliable benchmarks, it is still unclear whether an LLM understands pragmatics or not. To facilitate this research, we propose a Pragmatic Understanding Benchmark (PUB) over four major pragmatic phenomena:
1. Implicature: understanding what is implied in a statement even though it is not literally expressed.
2. Presupposition: an implicit assumption that is taken for granted before the use of a statement.
3. Deixis: a phenomenon in which certain phrases within a sentence rely on contextual cues, such as the speaker, the listener, or the surrounding context, to convey their meaning effectively.
4. Reference: how language points to things, people, place, time, etc. in accordance with the content and structure outlined in the Handbook of Pragmatics (Horn and Ward, 2004).

PUB includes 22,000 examples, leveraging existing data, and introduces three new datasets with 6,100 newly annotated examples. Human evaluation of a subset of these datasets is conducted to assess performance against established LLMs. The benchmark comprises fourteen tasks that evaluate pragmatics as an MCQA task since MCQA evaluation is more closely related to question-answering abilities in conversations (Robinson and Wingate, 2023). We carefully curate the existing datasets to balance them and formulate prompts for these tasks, which are more natural and better suited to evaluate LLMs. Following ((Brown et al., 2020), (Robinson and Wingate, 2023)), we evaluate the pragmatic abilities of LLMs using Multiple Choice Prompting (MCP) and Cloze prompting (CP). To validate the model's confidence in its choices we also calculate the Proportion of Plurality Agreement (PPA) on 3 tasks similar to (Robinson and Wingate, 2023).

Our contributions are: (1) a comprehensive and unified dataset for 14 distinct tasks in pragmatics (Figure: 3), containing 28k data points; to the best of our knowledge this is the first dataset-linguistically motivated and well-grounded- to test pragmatic capabilities of LLMs. (2) a systematic evaluation of 6 variations of llama-2, t5, flan-t5, and GPT-3.5, on the *fourteen* mentioned tasks. (3) a study of human performance on a sample of the dataset to highlight the performance gap between LLMs and humans. (4) insight emerging from (3) to uncover strengths and weaknesses of LLMs vis-a-vis humans. These contribution points- we hope- will assist researchers in improving the interactive abilities of LLMs.

2 Related Work

Pragmatics is very crucial in the domain of linguistics, where it plays a critical role in understanding meaning (Allwood, 1981). In linguistic terms, pragmatics deals with the study of context-dependent aspects of meaning that are systematically abstracted away from, in the construction of content or logical form (Horn and Ward, 2004). Some of the basic subfields of pragmatics include *implicature*, *presupposition*, *speech acts*, *reference*, *deixis*, *definiteness*, and *indefiniteness*.

Over the years, many researchers have devoted their research to studying such pragmatic phenomena for machine learning. To study implicatures, Louis et al. (2020) employ indirect answers in polar questions, Zheng et al. (2021) utilize hierarchical grammar models for understanding implicature and deictic reference in simple conversations, Jeretic et al. (2020) employ Natural Language Inference (NLI) to grasp scalar implicatures, Deng et al. (2014) leverage implicature rules for optimizing sentiment detection, and Lahiri (2015) develop a sentence-level corpus with implicature ratings. Whereas for presupposition, Kim et al. (2022) use search engine queries that may contain questionable assumptions that are closely related to presupposition. Kabbara and Cheung (2022) also reveals that Transformer models exploit specific structural and lexical cues as opposed to performing some kind of pragmatic reasoning.

A recent comparison of pragmatic understanding between humans and models, conducted by Hu et al. (2023), shows that language models struggle to comprehend humor, irony, and conversational maxims (Grice, 1975). In the most recent work, Ruis et al. (2023) have studied implicature recovery in polar questions and answers. These approaches have offered only a restricted understanding of the shortcomings exhibited by these models by either evaluating only a single phenomenon or using a smaller number of samples to make it quantifiable. Other existing works (Deng et al., 2014; Sileo et al., 2022; Qi et al., 2023) do not comprehensively cover all important domains of pragmatics to evaluate LLMs. To the best of our knowledge, we are the first ones to combine major aspects of pragmatics to create a quantifiable benchmark.

3 Datasets and Tasks

We describe the datasets used and curated for creating PUB in section 3.1. Various tasks for evaluation of LLMs is introduced in section 3.2.

3.1 Datasets

With the help of language experts, we selected existing datasets covering important pragmatic aspects. Specifically, we select Circa (Louis et al., 2020), GRICE (Zheng et al., 2021), FigQA (Liu et al., 2022), FLUTE (Chakrabarty et al., 2022), IMPPRES (Jeretic et al., 2020), and NOPE (Parrish et al., 2021). We adapted datasets for various tasks (in MCQA format) with necessary changes and also made new ones where needed for specific purposes. Annotation details are discussed in Appendix A.

Overview of newly annotated datasets is discussed below:

1. **CircaPlus** is a newly annotated dataset containing 2.5k human written implied meanings based on the indirect responses present in Circa dataset (Louis et al., 2020).
2. **DialogAssumptions** is a new dataset containing 2.5k pairs of expert-annotated presuppositions based on a subset of dialogues from the Dailydialog dataset (Li et al., 2017). While current presupposition datasets are built around trigger words present in sentences, to our understanding, there hasn't been a resource addressing presuppositions in conversational contexts where trigger words are absent. Hence, we developed this dataset specifically to fill this gap.
3. **MetoQA** is a novel dataset comprising 1100 multiple-choice questions based on the linguistic phenomenon called metonymy. Metonymy is a figure of speech in which one word or phrase is substituted with another word or phrase with which it is closely associated or related. Unlike a metaphor, where one thing is said to be another (e.g., “Life is a journey"), in metonymy, the substitution is based on a real, often contiguously related, connection between the two terms (e.g., “These are my hired guns").

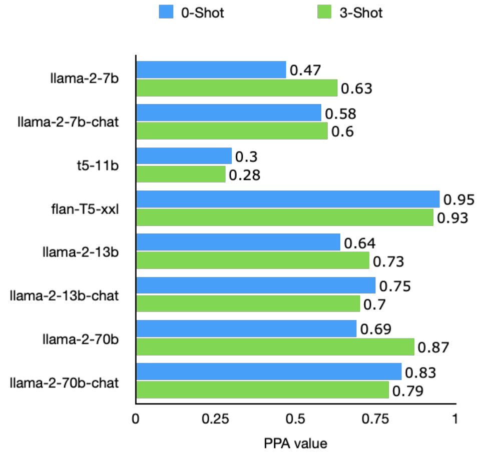


Figure 2: Comparison of Proportion of Plurality Agreement (PPA). Results are averaged across Task 4, 11, and 14, each representing a pragmatic domains. Vanilla LLMs show improved consistency with a few shots, while instruction-tuned models show no improvement.

3.2 Tasks

Each task incorporated within PUB is structured to evaluate distinct domains of pragmatics. Owing to the importance of Implicature in pragmatics, this benchmark includes a greater focus on Implicature, with ten tasks designed to thoroughly evaluate models' abilities. Presupposition and Reference are covered through two tasks each. Figure 3 contains details and illustrations for each task introduced in PUB. Additional description for all tasks is given in Appendix B.

