DO LLMS UNDERSTAND PRAGMATICS? AN EXTENSIVE BENCHMARK FOR EVALUATING PRAGMATIC UNDERSTANDING OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are typically evaluated based on semantic understanding and are believed to be capable of handling general language processing. While LLMs can mimic human-like responses, they still are a contraption in their pragmatic or contextual understanding of language. To test this hypothesis, we subject LLMs to the complex task of pragmatics. We conducted evaluation across *fourteen* tasks spanning *four* domains of pragmatics namely, Implicature, Presupposition, Reference, and Deixis. For each task, we curated high-quality test sets, consisting of Multiple Choice Question Answers (MCQA). We evaluate a wide range of LLMs with different types and sizes. Our findings reveal that LLMs with no instruction fine-tuning have near-random accuracy on many tasks. The performance gradually increases with the increase in model capacity. Additionally, we create a unified benchmark enabling the research community to better assess the underlying pragmatic understanding of the language models.

1 INTRODUCTION

With an increase in understanding of how to better train LLMs, we have now started to see models which are trained over trillions of tokens and over billions of parameters, (Chronological order: GPT-3 (Brown et al., 2020), BLOOM (Scao et al., 2022), PaLM (Chowdhery et al., 2022), LLAMA-2 Touvron et al. (2023), others) which have shown remarkable abilities on many downstream tasks like NLU (GLUE (Wang et al., 2019b), MultiNLI (Williams et al., 2018)), Text generation (LAMBADA, Wikitext), Code synthesis (APSS, HumanEval (Chen et al., 2021)), QA (Natural Questions, ARC, OpenbookQA (Mihaylov et al., 2018), SQuAD (Rajpurkar et al., 2018)), Reasoning (SuperGLUE Wang et al. (2019a), GSM8k (Cobbe et al., 2021), Strategy QA (Geva et al., 2021)), etc. Moreover, these language models often show a correlation between their size and their performance, which is referred to as the scaling law Kaplan et al. (2020).

But as we move towards more and more complex language models, we need to ask the question: How much do LLMs understand what humans actually mean during conversations? 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 which deals with understanding meaning in context or in change of context (Grice, 1975). While semantics involves the study of words and their meanings in a language, *pragmatics* extends this inquiry by considering words' meanings within the context in which they are used. For example, consider a situation where Alice asks Bob whether he would like to meet today, and Bob responds by saying, "*I have a few things to take care of.*" In this context, Bob doesn't explicitly state whether he would like to meet or not, but it is implicit from his answer that he cannot meet today. Here, the implicature arises because Bob's response is less informative than Alice might have expected if he were entirely available. By adhering to the Maxim of Quantity Grice (1975), Bob indirectly conveys that he might not have all his time free without explicitly saying so. This illustrates how the same statement can have a different interpretation semantically and pragmatically depending on the context. It is easier for humans to capture this type of phenomenon but not so much for language models. *Pragmatic competence* can be defined as the ability to understand a speaker's intended meaning and the additional information conveyed implicitly.

In linguistic theory, a significant milestone in the development of a systematic framework for pragmatics was Grice (1975) work, which showcased how a structured approach to language use facilitates a more simplified and elegant description of language structure. From then, pragmatics has emerged as a crucial subfield of linguistics that deals with phenomena such as implicature, presupposition, speech acts, reference, and deixis. The handbook of pragmatics (Horn & Ward, 2004) or Grice (1975) gives a comprehensive overview of both the traditional and the extended goals of theoretical and empirical pragmatics.

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 actually meant by a statement. To facilitate this research, we propose an LLM understanding evaluation benchmark 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 this unified benchmark to evaluate the Pragmatic abilities of LLMs, we devise tasks upon existing datasets for Implicature, Reference, Deixis, and Presupposition and provide four new datasets comprising 6100 newly annotated examples, along with human evaluation on a sample of these datasets to compare performance with existing 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 & 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. More information can be found in 1 and Appendix B.

