PUB: A <u>Pragmatics Understanding Benchmark for Assessing LLMs</u>' Pragmatics Capabilities

Settaluri Lakshmi Sravanthi[°], Meet Doshi[°], Tankala Pavan Kalyan[°],

Pushpak Bhattacharyya[°], Rudra Murthy[§], Raj Dabre[‡]

[°]CFILT, Indian Institute of Technology Bombay

[§]IBM Research

[‡]NICT, Japan

{sravanthi,meetdoshi,pb}@cse.iitb.ac.in,

190020124@iitb.ac.in, rmurthyv@in.ibm.com, prajdabre@gmail.com

Abstract

LLMs have demonstrated remarkable capability for understanding semantics, but they often struggle with understanding pragmatics. To demonstrate this fact, we release a Pragmatics Understanding Benchmark (PUB) dataset consisting of fourteen tasks in four pragmatics phenomena, namely, Implicature, Presupposition, Reference, and Deixis. We curated high-quality test sets for each task, consisting of Multiple Choice Question Answers (MCQA). PUB includes a total of 28k data points, 6.1k of which have been created by us, and the rest are adapted from existing datasets. We evaluated nine models varying in the number of parameters and type of training. Our study indicates that fine-tuning for instructionfollowing and chat significantly enhances the pragmatics capabilities of smaller language models. However, for larger models, the base versions perform comparably with their chat-adapted counterparts. Additionally, there is a noticeable performance gap between human capabilities and model capabilities. Furthermore, unlike the consistent performance of humans across various tasks, the models demonstrate variability in their proficiency, with performance levels fluctuating due to different hints and the complexities of tasks within the same dataset. Overall, the benchmark aims to provide a comprehensive evaluation of LLM's ability to handle real-world language tasks that require pragmatic reasoning.

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



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

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 the realm of Natural Language Processing (NLP), Large Language Models (LLMs) (GPT-3 (Brown et al., 2020), BLOOM (Scao et al., 2022), PaLM (Chowdhery et al., 2022), LLAMA-2 (Touvron et al., 2023), others) have emerged as a transformative force in recent years. LLMs have shown remarkable abilities on many downstream tasks like NLU (GLUE (Wang et al., 2019b), MultiNLI (Williams et al., 2018)), Text generation (LAM-BADA, Wikitext), Code synthesis (APSS, HumanEval (Chen et al., 2021)), QA (Natural Questions, ARC, OpenbookQA (Mihaylov et al., 2018),

^{*} Equal contribution in coding and experiments

SQuAD (Rajpurkar et al., 2018)), Reasoning (SuperGLUE (Wang et al., 2019a), GSM8k (Cobbe et al., 2021), Strategy QA (Geva et al., 2021)), etc.

As LLM's capabilities have expanded, they are now being utilized in practical real-world applications like chatbots, search engines, and web browsers. Given the increased interaction between humans and LLMs, the following research questions need to be answered:

a. How much do LLMs understand what humans mean during conversations?

b. Is there a correlation between a model's pragmatics abilities and its scale?

c. Do LLMs that are optimized for dialogue usecases exhibit superior pragmatic abilities?

d. Despite operating on the same dataset, do LLMs demonstrate varying task sensitivity?

e. How do the pragmatic abilities of LLMs compare concerning world knowledge involvement?f. Do they understand the same implied meaning and make the same assumptions as us?

To answer these questions we lean towards the domain of *pragmatics*. 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. Most benchmarks until now deal only with abilities like problem-solving (Cobbe et al., 2021) or semantic understanding (GLUE (Wang et al., 2019b), BigBench (Srivastava et al., 2022), etc.) where LLMs have started to come close or be at par with human benchmarks. Despite the recent progress, we notice that there is still a lot of pragmatic understanding gap between what the language model understands and what was meant by a statement. To facilitate this research, we propose a Pragmatic Understanding Benchmark (PUB) over four major Pragmatic phenomena, namely, Implicature (Understanding what is suggested or implied in a statement even though it is not literally expressed), Presupposition (An implicit assumption that is taken for granted before the use of a statement), Deixis (a phenomenon in which certain words or phrases within a sentence or discourse rely on contextual cues, such as the speaker, the listener, or the surrounding context, to convey their meaning effectively) and 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).

In PUB, we've constructed tasks based on datasets focusing on Implicature, Presupposition, Deixis and Reference. The benchmark 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) 3 tasks similar to (Robinson and Wingate, 2023), this way we can evaluate the model's certainty in its predictions to achieve higher performance.

Our contributions are: (1) a comprehensive and unified dataset for 14 distinct tasks in pragmatics (Figure: 2), 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 LLMs¹. (2) a systematic evaluation of 6 variations of llama-2, t5, Flan-t5, and GPT-3.5, on the 14 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 contextdependent aspects of meaning that are systematically abstracted away from, in the construction of content or logical form (Horn and Ward,

¹The benchmark is available at https: //huggingface.co/datasets/cfilt/PUB

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.