4 Evaluation strategy

We have selected two evaluation methods namely length normalized cloze prompting (Brown et al., 2020) and Multiple Choice Prompting (MCP) (Robinson and Wingate, 2023). Since MCP is also dependent on the multiple choice symbol binding ability of LLMs, we have computed the Proportion of Plurality Agreement (PPA) (Robinson and Wingate, 2023) to ensure the model's consistency across possible orders of answer options. The models under investigation include flan-t5-xxl (Chung et al., 2022); llama-2 : 7b, 7b-chat, 13b, 13b-chat, 70b, 70b-chat (Touvron et al., 2023); t5 (Raffel et al., 2020) and GPT-3.5 (Brown et al., 2020)

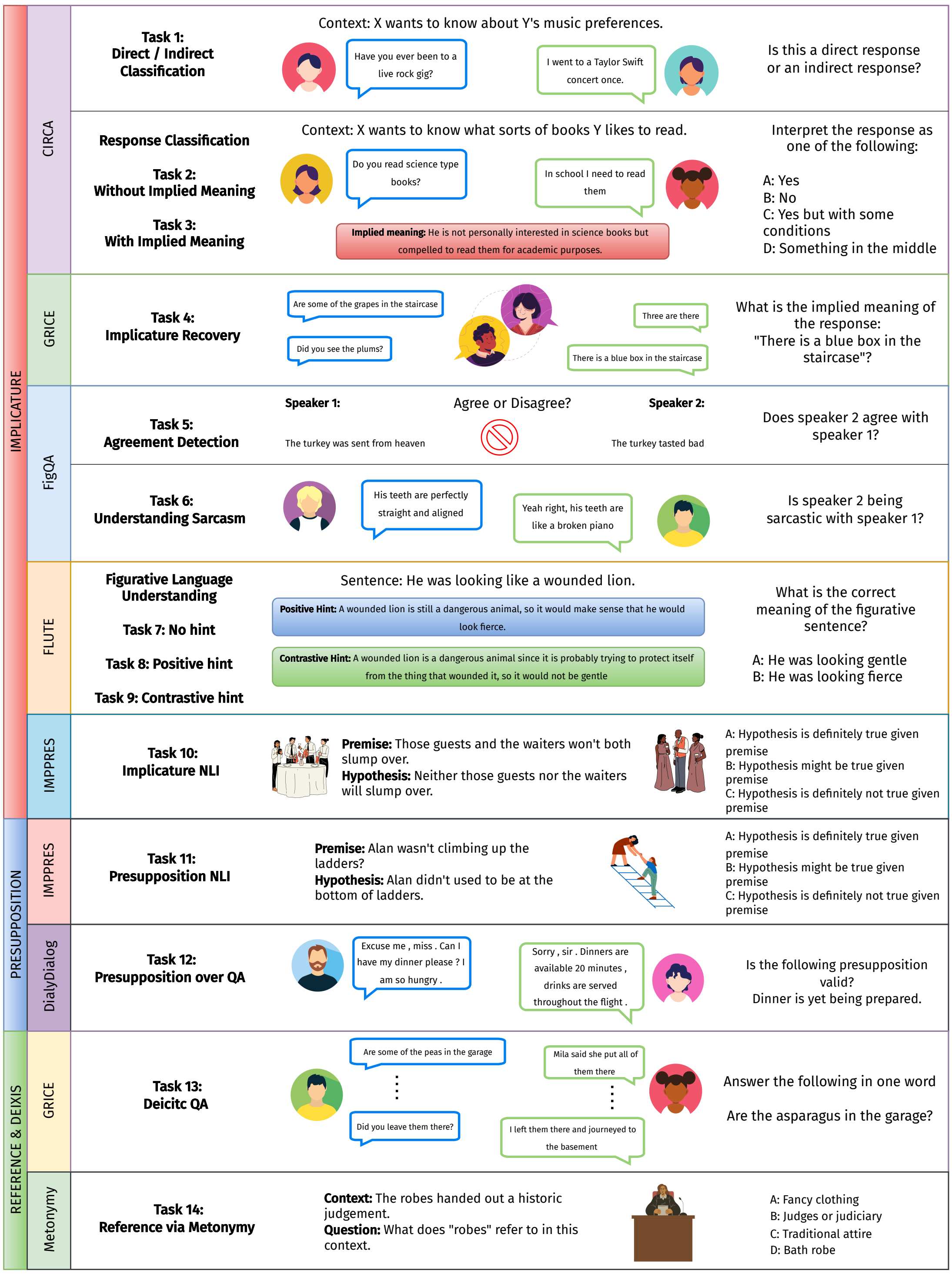

IMPLICATURE
PRESUPPOSITION
REFERENCE & DEIXIS
CIRCA
Task 1: Direct / Indirect Classification
Context: X wants to know about Y's music preferences.
Have you ever been to a live rock gig?
I went to a Taylor Swift concert once.
Is this a direct response or an indirect response?
Response Classification
Task 2: Without Implied Meaning
Task 3: With Implied Meaning
Context: X wants to know what sorts of books Y likes to read.
Do you read science type books?
In school I need to read them
Implied meaning: He is not personally interested in science books but compelled to read them for academic purposes.
Interpret the response as one of the following:
A: Yes
B: No
C: Yes but with some conditions
D: Something in the middle
GRICE
Task 4: Implicature Recovery
Are some of the grapes in the staircase
Did you see the plums?
Three are there
There is a blue box in the staircase
What is the implied meaning of the response: "There is a blue box in the staircase"?
FigQA
Task 5: Agreement Detection
Speaker 1:
Agree or Disagree?
Speaker 2:
The turkey was sent from heaven
The turkey tasted bad
Does speaker 2 agree with speaker 1?
Task 6: Understanding Sarcasm
His teeth are perfectly straight and aligned
Yeah right, his teeth are like a broken piano
Is speaker 2 being sarcastic with speaker 1?
FLUTE
Figurative Language Understanding
Task 7: No hint
Task 8: Positive hint
Task 9: Contrastive hint
Sentence: He was looking like a wounded lion.
Positive Hint: A wounded lion is still a dangerous animal, so it would make sense that he would look fierce.
Contrastive Hint: A wounded lion is a dangerous animal since it is probably trying to protect itself from the thing that wounded it, so it would not be gentle
What is the correct meaning of the figurative sentence?
A: He was looking gentle
B: He was looking fierce
IMPPRES
Task 10: Implicature NLI
Premise: Those guests and the waiters won't both slump over.
Hypothesis: Neither those guests nor the waiters will slump over.
A: Hypothesis is definitely true given premise
B: Hypothesis might be true given premise
C: Hypothesis is definitely not true given premise
IMPPRES
Task 11: Presupposition NLI
Premise: Alan wasn't climbing up the ladders?
Hypothesis: Alan didn't used to be at the bottom of ladders.
A: Hypothesis is definitely true given premise
B: Hypothesis might be true given premise
C: Hypothesis is definitely not true given premise
DialyDialog
Task 12: Presupposition over QA
Excuse me , miss . Can I have my dinner please ? I am so hungry .
Sorry , sir . Dinners are available 20 minutes , drinks are served throughout the flight .
Is the following presupposition valid?
Dinner is yet being prepared.
GRICE
Task 13: Deicitc QA
Are some of the peas in the garage
Did you leave them there?
Mila said she put all of them there
I left them there and journeyed to the basement
Answer the following in one word
Are the asparagus in the garage?
Metonymy
Task 14: Reference via Metonymy
Context: The robes handed out a historic judgement.
Question: What does "robes" refer to in this context.
A: Fancy clothing
B: Judges or judiciary
C: Traditional attire
D: Bath robe

Figure 3: Illustration of each task from PUB. The dataset used for each task is prepended to each row in the figure. Related tasks are grouped together. This is followed by the task name, an illustration and a prompt example. Verbal descriptions for these tasks is mentioned in Appendix B. Prompts used to evaluate LLMs are given in Appendix C.

4.1 Prompting LLMs

We do a *zero-shot* and a *3-shot evaluation* for each of the above mentioned strategies. The OpenAI model is evaluated only using MCP. For Zero-shot prompts, all the instances of the data were used as is. For Few-shot prompts, a dev set of 20 examples was created for each task. These 20 examples were selected to ensure a balanced representation of options. These examples were randomly selected from the entire dataset for tasks with unique options for each question. Three samples were randomly selected from this dev set for 3-shot eval-

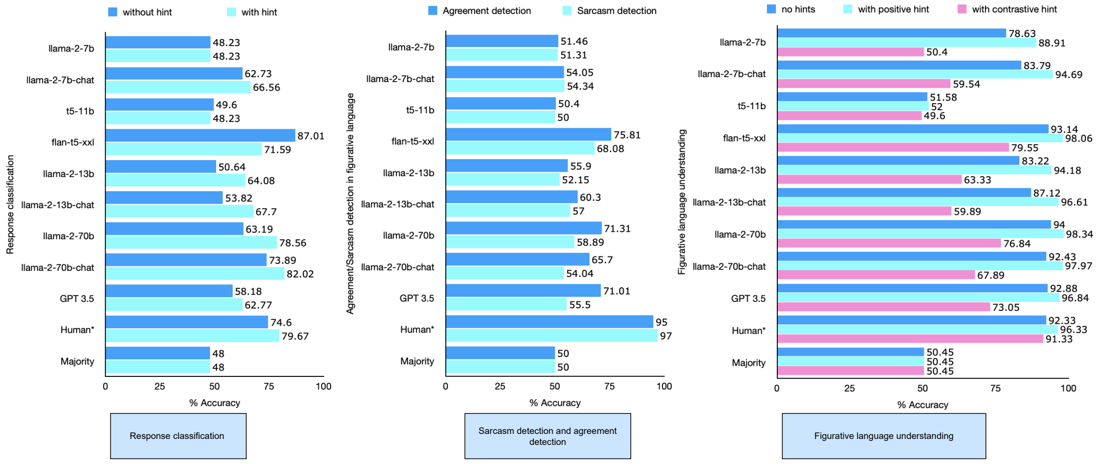


Figure 4: Results for tasks 2 & 3, tasks 5 & 6 and tasks 7, 8 & 9. The results presented in this table are the maximum across all types of evaluations (0-shot and 3-shot Cloze and MCQA) performed on the models.

uation. The remaining instances of the data, other than the dev set, were used to evaluate the model. Prompts for each task are given in Appendix C.