Following (GPT-3 (Brown et al., 2020), Joshua Robinson & David Wingate (Robinson & 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 evaluate the Proportion of Plurality Agreement (PPA) agreement on 3 tasks similar to (Joshua Robinson & David Wingate), this way we can evaluate the model's certainty in its predictions to achieve higher performance. We do this evaluation for an array of different models ranging from the smallest Flan-t5-small 60M Chung et al. (2022) to GPT-3 Brown et al. (2020) and Falcon 180B, which vary in size, tuning mechanism, and amount of pretraining corpora. We argue that understanding pragmatics is not an emergent ability in LLMs, yet the majority of these models exhibit a notable disparity in pragmatic understanding when compared to humans. They even encounter challenges with certain fundamental pragmatic tasks that humans can solve effortlessly.

Our primary contributions include (1) A comprehensive and unified benchmark that consists of 14 distinct tasks, collectively containing an average of 6k data points. The primary objective of this benchmark is to assess the pragmatic performance of LLMs across a range of scenarios and linguistic contexts. (2) A systematic evaluation of various LLMs, each employing different prompting styles, across the tasks included in our proposed benchmark. (3) Human performance on a sample of the benchmark to highlight the gap between LLMs and humans. The benchmark we have introduced serves as a valuable evaluation framework for assessing LLMs' performance in pragmatic tasks. We hope that this benchmark will help researchers in improving LLMs' conversational abilities with humans.

2 PRAGMATICS BENCHMARK

In this section, we will introduce the tasks that are used in the evaluation benchmark. A few existing datasets cover various pragmatic phenomena. With the help of linguistic experts, we selected the existing datasets that cover at least one important pragmatic phenomenon. More 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). However, these datasets do not explicitly evaluate implicature, presupposition, and reference understanding in conversations. Therefore we created 4 new datasets on top of existing conversational datasets

like CIRCA, DailyDialog (Li et al., 2017), and Convokit (Chang et al., 2020) comprising a total of 6100 newly annotated data points. We reframe all tasks according to expert prompts and instructions as an MCQA task Robinson & Wingate (2023). More details and examples can be found in Figure 1. The annotation details are presented in Appendix B.

We list down details of the unified benchmark below:

Circa is a dataset containing 34,628 crowdsourced pairs of polar questions and indirect answers in English. It extends beyond binary yes/no responses to include conditionals, uncertain, and middle-ground answers. We use existing declarative direct answers and indirect answers from Circa to formulate Task 1, aiming to identify whether LLMs can understand the difference between a direct and indirect response. For interpreting indirect answers to polar questions, we provide the original task to LLMs along with instructions and class labels in Task 2.

Grice is a grammar-based dialogue dataset designed for implicature recovery, coreference resolution, and conversational reasoning. The dataset is methodically created using a hierarchical grammar model. We present the original problem formulation as provided by Zheng et al. (2021) in Task 4 for implicature recovery and Task 12 for deictic conversational reasoning. In these tasks, we filter the dataset into four types of deixis, namely person, spatial, temporal, and discourse deixis, using common types of deictic terms. Our filtering steps are listed in Appendix A.

FigQA is a Winograd-style nonliteral language understanding dataset. It consists of 10,256 paired figurative phrases with divergent meanings. We use the existing FigQA dataset to formulate a new figurative agreement detection task between sentences in Task 5 and a new sarcasm detection task in Task 6.

FLUTE, is an NLI-style semi-synthetic dataset that contains 9k literal, figurative sentence pairs with entail/contradict labels and the associated explanations, spanning four categories: Sarcasm, Simile, Metaphor, and Idioms. We reframe the FLUTE's Recognizing Textual Entailment (RTE) task as an MCQA task to check whether language models can accurately capture figurative meanings.

IMPPRES, is an NLI-style semi-automatically generated dataset containing 25.5K pairs of sentences containing well-studied pragmatic inference types with entail/contradict/neutral labels. This dataset is generated according to linguist-crafted templates. We prompt IMPPRES as an MCQA task similar to FLUTE.

NOPE, is an NLI-style dataset containing 2386 human annotated presuppositions from the sentences extracted from the COCA dataset (Davies, 2010) using 10 different types of presupposition triggers. We do not utilize the 346 adversarial examples provided by the authors. We combine IMPPRES and NOPE together for evaluation.

CircaPlus is our newly annotated dataset used in Task 3 to determine if LLMs can interpret direct answers with the assistance of implied meanings written by humans. For this, we sampled a test set from Circa and annotated 2,651 implied meanings based on the indirect responses. Considering the subjectivity inherent in implicature, we employ two expert English linguists for the annotation process and implement double-blind checking for the annotations.