Recent studies (Hu et al., 2023; Ruis et al., 2023) highlight language models' struggle with humor, irony, and conversational maxims. Previous evaluations either focused on singular phenomena or had limited sample sizes, like (Deng et al., 2014; Sileo et al., 2022; Qi et al., 2023). To the best of our knowledge, we are the first ones to combine major aspects of pragmatics to create a quantifiable benchmark.

3 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. Details of newly annotated datasets are discussed below:

- 1. **CircaPlus** is our 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 pre-

suppositions 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").

4 Tasks

In this section, we describe each task and the associated dataset. Each task incorporated within PUB is structured to evaluate distinct domains of pragmatics. Figure 2 contains examples from each task.

4.1 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 classificationThis 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 labelbalanced set of 2,500 data points sourced from the Circa dataset for this purpose.



Figure 2: Examples of each task from PUB, The tasks are divided across *four* domains of pragmatics (Implicature, Presupposition, Reference, and Deixis). Our proposed benchmark builds upon existing pragmatic datasets and combines our newly annotated datasets comprising 6k annotations to complete the pragmatic evaluation test suite with 28k examples. We have reformatted the existing datasets into MCQA prompts that explicitly test these abilities.

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 2 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.

4.2 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.

4.3 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.



Figure 3: Comparison of various models' multiple choice symbol binding using PPA. Results averaged across Task 4, 11, and 14, representing different pragmatic domains.

5 Methodology

We have selected two evaluation methods namely length normalized Cloze prompting (Brown et al., 2020) and Multiple Choice Prompting (MCP) (Robinson and Wingate, 2023) considering the capabilities of all the models. We have also computed the Proportion of Plurality Agreement (PPA) (Robinson and Wingate, 2023) for all the models to ensure the model's consistency across possible orders of answer options. The results for PPA are presented in Figure 3. We see that vanilla LLMs show improved consistency with a few shots, while instruction-tuned models don't benefit from additional examples. The models under investigation include flan-t5-xxl (Chung et al., 2022), llama-2 (Touvron et al., 2023), t5 (Raffel et al., 2020), and GPT-3.5 Brown et al. (2020).

5.1 Sampling for few-shot prompts

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. These 20 examples were selected to ensure a balanced representation of options. For tasks that have unique options for each question, 20 examples were randomly selected from the entire dataset. Depending on the value of k for k-shot prompt, k samples were randomly selected from this dev set. The remaining instances of the data, other than the dev set, were used to evaluate the model.

5.2 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 for each task. Each of the 3



Figure 4: Results (accuracy) 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.

human evaluators evaluated these 100 samples for 14 tasks. In total, we have performed 4,200 human evaluations. The samples were chosen to ensure a balanced representation of all option types. 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 random human performance on complex pragmatic tasks. These evaluators are presented with the same prompt as the *0-shot* MCP presented to the LLMs.

6 Results and Analysis

We evaluate all the open-source models using both the evaluation methods, i.e. length normalized cloze prompt method and multiple choice prompts. In each of these methodologies, we do a *zero-shot* evaluation and a *3-shot evaluation*. The OpenAI model is evaluated using MCP.

6.1 Results

The results of our experiments are presented in Figures 4, 5. Based on these results, we try to address the questions raised in the introduction.

How much do LLMs understand what humans mean during conversations? To evaluate how well LLMs understand what humans intend during conversations, tasks related to implicature and reference offer pertinent insights. We observe that the models perform moderately in the classification of a response as direct or indirect. They struggle to interpret the meaning of the indirect response. Notably, except for the *llama-70b-chat* model, this trend persists across the models evaluated. Furthermore, in this specific task, a slight but noticeable increase in performance is observed across most models when a hint is provided. Interestingly this pattern aligns closely with human performance. The performance trend remains the same in task 4, focusing on resolving implicature in non-polar question-answer scenarios. Even though, NLI is an established task in NLP, it is observed that models perform poorly on making pragmatic inferences. Figure 1 shows that the average performance on implicature and reference tasks is similar.

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. Interestingly, LLMs demonstrate stronger performance in agreement detection compared to sarcasm detection (on average there is a 13% performance gap in models > 13b parameters) tasks within the same dataset. 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 to contrastive there is a drastic decrease (on an average of 20%) in the accuracy levels for this task across all the models. Interestingly, the inclusion of a positive



Figure 5: Results for Task 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.

hint, which has a higher lexical overlap with the correct answer, seems to boost the performance of the model. In contrast, the model's performance appears to decrease 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 overall performance depicted 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 chatoptimized variants of *llama* slightly outperform the base models on most of the tasks. However, there is a notable performance discrepancy 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, bridging the gap toward human-like understanding and processing of language.