4.2 Human evaluation

To compare the performance of these LLMs with humans, we selected 100 examples from the complete evaluation set for each task. We employed three human evaluators to answer these 100 samples from each task, resulting in a total of 4,200 human evaluations. The evaluators are fluent English speakers and have graduated from a technical university where English is the medium of instruction. It is important to note that the human evaluation does not reflect expert human reference but rather the performance of a random human on complex pragmatic tasks. These evaluators are presented with the same prompt as the *0-shot* MCP presented to the LLMs.

5 Results and Analysis

The results of our experiments are presented in Figures 4, 5. Only the maximum across all evaluation strategies is reported in these figures. Detailed results are given in Appendix E. The results for PPA are presented in Figure 2. Based on these results, the following section aims to answer the questions regarding the pragmatic capabilities of LLMs.

5.1 Main Results

How much do LLMs understand what humans mean during conversations? To evaluate how well LLMs understand implied meanings of conversations, implicature and reference tasks offer pertinent insights. We observe that the models perform moderately in classifying a response as direct or indirect (Task 1). They also struggle to interpret the meaning of the indirect response (Task 2). However, *llama-70b-chat* is an exception to this trend. Similar to humans, a noticeable increase in performance is observed when a hint is provided for indirect response interpretation (Task 3). The performance of models is indifferent to both polar and non-polar question answers in resolving implicatures (Task 2 vs. Task 4). Even though NLI is an established task in NLP, it is observed that models perform poorly in making pragmatic inferences (Task 5). Finally, as in Figure 1, the average performance on implicature and reference is similar, suggesting that these models do not fully interpret human conversations.

Despite operating on the same dataset, do LLMs demonstrate varying task sensitivity? While it's known that LLMs are sensitive to the wording of prompts (Webson and Pavlick, 2021), this investigation aims to explore their task sensitivity. Specifically, we want to understand how altering the order of speakers, asking a different question or giving a different hint impacts the model's performance. Although derived from the same dataset, LLMs demonstrate stronger performance in agreement detection (Task 5) over sarcasm detection (Task 6) (on average, there is a 13% performance gap in models $\geq 13b$ parameters). The tasks designed on flute dataset (Chakrabarty et al., 2022) shed light on the model's susceptibility to distractions. We can observe that with a change in the hint from positive (Task 8) to contrastive (Task 9), there is a drastic decrease (on an average

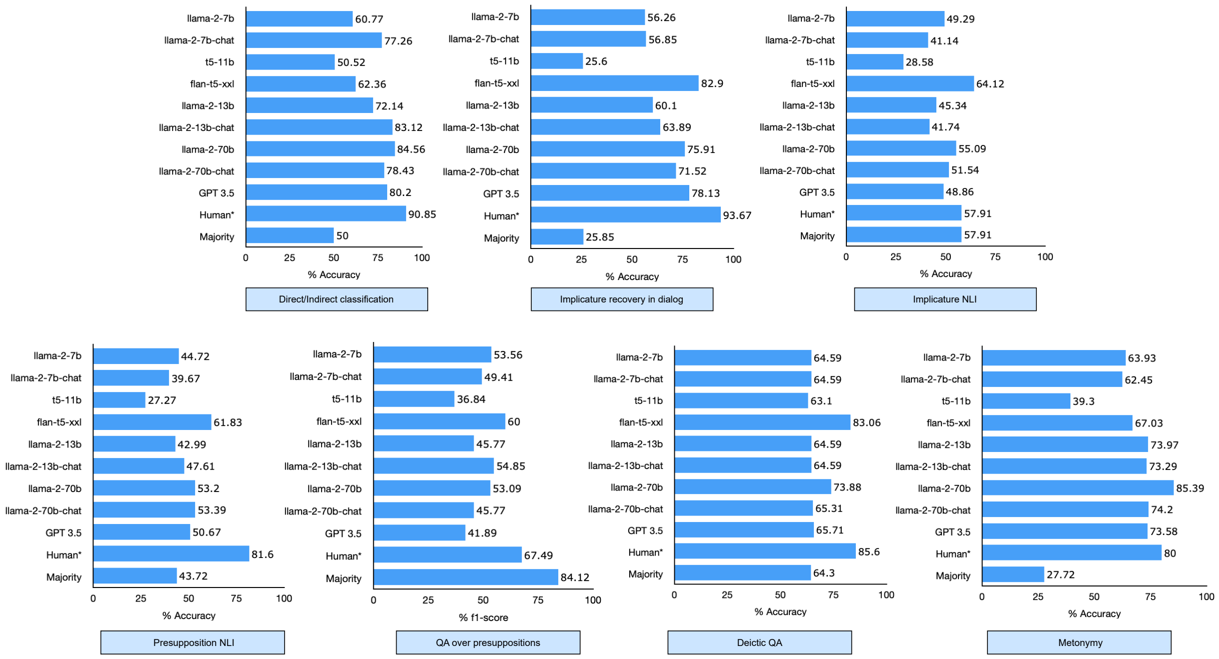


Figure 5: Results for tasks 1, 4, 10, 11, 12, 13 and 14. The results presented in this table are the maximum across all types of evaluations (0-shot and 3-shot Cloze and MCQA) performed on the models.

of 20%) in the accuracy levels. Interestingly, the inclusion of a positive hint, which has a higher lexical overlap with the correct answer, seems to boost the performance of the model. However, the model's performance decreases when a contrastive hint is introduced. This observed pattern brings into question the pragmatic abilities of these models, suggesting that their understanding and interpretation of language may be more significantly influenced by the presence and nature of linguistic cues than by inherent logic.

Does a Model's Scale Correlate with Its Pragmatic Abilities? The performance shown in Figure 1 hints at a possible correlation between a model's scale and its pragmatic capabilities. However, given the model's vulnerability to task sensitivity, even the largest models display perplexity, as previously discussed. Consequently, concluding that pragmatics is an emergent ability might be premature due to observed inconsistencies, even among models at the extremes of the scale.

Do LLMs that are optimized for dialogue use cases exhibit superior pragmatic abilities? From the experiments, it is evident that the chat-optimized variants of *llama* slightly outperform the base models on most of the tasks. There is a notable performance gap between models like *t5-11b* and *flan-t5-xxl*, with the instruction-tuned *flan-t5-xxl* model approaching near-human-level performance in many of the tasks. This suggests that instruction tuning can significantly enhance a model's ability to handle complex language tasks, bringing it closer to human-like language comprehension.

How do the pragmatic abilities of LLMs compare concerning world knowledge involvement? All implicature tasks (except Tasks 1 and 4) involve a certain degree of world knowledge. In Reference, while the metonymy task (Task 14) requires world knowledge, the Deixis task (Task 13) does not. The model's below-par performance is not primarily due to a lack of world knowledge. Instead, it appears to stem from a deficiency in their innate pragmatic abilities. This is evident because even in tasks not reliant on world knowledge, like Deixis, the model's performance isn't on par with tasks involving world knowledge. It suggests that the challenge lies more in the model's pragmatic processing than its knowledge base.