DialogAssumptions is our newly annotated dataset for Task 12, containing 2.5k pairs of expert-annotated presuppositions based on a subset of dialogues from the Dailydialog dataset (Li et al., 2017). It is designed to evaluate whether the presuppositions that humans have before a sentence is uttered in a conversation are being properly understood by language models.

MetoQA is our newly annotated dataset used in Task 15 that consists of 744 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").

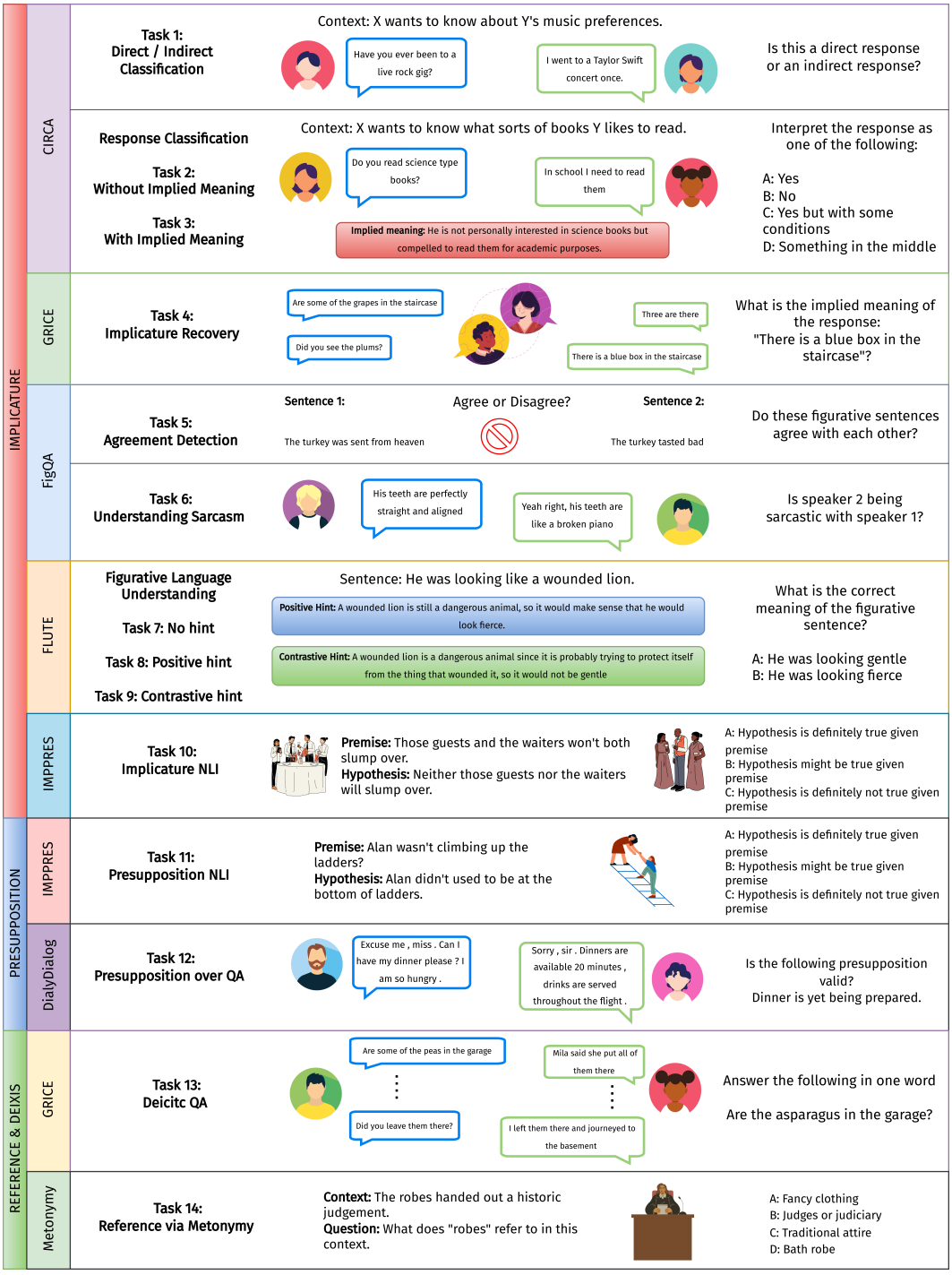

Figure 1: Examples of each task from our Pragmatic benchmark, The tasks are divided across *four* domains of pragmatics (Implicature, Presupposition, Reference, and Deixis). Our proposed benchmark builds upon existing pragmatic datasets and combines our newly annotated datasets comprising of 6k annotations to complete the pragmatic evaluation test suite. We have reformatted the existing datasets into MCQA prompts that explicitly test these abilities. The data annotation process and the formulation of the tasks are provided in Appendix B. The prompts for each task are given in Appendix D. We also perform human evaluation on our benchmark to compare the performance of LLMs.

3 EXPERIMENTAL SETUP

In this section, we describe different evaluation methods and models used. We have selected two evaluation methods namely Cloze prompting and Multiple Choice Prompting taking the capabilities of all the models into consideration.

3.1 CLOZE PROMPTING (CP)

In the cloze prompting approach, a question is given to an LLM, and the model independently scores each potential answer. The answer with the highest probability is selected by the model. Brown et al. (2020) acknowledged that the probabilities of answers could be affected by particularly frequent or rare tokens or sequences of different lengths, so they employed two normalization methods. One method involves normalizing the probability of a sequence for its length by taking the n^{th} root; $P(x_1, x_2, \ldots, x_n) = \sqrt[n]{\prod_{i=1}^{n} P(x_i)}$. The length normalization strategy requires N forward passes through LLMs as compared to 2N forward passes in the other normalization strategy. Since our primary goal is to evaluate the pragmatic abilities and not the normalization strategy, in this paper, we follow length normalization for all the evaluations involving the cloze prompting approach.