How do the pragmatic abilities of LLMs compare concerning world knowledge involvement? In implicature tasks, excluding task-1 (Direct/Indirect classification) and task-4 (Implicature recovery in dialog context), the other tasks involve a certain degree of world knowledge. While the Metonymy task requires world knowledge, the Deixis task does not. Upon reviewing the outcomes, it becomes evident that the model's belowpar 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 rather than their 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

Task No.	GT-Human	Human-LLM
Task 1	0.829	0.749 (- <mark>0.08</mark>)
Task 2	0.681	0.421 (- <mark>0.26</mark>)
Task 3	0.754	0.550 (- <mark>0.20</mark>)
Task 5	0.901	0.515 (- <mark>0.39</mark>)
Task 6	0.940	0.340 (- <mark>0.60</mark>)
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 (<mark>-0.14</mark>)

Table 1: Phi coefficient (ϕ) correlations among Ground Truth (GT), Human evaluator (Human), and LLaMA-2-base-70B (LLM) 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 GT-Human and Human-LLM for each task.

meanings from the speaker. However, the models exhibit shortcomings in capturing the speaker's assumptions, known as presuppositions, as evidenced by the results of presupposition tasks (on average there is a performance gap of $\sim 15\%$ between humans and best performing model). Notably, the model's sensitivity to hints and task variations is 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 also observed in tasks concerning figurative language understanding with models showing an average gap of $\sim 25\%$ and human performance only differs by 1%.

6.2 Error Analysis

In this section, we look into cases where LLMs fall short in simple pragmatic understanding tasks that humans do with ease. More specifically, we consider the LLaMA-2-70b base model due to its consistently high performance across various tasks and models. For implicature understanding, we see that the model fails to understand the meaning of the response when the response involves complex language phenomena like phrases, expressions, assumptions, or instances where common sense is needed, etc. 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 the correlation between human evaluators and LLMs, we report the Phi

coefficient ϕ (Matthew's correlation coefficient) in Table 1 between LLMs (LLaMA-2-70b-base) vs human evaluators (Human-LLM) and compare it with ground truth vs human evaluators (GT-Human). 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. Although we see that for some tasks the correlation values are more than random in Human-LLM, meaning they do make some similar mistakes when compared with GT-Human to see that still there is a large difference and LLMs do not always make the same mistakes as humans. This can be seen in Task 3, where the performance is the same for the LLM and human is the same but there is a correlation gap. This can also be seen in Figure 6 where LLMs do make different mistakes than humans during classification.



Figure 6: Confusion matrix comparing ground truth with Language Models (LLMs) and ground truth with humans, revealing LLMs' tendency to misclassify positive labels as negatives. Here GT refers to ground truth.

For the task of response classification, we see examples where the model thinks that the response is true given some conditions are met but humans do not consider the context as a condition but rather as an auxiliary information. See examples below

```
Task 2
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.
Chosen answer: Yes, but with some conditions.
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.
Chosen answer: Something in the middle
```

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. For understanding implicature in figurative language, we often see examples where metaphors, hyperbole, and tautological statements exist but are in agreement with the speaker.

Task 6
Speaker_1: The book is a quick, entertaining
 read
Speaker_2: True, Reading the book is a fun
 little jog
Chosen answer: Sarcastic disagreement

We see that in Tasks 5 and 6 the model often confuses agreements with figurative language as sarcastic disagreement but can correctly differentiate sarcastic statements from statements that agree with the speaker, as shown below. Using distractors in figurative language understanding tasks shows how vulnerable LLMs are in their pragmatic abilities. We see that adding a distractor hint in the task confuses LLM and in many cases falls short whereas humans are more robust and see that the hint is contrasting and helps distinguish both meanings of the sentence in the context and choose the correct one.

```
Task 9
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.
Chosen answer: The ex-slave was so close to
    getting her freedom, but she died before
    that.
```

In instances of presupposition, we observe a recurring pattern where the model erroneously interprets negatives as positives. In the following example, 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 "kneedeep in cockroaches" signifies unhygienic conditions, deeming it an invalid presupposition.

```
Task 12
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 unhygienic.
Chosen answer: Invalid
```

Although LLaMA-2 achieves better results compared to humans in Metonymy understanding, it makes trivial mistakes where humans get it right. But humans fail in cases when reference is one which 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 a bit of world knowledge to understand references which humans learn over time. A few examples are given below where the LLM takes the semantic meaning of the reference instead of the pragmatic one.

```
Task 14
Context: The chisel sculpted the masterpiece
Question: what does "chisel" refer to?
Chosen answer: Blade
Context: I drive a BMW today
```

Question: What does "BMW" stand for? Chosen answer: The Brand BMW

From this error analysis, we find that LLMs don't make the same mistakes as humans and get confused easily, but more importantly, LLMs fail in trivial cases where humans easily understand the underlying pragmatic answer. More insight into why LLMs fail in such cases is required but we leave that for future research work.

7 Conclusion

In this study, we introduce the Pragmatic Understanding Benchmark (PUB) designed to assess pragmatic comprehension in LLMs. We offer a comprehensive analysis, providing insights into various aspects of pragmatic understanding within LLMs. Our findings 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 pragmatic understanding. Notably, smaller models, particularly the instruction-tuned variants, outperform their base counterparts, but this advantage diminishes as models scale up, with base and instructiontuned 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.

8 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.

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.
- 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.
- 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 transformerbased 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.
- 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*.
- 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.
- 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 openaccess multilingual language model. CoRR, abs/2211.05100.
- Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022. A pragmaticscentered 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 generalpurpose 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.

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