Do they understand the same implied meaning and make the same assumptions as humans? The models demonstrate relatively stronger performance in tasks related to implicature and reference, both of which involve inferred meanings

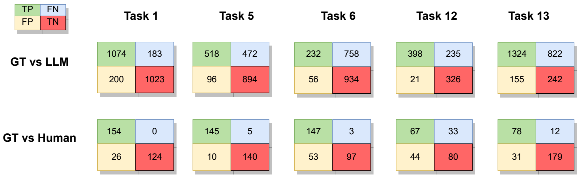
Task No.	GT-Human	Human-LLM
Task 1	0.829	0.749 (-0.08)
Task 2	0.681	0.421 (-0.26)
Task 3	0.754	0.550 (-0.20)
Task 5	0.901	0.515 (-0.39)
Task 6	0.940	0.340 (-0.60)
Task 10	0.402	0.374 (-0.03)
Task 11	0.565	0.269 (-0.30)
Task 12	0.350	0.327 (-0.02)
Task 13	0.685	0.544 (-0.14)

Table 1: Comaprison of Matthew's correlation coefficient (ϕ) for Human-GT and Human-LLM (llama-2-base-70b) across 300 examples. Tasks 1-10 examine Implicature, Tasks 11-12 assess Presupposition, and Task 13 focuses on Reference and Deixis. Red text indicates correlation differences between Human-GT and Human-LLM for each task.

from the speaker. However, the models exhibit shortcomings in capturing the speaker's assumptions, known as presuppositions. On average, there is a performance gap of $\sim 15\%$ between humans and the best-performing model on these tasks. Notably, the model's sensitivity to hints and task variations is also an important aspect. Human performance remains consistent across sarcasm detection and agreement detection tasks, whereas the models show significant performance discrepancies in these tasks (with an average difference of 13%). Similarly, this gap is observed in tasks concerning figurative language understanding, with models showing an average gap of $\sim 25\%$ and human performance only differs by 1%.

5.2 Error Analysis

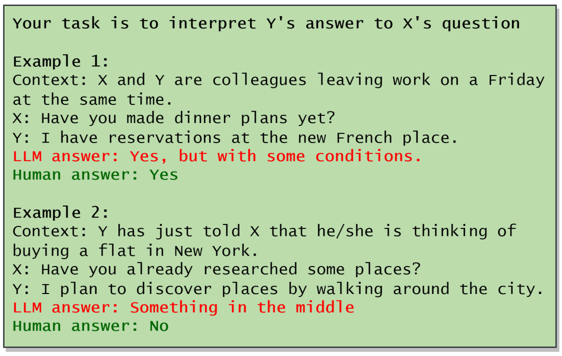
In this section, we analyse cases where LLMs fall short in simple pragmatic understanding tasks that humans do with ease. Specifically, we consider the llama-2-70b base model due to its consistently high performance across various tasks. We compare mistakes of humans and LLMs to see if there is any correlation in pragmatic understanding and if so, is it significant? To see this correlation, we report ϕ (Matthew's correlation coefficient) in Table 1 comparing human-LLMs (llama-2-70b-base) and human-GT correlation values. ϕ ranges from -1 to 1 where 1 means total agreement, 0 means the predictions are random with respect to the actual values, and -1 means total disagreement. For most tasks, the human-LLM correlation values are above random. This suggests that models make some mistakes similar to humans, but this is far from a human-ground truth correlation. For instance, the performance of LLMs is comparable to humans for response classification with implied meaning (Figure 4 - Task 3), but the correlation values say otherwise. This is further supported by Figure 6 showing that LLMs do make different mistakes than humans during classification.



	Task 1	Task 5	Task 6	Task 12	Task 13
GT vs LLM	1074 183 200 1023	518 472 96 894	232 758 56 934	398 235 21 326	1324 822 155 242
GT vs Human	154 0 26 124	145 5 10 140	147 3 53 97	67 33 44 80	78 12 31 179

Figure 6: Confusion matrix comparing mistakes of LLMs vs. Humans against ground truth answers. These tasks are chosen to have binary and consistent options for all questions in the task.

Now, we present examples for each pragmatic phenomenon to understand the pragmatic abilities of LLMs qualitatively. For response classification (Figure 7), the model selects that the response is true given some conditions are met, unlike humans, who consider the context only as auxiliary information (Example 1). We also encounter examples where Y's response is what we call a "polite decline" since there isn't a direct no in the response but an implied No in a tactful manner (Example 2).

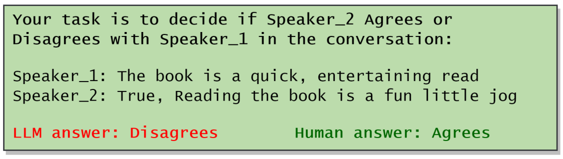


```
Your task is to interpret Y's answer to X's question

Example 1:
Context: X and Y are colleagues leaving work on a Friday
at the same time.
X: Have you made dinner plans yet?
Y: I have reservations at the new French place.
LLM answer: Yes, but with some conditions.
Human answer: Yes

Example 2:
Context: Y has just told X that he/she is thinking of
buying a flat in New York.
X: Have you already researched some places?
Y: I plan to discover places by walking around the city.
LLM answer: Something in the middle
Human answer: No
```

Figure 7: Examples of response classification (Task 2)

For understanding implicature in figurative language, we often see responses where metaphors, hyperbole, and tautological statements exist but are in agreement with the speaker. Figure 8 shows that the model often confuses agreements with figurative language as sarcastic disagreement (Task 5) but can correctly differentiate sarcastic statements from statements that agree with the speaker (Task 6).

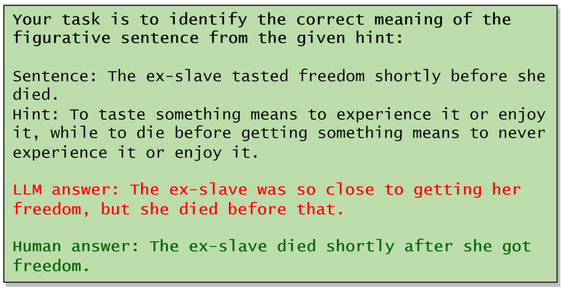
Using distractors in figurative-language under-



```
Your task is to decide if Speaker_2 Agrees or
Disagrees with Speaker_1 in the conversation:

Speaker_1: The book is a quick, entertaining read
Speaker_2: True, Reading the book is a fun little jog

LLM answer: Disagrees        Human answer: Agrees
```

Figure 8: Example of agreement detection in figurative language (Task 5)

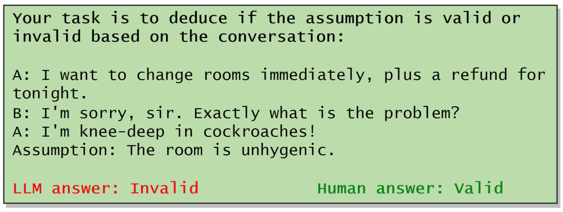
standing tasks shows LLMs' vulnerability in their pragmatic abilities. Humans are robust to see that the hint is contrasting and helps distinguish the options in the context and choose the correct one (Figure 9).



```
Your task is to identify the correct meaning of the
figurative sentence from the given hint:

Sentence: The ex-slave tasted freedom shortly before she
died.
Hint: To taste something means to experience it or enjoy
it, while to die before getting something means to never
experience it or enjoy it.

LLM answer: The ex-slave was so close to getting her
freedom, but she died before that.

Human answer: The ex-slave died shortly after she got
freedom.
```

Figure 9: Example of Figurative language understanding task with contrastive hint (Task 9)

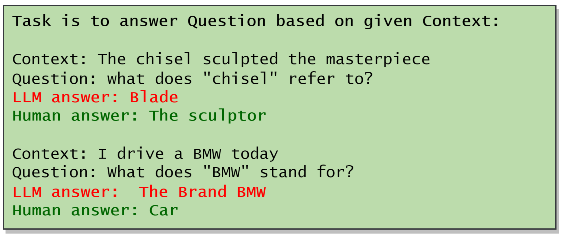
In instances of presupposition, we observe a recurring pattern where the model erroneously interprets negatives as positives. In Figure 10, Speaker A expresses frustration about the unsanitary condition of the room, attributing it to the presence of cockroaches. However, the model incorrectly dismisses the notion that being "knee-deep in cockroaches" signifies unhygienic conditions, deeming it an invalid presupposition.



```
Your task is to deduce if the assumption is valid or
invalid based on the conversation:

A: I want to change rooms immediately, plus a refund for
tonight.
B: I'm sorry, sir. Exactly what is the problem?
A: I'm knee-deep in cockroaches!
Assumption: The room is unhygenic.

LLM answer: Invalid             Human answer: Valid
```