3.2 MULTIPLE CHOICE PROMPTING (MCP)

In Multiple Choice Prompting, a question and its candidate answers, each associated with a symbol, are combined into a single prompt for an LLM. The model is structured to predict only one token (e.g., "A", "B", etc.). The model's answer is the answer choice corresponding to the token with the highest probability. Consequently, the probabilities of these symbols act as a substitute for the probabilities of each answer. A notable limitation of this evaluation method is that models exhibiting suboptimal performance in the context of *multiple choice symbol binding* (MCSB) tend to yield inferior results Robinson & Wingate (2023). Therefore we also perform the Proportion of Plurality Agreement (PPA) experiments for all the models to estimate the MCSB abilities of these models.

3.3 PROPORTION OF PLURALITY AGREEMENT (PPA)

When presenting a multiple-choice question, the potential answers must be arranged in a specific sequence. In general, human responses to such questions exhibit order-invariance, meaning that the order of the options does not affect the answer selection. Robinson & Wingate (2023) have proposed a method to verify if *LLMs* exhibit the same characteristics. Given a question with *n* answer options, there are $n!$ different ways these options can be associated with an ordered, fixed set of symbols. To compute PPA, the model is presented with the question using each unique permutation of the answer options. For every permutation, the model assigns a probability to each answer, and the answer with the highest probability is recorded. Subsequently, the PPA for the question is calculated as the ratio of permutations that selected the plurality answer to the total number of permutations. PPA measures order invariance irrespective of the model's ability to perform a task. A model with consistent answers across possible orders of answer options will have a high PPA, even if it performs poorly on the task. For a dataset with n answer choices per question, the baseline PPA is $1/n$.

3.4 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 three human evaluators evaluated these 100 samples. 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 exact same prompt as the MCP presented to the LLMs. These MCP prompts can be found in Appendix D.