Figure 10: Example of the presupposition task (Task 12)

Although llama-2 achieves better results than humans in Metonymy understanding, it makes trivial mistakes where humans get it right. Humans fail, too, when a reference is one that they are not familiar with, but LLMs, due to access to vast and diverse sources of texts, get it right. This task requires common sense and world knowledge to understand references that humans learn over time. Figure 11 shows examples where the LLM takes the semantic meaning of the reference instead of the pragmatic one.



```
Task is to answer Question based on given Context:

Context: The chisel sculpted the masterpiece
Question: what does "chisel" refer to?
LLM answer: Blade
Human answer: The sculptor

Context: I drive a BMW today
Question: What does "BMW" stand for?
LLM answer:  The Brand BMW
Human answer: Car
```

Figure 11: Examples from Metonymy (Task 14)

This error analysis shows that LLMs don't make the same mistakes as humans. Importantly, LLMs fail in trivial cases where humans easily understand the underlying pragmatic answer. More insight into why LLMs fail in such cases is in the scope of our future research work.

6 Conclusion

In this study, we introduce the Pragmatic Understanding Benchmark (PUB) designed to assess pragmatic comprehension in LLMs. We offer a detailed analysis, providing insights into various aspects of pragmatic understanding within LLMs. Our observations reveal that pragmatic understanding in LLMs can be enhanced through instruction-tuning of these models. Interestingly, even without specific fine-tuning, language models at scale exhibit equivalent performance. Notably, smaller models, particularly the instruction-tuned variants, outperform their base counterparts, but this advantage diminishes as models scale up, with base and instruction-tuned models showing comparable performance. Despite advancements, LLMs are yet to attain human-level performance, especially in tasks requiring a deep understanding of language context. The observed variability in model performance across different tasks within the same dataset highlights the complexity of achieving human-like pragmatic understanding in LLMs. The PUB benchmark thus provides a clear indication of where LLMs currently stand and the strides still needed to reach human parity in language understanding. We hope that this benchmark will aid researchers in improving LLMs' conversational abilities with humans.

Limitations

Our work addresses an important benchmark that can be used to understand and improve the chat capabilities of language models. While we carefully put together a benchmark for evaluation, it's important to note that there might be biases present that may show up in evaluations. Furthermore, we employed different sampling techniques to avoid evaluation bias for different classes. Although we tried our best to evaluate the models consistently, the models are sensitive to prompt wordings. For the same prompts too, the models are not consistent with the answers when changed the order of options as mentioned in PPA. Therefore there can be slight variations in the performances when trying to reproduce the results. The human evaluation scores reported in the paper are done by graduate students who are proficient in English and language understanding, the results may vary for different sets of human evaluators. The inconsistency of language models is another issue for MCQA results (Robinson and Wingate, 2023), since inconsistency in answers can lead to false results but until better evaluation methods arrive, we rely on the methods currently used in the paper.

Ethics Statement

This study adhered to the ACL Ethics Policy. All annotators and human evaluators received fair compensation. Our datasets solely serve the purpose of evaluating the pragmatic comprehension of LLMs. We make our dataset available for research and educational purposes, with no expectation of it being misused for malicious intent.

References

Jens Allwood. 1981. On the distinctions between semantics and pragmatics. In *Crossing the Boundaries in Linguistics: Studies Presented to Manfred Bierwisch*, pages 177–189. Springer.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7139–7159. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Lingjia Deng, Janyce Wiebe, and Yoonjung Choi. 2014. Joint inference and disambiguation of implicit sentiments via implicature constraints. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 79–88. ACL.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. 2021. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*.

Laurence R Horn and Gregory L Ward. 2004. *The handbook of pragmatics*. Wiley Online Library.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4194–4213. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8690–8705. Association for Computational Linguistics.

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 779–785. International Committee on Computational Linguistics.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2022. (qa)^2: Question answering with questionable assumptions. *CoRR*, abs/2212.10003.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Shibamouli Lahiri. 2015. Squinky! A corpus of sentence-level formality, informativeness, and implicature. *CoRR*, abs/1506.02306.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4437–4452. Association for Computational Linguistics.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. " i'd rather just go to bed": Understanding indirect answers. *arXiv preprint arXiv:2010.03450*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *ArXiv*, abs/1609.07843.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro

Pezzelle, Marco Baroni, Gemma Boleda, and R. Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *ArXiv*, abs/1606.06031.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in english. In *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 349–366. Association for Computational Linguistics.

Peng Qi, Nina Du, Christopher D. Manning, and Jing Huang. 2023. Pragmaticqa: A dataset for pragmatic question answering in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6175–6191. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022. A pragmatics-centered evaluation framework for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2382–2394. European Language Resources Association.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

George Yule. 1996. *Pragmatics*. Oxford university press.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2074–2085. Association for Computational Linguistics.

A Annotation details

For CircaPlus, Considering the subjectivity inherent in implicature, we employ two expert English linguists for the annotation process and implement double-blind checking for the annotations. For DialogAssumptions, all the conversations from the DailyDialog dataset were given to 2 linguistic experts. These experts were asked to add presuppositions to random dialog turns from the datasets. The annotators were also instructed to create false presuppositions and mark them as invalid. Only those examples that are in agreement of both the experts are considered. The Metonymy dataset is curated by four graduates of Literature and two linguist experts from a reputable university. The annotators are given basic examples from Wikipedia and a list of metonymic words as references. We encouraged the annotators to discover new metonymic words in order to avoid repetition in the data. They create these examples from scratch while referring to the provided instructions and examples. All the examples were verified by the experts.

B Tasks

B.1 Implicature

Implicature, an unspoken aspect of a speaker's meaning, extends beyond the literal content in a speaker's message. Understanding implicature is crucial for LLMs, as it allows them to interpret context, discern implied messages, and produce responses that surpass literal text, ensuring more contextually suitable, human-like, and meaningful interactions. Owing to the importance of implicature in pragmatics we have designed *ten* tasks that thoroughly test the LLM's abilities to capture this phenomenon.

Task 1 - Direct/Indirect classification This task evaluates language models' capability to distinguish between direct and indirect responses, crucial for understanding user intentions in dialogue systems. The model receives context, a question, and a response (that can be direct or indirect) and then selects between two options: A) Direct answer and B) Indirect answer. We utilized a label-balanced set of 2,500 data points sourced from the Circa dataset for this purpose.

Task 2 and 3 - Response classification without implied meaning and with implied meaning: Task 2 involves categorizing indirect answers using five labels. The model receives context, a question, and an indirect answer and must choose the most fitting label from options A) Yes, B) No, C) Yes, subject to conditions, D) In the middle, neither yes nor no, E) Other. This task evaluates LLMs' ability to comprehend indirect responses, specifically within polar Question and Answer scenarios, utilizing the Circa dataset. Task 3, an extension of Task 2, introduces implied meanings as additional cues to assist LLMs in interpreting indirect answers. The implied meaning acts as a chain-of-thought prompt for understanding indirect responses, assessed using the CircaPlus dataset. Both tasks involve evaluating 2,500 data points.

Task 4 - Implicature recovery Task 4 differs from tasks 2 and 3 by focusing on implicature recovery in non-polar Question and Answer contexts. In this task, we present the conversation which is a sequence of QAs $(Q_1, A_1), (Q_2, A_2), .., (Q_n, A_n)$ and four choices for the implied meaning of A_n. The task for the model is to select an appropriate choice that resolve's the implicature to its explicit form, *i.e.*, to perform implicature recovery. We use 2000 data points from the Grice dataset for this task.

While prior tasks have focused on understanding implied meanings in conversations devoid of figurative language, it's important to note that figurative language is a common feature in human communication (Lakoff and Johnson, 2008). Understanding the underlying meanings when such language is used in dialogue is crucial. Therefore, to provide a comprehensive benchmark, we are introducing tasks that focus on understanding implied meanings in conversations where figurative language is present.

Task 5 and 6 - Agreement detection and Understanding sarcasm Task 5, "Agreement Detection", and Task 6, "Understanding Sarcasm", are both designed to evaluate a language model's ability to comprehend and interpret figurative language within a dialogue. In Task 5, the model is given a conversation between two speakers, a question, and two options: A: Agrees and B: Disagrees. Speaker 1 uses figurative language, and Speaker 2 responds either in agreement or disagreement. The model's objective is to accurately determine if the second speaker concurs with the first. Task 6 flips the roles from Task 5. Here, Speaker 1 makes a statement, and Speaker 2 responds with 'yes', but continues the sentence using figurative language to either agree or disagree (refer to Figure 3 for examples). The model is then tasked with correctly determining if the second speaker is in agreement with the first or is being sarcastic. Modifications are applied to the (Liu et al., 2022) dataset to accommodate both tasks. The evaluation involves 2000 data points for each of the tasks.

Task 7, 8 and 9 - Figurative language understanding using positive and contrastive hints Tasks 7, 8, and 19 are formulated based on the FLUTE dataset (Chakrabarty et al., 2022). The FLUTE dataset consists of sentences or premises in figurative language and their corresponding hypotheses in simple language. For each premise, there are two types of hypotheses: one that entails and another that contradicts. Additionally, the dataset includes separate explanations for the entailment and contradiction. In Task 7, the objective is to test if the figurative language is correctly understood. The model must choose between an entailed sentence or a contradictory sentence as the meaning of the premise. In Task 8, the model is provided with an explanation of the entailment, which is referred to as a positive hint as it explains why the entailment option is the correct meaning of the premise. In Task 9, an explanation of the contradictory statement is provided, along with an explanation of why it is not the correct meaning of the figurative sentence. This is considered a contrastive hint. Through these tasks, we aim to test if the models understand the task or if their responses rely on the semantic overlap with the positive hint. The evaluation involves 1770 data points for each of the tasks.

Task 10 - Implicature NLI Given that Natural Language Inference (NLI) is a well-established task in the training and evaluation of language models, we have incorporated the NLI task to assess whether the models are capable of making inferences when implicatures are involved. We use 2100 data points from IMPRESS(Jeretic et al., 2020) dataset for this task.

B.2 Presuppositions

Presuppositions in a sentence are the underlying assumptions or facts that are implicitly accepted as true by the speaker when making a statement.

Task 11 - Presupposition NLI In this task, we approach presupposition verification by framing it as Natural Language Inference (NLI), with an objective akin to that of task 10. We use 1800 data points from IMPRESS (Jeretic et al., 2020) NOPE (Parrish et al., 2021) dataset for this task.

Task 12 - QA over presupposition This task aims to test the ability of the language models on how well they can capture the speaker's assumptions in a dialog. We provide the model with a conversation (set of dialogues between two people), presupposition on the conversation, and two options A.

Valid and B. Invalid. The task for the model is to determine if the given presupposition is valid or invalid based on the conversation. We use 2500 data points from the newly annotated DialogAssumptions dataset for this task.

B.3 Reference

Deixis, which involves the act of pointing through language, encompasses expressions that are often among the earliest spoken by very young children. These expressions, such as person deixis ('me', 'you'), spatial deixis ('here', 'there'), or temporal deixis ('now', 'then') (Yule, 1996), are indicative of individuals, locations, or times. Deixis is a type of reference closely linked to the speaker's context.

Task 13 - Diectic QA This task is designed to access the model's capabilities in resolving references where deictic terms are used. The model is provided with a conversation containing deictic expressions, a polar question regarding reference resolution, and two answer options: A. "Yes" and B. "No.". The model's objective is to accurately determine and provide the correct response to the polar question within the context of the conversation. We selected all the questions and corresponding conversations from the GRICE dataset (Zheng et al., 2021) that have Yes/No answers. These questions were then filtered using a manually curated list of deictic terms. A total of 2000 data points are used for this task.

Task 14 - Referential metonymy The task aims to test the model's abilities to understand language use that involves referring to a target object/individual in terms of a distinctive or saliently associated feature. The model is presented with a context featuring metonymic references, along with a question and four possible options. The task requires the model to choose the most suitable option that correctly resolves the reference in response to the question. We use 1100 data points from the newly annotated MetoQA dataset for this task.

C Prompts used for each task

In this section we provide prompts used for each task. Any typos in the shown examples are present in the datasets they are drawn from. The examples presented here are Multiple Choice Prompts (MCPs). Cloze Prompts (CPs) can be obtained by removing the options from the MCPs.

```
Your task is to label the 'Response' as
an Indirect or Direct answer based on the
Context and Question:

Context: X wants to know what activities
Y likes to do during weekends.
Question: Are you a fan of bars?
Response: I love to drink beer at pubs.
Options:
A: Direct answer
B: Indirect answer
Correct option=
```

Figure 12: Prompt example for Task 1

```
Your task is to interpret Y's answer to
X's question into one of the options:
A: Yes
B: No
C: Yes, subject to some conditions
D: In the middle, neither yes nor no
E: Other

Context: X and Y are childhood neighbours
who unexpectedly run into each other at
a cafe.
X: Would you like to exchange numbers?
Y: I'll get my contacts open here.
Options:
A: Yes
B: No
C: Yes, subject to some conditions
D: In the middle, neither yes nor no
E: Other
Correct option=
```

Figure 13: Prompt example for Task 2

```
Your task is to interpret Y's answer to X's question into one of the options:
A: Yes
B: No
C: Yes, subject to some conditions
D: In the middle, neither yes nor no
E: Other

Context: X and Y are childhood neighbours who unexpectedly run into each other at a cafe.
X: Would you like to exchange numbers?
Y: I'll get my contacts open here.
Implied meaning: He likes to exchange numbers
Options:
A: Yes
B: No
C: Yes, subject to some conditions
D: In the middle, neither yes nor no
E: Other
Correct option=
```

Figure 14: Prompt example for Task 3

```
Your task is to understand the implied meaning in Speaker_2's last response and give the explicit meaning:

Speaker_1: did Liam leave the watermelons in the attic
Speaker_2: no, he didn't
Speaker_1: did Jackson leave the watermelons there
Speaker_2: he said he was not there
Speaker_1: where can I get them
Speaker_2: the watermelons are in the bathroom or the laundry
Speaker_1: what about the cherries
Speaker_2: they are in the kitchen
Speaker_1: did you see the cabbages
Speaker_2: there is a blue bathtub in the bathroom
Speaker_1: did you place the cabbages there
Speaker_2: no, I didn't
Speaker_1: are all of them there
Speaker_2: some are there
Speaker_1: how many cherries are in the kitchen
Speaker_2: there are at least one there
Speaker_1: did Liam put the cherries there
Speaker_2: he put them there and walked to the bathroom
Options:
A: Liam put the cherries in the kitchen and then walked to the bathroom
B: Liam didn't put the cherries in the kitchen
C: I put the cherries in the kitchen
D: Liam put the cherries in the kitchen
Correct option=
```

Figure 15: Prompt example for Task 4

```
Your task is to decide if Speaker_2 Agrees or Disagrees with Speaker_1 in the conversation:

Speaker_1: The chair was comfortable like a pillow.
Speaker_2: The chair was uncomfortable.
Options:
A: Agrees
B: Disagrees
Correct option=
```

Figure 16: Prompt example for Task 5

```
Your task is to decide if Speaker_2 Agrees
or is being Sarcastic with Speaker_1 in
the conversation:

Speaker_1: The chair was uncomfortable.
Speaker_2: Yeah, The chair was
comfortable like a pillow.
Options:
A: Agrees
B: Sarcastic
Correct option=
```

Figure 17: Prompt example for Task 6

```
Your task is to identify the correct
meaning of the figurative sentence:

Sentence : To add insult to injury, a boy
was leading a handsome sheep on a string
behind him.
Options:
A: To make things worse, a boy was leading
a handsome sheep on a string behind him.
B: In order to make things a lot better,
a boy was leading a handsome sheep on a
string behind him.
Correct option=
```

Figure 18: Prompt example for Task 7

```
Your task is to identify the correct
meaning of the figurative sentence from
the given hint:

Sentence : To add insult to injury, a boy
was leading a handsome sheep on a string
behind him.
Hint : To add insult to injury means to
make a bad situation worse, and in this
sentence the boy leading the sheep makes
the situation worse.
Options:
A: To make things worse, a boy was leading
a handsome sheep on a string behind him.
B: In order to make things a lot better,
a boy was leading a handsome sheep on a
string behind him.
Correct option=
```

Figure 19: Prompt example for Task 8

```
Your task is to identify the correct
meaning of the figurative sentence from
the given hint:

Sentence : To add insult to injury, a boy
was leading a handsome sheep on a string
behind him.
Hint : To add insult to injury means to
make a bad situation worse, but in this
sentence the boy leading the sheep makes
the situation better.
Options:
A: To make things worse, a boy was leading
a handsome sheep on a string behind him.
B: In order to make things a lot better,
a boy was leading a handsome sheep on a
string behind him.
Correct option=
```

Figure 20: Prompt example for Task 9

```
Premise: Amy could prevent Stephen from hiding.
Hypothesis: Amy couldn't prevent Stephen from hiding.
Options:
A: Hypothesis is definitely true given premise
B: Hypothesis might be true given premise
C: Hypothesis is definitely not true given premise
Correct option=
```

Figure 21: Prompt example for Task 10

```
Premise: Natalie hasn't discovered where Tracy worries.
Hypothesis: Tracy doesn't worry.
Options:
A: Hypothesis is definitely true given premise
B: Hypothesis might be true given premise
C: Hypothesis is definitely not true given premise
Correct option=
```

Figure 22: Prompt example for Task 11

```
Your task is to deduce if the Assumption is valid or invalid based on the conversation:

Conversation:
A: Say , Jim , how about going for a few beers after dinner ?
Assumption: Jim exists.
Options:
A: Valid
B: Invalid
Correct option=
```

Figure 23: Prompt example for Task 12

```
Your task is to answer the given question based on the conversation:

Conversation:
Speaker_1: did you go to the basement
Speaker_2: I walked to the cellar
Speaker_1: did you see the beans
Speaker_2: I have no idea
Speaker_1: what about the pumpkin
Speaker_2: it is in the hallway
Speaker_1: did you see the celeries
Speaker_2: there is a green pantry in the cellar
Speaker_1: did Mason place the celeries there
Speaker_2: he placed them there and walked to the hallway
Speaker_1: did he put the peaches in the cellar
Speaker_2: no, he didn't
Speaker_1: did Lily place them in the cellar
Speaker_2: no, she didn't
Speaker_1: where can I get the melons
Speaker_2: there is a red bottle in the cellar
Speaker_1: are all of them there
Speaker_2: yes
Speaker_1: where are the peaches
Speaker_2: the peaches are in the basement
Question: are the melons in the cellar?
Options:
A: yes
B: no
Correct option=
```

Figure 24: Prompt example for Task 13

```
Your task is to answer the Question based on the given Context:

Context: She is attracted to blue jacket
Question: What does "blue jacket" refer to?
Options:
A: Colour
B: Jacket
C: Sailor
D: Sea
Correct option=
```

Figure 25: Prompt example for Task 15

D Discussion about Data Leakage

LLMs have been trained on a vast amount of openly available data. However, this abundance of data raises concerns about the evaluation sets, as they can yield biased results when exposed to similar data during testing. We assess a wide range of models, which introduces the risk of data leakage. While we cannot conduct exhaustive collision checks with the training corpora of all these models due to their immense size, we have performed several studies to reduce the risk of data leakage in their fine-tuning datasets. Firstly, we have identified that Circa, Imppres, and DailyDialog are components of instruction-tuning datasets, such as Super Natural Instructions (Wang et al., 2022) and Flan (Wei et al.), on which flan-t5 is fine-tuned. GPT-3.5 model may also include them. Secondly, despite the potential for data leakage, flan-t5 demonstrates competitive performance on datasets it has never encountered before, such as Task 14, which is an entirely new dataset.

Since these datasets are available on public websites, it is likely that some part of the data might be seen in the pertaining corpora of these models, but we suspect the following reasons why data leakage does not affect our results for other models. First, we see that the models perform consistently on new data, and we do not notice a surge in numbers for a particular model on these tasks. Secondly, similar to Robinson and Wingate (2023), we see that shuffling candidate answers does not cause a dip in PPA performance , and if data leakage would have impacted our results then we would see more probability assigned to the correct answer regardless of the order of options as claimed by Robinson and Wingate (2023).

LLMs have been trained on a vast amount of openly available data. However, this abundance of data raises concerns about the evaluation sets, as they can yield biased results when exposed to similar data during testing. We assess a wide range of models, which introduces the risk of data leakage. While we cannot conduct exhaustive collision checks with the training corpora of all these models due to their immense size, we have performed several studies to reduce the risk of data leakage in their fine-tuning datasets. Firstly, we have identified that Circa, Imppres, and DailyDialog are components of instruction-tuning datasets, such as Super Natural Instructions (Wang et al., 2022) and Flan (Wei et al.), on which flan-t5 is fine-tuned. GPT-3.5 and Falcon models may also include them, to the best of our knowledge. Secondly, despite the potential for data leakage, flan-t5 demonstrates competitive performance on datasets it has never encountered before, such as Task 14, which is an entirely new dataset.

Since these datasets are available on public websites, it is likely that some part of the data might be seen in the pertaining corpora of these models, but we suspect the following reasons why data leakage does not affect our results for other models. First, we see that the models perform consistently on new data, and we do not notice a surge in numbers for a particular model on these tasks. Secondly, similar to Robinson and Wingate (2023), we see that shuffling candidate answers does not cause a dip in PPA performance , and if data leakage would have impacted our results then we would see more probability assigned to the correct answer regardless of the order of options as claimed by Robinson and Wingate (2023).

E Results

In this section, we presented results of all evaluation strategies in both 0-shot and 3-shot settings in tables 2,3,4,5,6,7,8,9,10,11,12,13,14,15

Direct/Indirect classification	0-shot CP	0-shot MCQA	3- CP	3-shot MCQA
task_0-flan-t5-xxl	50.68	62.36	54.03	62.02
task_0-llama-2-13b	49.32	51.76	57.86	72.14
task_0-llama-2-13b-chat	49.32	83.12	64.56	75.44
task_0-llama-2-70b	49.32	62.84	50.28	84.56
task_0-llama-2-70b-chat	49.32	77.28	64.15	78.43
task_0-llama-2-7b	49.32	17.36	55.65	60.77
task_0-llama-2-7b-chat	49.32	57.64	60.12	77.26
task_0-t5-11b	44.20	50.52	50.28	44.48
task_0-gpt 3.5	-	-	80.20	73.87

Table 2: Results for Task 1 - Direct/Indirect classification

Response classification	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_1-flan-t5-xxl	3.27	85.28	6.29	87.01
task_1-llama-2-13b	48.00	50.64	48.23	39.34
task_1-llama-2-13b-chat	48.00	27.85	49.60	53.82
task_1-llama-2-70b	48.00	57.90	48.23	63.19
task_1-llama-2-70b-chat	48.00	66.16	48.23	73.89
task_1-llama-2-7b	48.00	10.85	48.23	25.91
task_1-llama-2-7b-chat	48.00	62.73	48.23	45.17
task_1-t5-11b	49.36	0.08	49.60	0.00
task_0-gpt 3.5	-	-	58.18	43.81

Table 3: Results for Task 2 - Response classification

Response classification with hint	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_2-flan-t5-xxl	3.51	71.59	8.37	70.19
task_2-llama-2-13b	48.00	51.08	48.23	64.08
task_2-llama-2-13b-chat	48.00	54.15	48.23	67.70
task_2-llama-2-70b	48.00	71.71	48.23	78.56
task_2-llama-2-70b-chat	48.00	80.29	48.23	82.02
task_2-llama-2-7b	48.00	43.93	48.23	33.02
task_2-llama-2-7b-chat	48.00	66.56	48.23	55.31
task_2-t5-11b	48.08	0.28	48.23	0.08
task_0-gpt 3.5	-	-	62.77	53.02

Table 4: Results for Task 3 - Response classification with hint

Implicature recovery in dialog context	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_3-flan-t5-xxl	73.30	82.90	70.81	82.63
task_3-llama-2-13b	54.10	46.90	60.10	57.27
task_3-llama-2-13b-chat	55.35	58.45	56.92	63.89
task_3-llama-2-70b	55.90	66.90	62.73	75.91
task_3-llama-2-70b-chat	50.30	67.15	56.01	71.52
task_3-llama-2-7b	53.05	37.05	56.26	36.46
task_3-llama-2-7b-chat	56.85	45.60	54.24	37.02
task_3-t5-11b	25.60	0.00	0	0.00
task_0-gpt 3.5	-	-	76.55	78.13

Table 5: Results for Task 4 - Implicature recovery in dialog context

Agreement detection in Conversations with figurative language	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_4-flan-t5-xxl	59.85	75.00	75.81	75.66
task_4-llama-2-13b	46.75	55.90	50.71	53.13
task_4-llama-2-13b-chat	44.45	60.30	54.80	58.13
task_4-llama-2-70b	47.55	70.95	55.25	71.31
task_4-llama-2-70b-chat	47.90	65.70	51.11	65.10
task_4-llama-2-7b	50.00	49.95	51.46	50.00
task_4-llama-2-7b-chat	50.05	54.05	51.97	51.21
task_4-t5-11b	49.60	5.75	50.40	1.67
task_0-gpt 3.5	-	-	70.25	71.01

Table 6: Results for Task 5 - Agreement detection in Conversations with figurative language

Sarcasm detection in Conversations with figurative language	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_5-flan-t5-xxl	62.45	61.70	68.08	58.18
task_5-llama-2-13b	50.00	52.15	50.15	51.62
task_5-llama-2-13b-chat	50.00	57.00	51.62	51.67
task_5-llama-2-70b	50.00	51.05	50.30	58.89
task_5-llama-2-70b-chat	50.00	50.35	50.66	54.04
task_5-llama-2-7b	50.00	49.85	50.30	51.31
task_5-llama-2-7b-chat	50.00	50.00	54.34	51.16
task_5-t5-11b	49.85	0.15	50.00	0.00
task_0-gpt 3.5	-	-	55.50	54.85

Table 7: Results for Task 6 - Sarcasm detection in Conversations with figurative language

Figurative language understanding with no hints	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_6-flan-t5-xxl	58.19	92.66	63.77	93.14
task_6-llama-2-13b	80.06	83.22	79.37	81.37
task_6-llama-2-13b-chat	80.40	87.12	79.26	85.89
task_6-llama-2-70b	80.51	92.77	81.43	94.00
task_6-llama-2-70b-chat	82.71	92.43	82.86	91.71
task_6-llama-2-7b	77.46	66.95	78.63	61.66
task_6-llama-2-7b-chat	76.38	83.79	78.00	79.09
task_6-t5-11b	51.58	18.08	50.86	14.57
task_0-gpt 3.5	-	-	92.88	93.03

Table 8: Results for Task 7 - Figurative language understanding with no hints

Figurative language understanding with positive hint	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_7-flan-t5-xxl	67.97	97.23	77.20	98.06
task_7-llama-2-13b	88.36	94.18	89.03	92.17
task_7-llama-2-13b-chat	88.98	96.61	89.66	95.66
task_7-llama-2-70b	90.28	96.84	91.54	98.34
task_7-llama-2-70b-chat	90.45	97.97	91.71	97.37
task_7-llama-2-7b	86.78	87.51	88.91	70.40
task_7-llama-2-7b-chat	87.68	94.69	88.91	90.17
task_7-t5-11b	51.64	4.97	52.00	14.97
task_0-gpt 3.5	-	-	96.84	97.94

Table 9: Results for Task 8 - Figurative language understanding with positive hint

Figurative language understanding with contrastive hint	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_8-flan-t5-xxl	50.56	79.55	50.69	74.97
task_8-llama-2-13b	48.25	63.33	46.46	61.60
task_8-llama-2-13b-chat	52.54	58.53	46.63	59.89
task_8-llama-2-70b	47.80	76.84	41.94	72.91
task_8-llama-2-70b-chat	49.94	63.84	43.83	67.89
task_8-llama-2-7b	47.06	45.59	47.26	50.40
task_8-llama-2-7b-chat	49.94	56.55	47.71	59.54
task_8-t5-11b	49.60	4.07	47.43	15.14
task_0-gpt 3.5	-	-	73.05	71.43

Table 10: Results for Task 9 - Figurative language understanding with contrastive hint

Implicature as NLI	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_9-flan-t5-xxl	14.29	63.05	58.07	64.12
task_9-llama-2-13b	17.43	14.24	45.34	23.87
task_9-llama-2-13b-chat	21.81	12.67	41.74	33.67
task_9-llama-2-70b	17.95	53.38	55.09	54.32
task_9-llama-2-70b-chat	16.67	50.86	51.34	51.54
task_9-llama-2-7b	49.29	14.29	44.19	13.50
task_9-llama-2-7b-chat	41.14	7.67	31.36	10.57
task_9-t5-11b	14.29	27.95	28.58	2.31
task_0-gpt 3.5	-	-	48.86	32.52

Table 11: Results for Task 10 - Implicature as NLI

Presupposition as NLI	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_10-flan-t5-xxl	24.72	61.83	45.68	60.77
task_10-llama-2-13b	42.67	24.72	42.99	41.81
task_10-llama-2-13b-chat	47.61	34.94	36.92	42.31
task_10-llama-2-70b	39.39	48.00	45.23	53.20
task_10-llama-2-70b-chat	44.72	53.39	48.71	51.96
task_10-llama-2-7b	44.72	24.72	36.92	26.88
task_10-llama-2-7b-chat	38.94	36.28	38.16	39.67
task_10-t5-11b	24.72		27.27	3.82
task_0-gpt 3.5	-	-	50.67	45.74

Table 12: Results for Task 11 - Presupposition as NLI

QA over presuppositions	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_13-flan-t5-xxl	60.00	42.74	48.68	44.79
task_13-llama-2-13b	84.12	84.00	84.39	50.06
task_13-llama-2-13b-chat	84.12	70.15	84.39	45.54
task_13-llama-2-70b	84.12	24.93	84.39	61.58
task_13-llama-2-70b-chat	84.12	21.50	84.39	34.45
task_13-llama-2-7b	84.12	82.83	84.39	71.37
task_13-llama-2-7b-chat	84.12	62.93	84.39	22.57
task_13-t5-11b	37.96	60.40	38.69	59.77
task_0-gpt 3.5	-	-	21.94	41.89

Table 13: Results for Task 12 - QA over presuppositions

Deixis	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_11-flan-t5-xxl	80.90	74.00	83.06	79.39
task_11-llama-2-13b	64.30	51.50	64.59	47.76
task_11-llama-2-13b-chat	64.30	63.80	64.59	54.59
task_11-llama-2-70b	64.30	73.40	64.59	73.88
task_11-llama-2-70b-chat	64.30	56.00	64.59	65.31
task_11-llama-2-7b	64.30	64.50	64.59	61.43
task_11-llama-2-7b-chat	64.30	62.60	64.59	48.37
task_11-t5-11b	41.70	63.10	0	31.53
task_0-gpt 3.5	-	-	64.50	65.71

Table 14: Results for Task 13 - Deixis

Metonymy	0-shot CP	0-shot MCQA	3-shot CP	3-shot MCQA
task_14-flan-t5-xxl	48.91	67.03	50.46	65.07
task_14-llama-2-13b	29.91	71.83	31.74	73.97
task_14-llama-2-13b-chat	31.22	71.18	30.14	73.29
task_14-llama-2-70b	30.13	72.93	32.19	85.39
task_14-llama-2-70b-chat	29.04	69.21	31.74	74.20
task_14-llama-2-7b	29.48	52.18	30.14	63.93
task_14-llama-2-7b-chat	29.69	62.45	30.14	61.87
task_14-t5-11b	39.30	26.20	27.63	20.09
task_0-gpt 3.5	-	-	73.58	73.97

Table 15: Results for Task 14 - Metonymy