3.5 MODELS

We consider a range of large language models with varying sizes, from 60 million to 175 billion parameters, to evaluate their pragmatic competence. The models under investigation include Falcon, Flan-T5 (Chung et al., 2022), Llama-2 (Touvron et al., 2023), Phi Gunasekar et al. (2023); Li et al. (2023), T5 Raffel et al. (2020), and GPT-3.5 Brown et al. (2020). We have chosen these models for their unique contributions to the field of natural language processing and their diverse capabilities. T5 is an encoder-decoder model known for its strong performance in various NLP tasks. We have included *3-billion* and *11-billion* versions of T5. Flan-T5 is an enhanced version of T5 that has been instruction fine-tuned on a mixture of tasks, offering improved performance across different tasks. We have included all five variants of Flan-T5 (small, base, large, xl, and xxl) to observe the accuracy across sizes. Llama-2 and LLama-2-chat are relatively large models that outperform open-source chat models on most benchmarks. Llama-2-chat is a variant of Llama-2 that has been supervised fine-tuned and iteratively refined using Reinforcement Learning from Human Feedback (RLHF). We have considered all variants of Llama-2 and Llama-2-chat to assess their *pragmatic competence*. We've opted for the *gpt-3.5-turbo-instruct* model among the available OpenAI models because it is one of the most recent releases, and it provides access to *log probabilities*. We have also included phi-1 and phi-1.5 which are of significantly smaller size but are trained on textbook data.

4 RESULTS

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 only using zero-shot MCP and 3-shot MCP for a cost-efficient evaluation. The results of our experiments can be found in Table 1 and Table 2. In the tables, we display the best figures (among zero-shot length normalized CP, 3-shot length normalized CP, zero-shot MCP and 3-shot MCP) for each model and task. The detailed results are available in Appendix D.

The tasks for implicature evaluation are meticulously designed to cover a spectrum of challenges. Even though there is remarkable proximity in performance between LLMs and humans in most of the tasks, the models still are much behind in unseen tasks like agreement detection in conversations with figurative language. Intriguingly, LLMs exhibit more robust performance in the task of sarcasm detection within the same dataset. This phenomenon can be attributed to the well-established nature of sarcasm detection as a task in the existing literature, whereas agreement detection in this context is less commonly encountered. The performance of LLMs on tasks on figurative language understanding with no hints, positive hints, and contrastive hints gives us an insight into how easily LLMs can be distracted. Whereas human performance is almost consistent on these tasks. The Deictic QA task tests how well LLMs can understand words in context and figure out what they refer to, checking their ability to grasp the context and identify references. In the Metonymy task, LLMs are evaluated on their capacity to recognize and interpret metonymic substitutions. Most of the models in these tasks fall below the baseline and human performance.

In the QA over presupposition task, many open-source models exhibit performance levels similar to the majority baseline. Still, their reliability is questionable due to a recurrent issue where these models consistently select only one label ("valid"). However, this behavior doesn't apply to GPT-3.5, which is an exception. Interestingly, GPT-3.5 falls considerably short of human performance in this task, underscoring the significant difficulties models encounter when tasked with presupposition comprehension.

PPA results for open-source models are presented in Figure 2. In many cases, the models show higher PPA scores when they are fine-tuned with specific instructions or filtered using RLHF compared to their "vanilla" versions. Additionally, within the LLAMA series of models, there's a noticeable pattern: larger model sizes tend to correspond with higher PPA values. These PPA values serve as indicators of the model's ability to maintain consistent performance regardless of the order in which answers are presented. These scores play a pivotal role in determining whether a particular model is suitable for Multiple Choice Prompts (MCP) evaluations.

Overall, both variants of the LLAMA-2 model with 70 billion parameters consistently exhibit strong performance across a wide range of tasks.

#Params	Model name	Task Number										
		1	2	3	4	5	6	7	8	9	10	Avg
-	Human	90.85	74.00	79.67	93.67	95.00	97.00	92.33	96.33	91.33	57.91	86.81
-	Baseline	50.68	48.54	48.00	25.85	50.00	50.00	50.45	50.45	50.45	78.09	50.25
60M	Flan-T5-small	50.69	37.77	37.99	**83.74**	50.81	50.46	51.53	57.18	76.23	35.35	53.18
220M	Flan-T5-base	51.20	69.88	56.98	76.97	56.87	62.69	77.80	88.25	60.46	57.19	65.83
770M	Flan-T5-large	72.24	54.22	50.28	79.30	65.50	73.43	87.57	95.71	71.13	62.97	71.24
1.3B	Phi-1	49.36	49.60	48.23	44.80	50.05	50.00	53.33	69.37	52.51	36.74	50.40
	Phi-1.5	50.69	49.60	48.23	54.65	50.00	53.09	70.17	82.57	52.43	28.90	54.03
3.5B	T5	61.72	47.95	47.75	28.45	50.61	50.76	51.69	54.29	47.77	49.42	49.04
	Flan-T5-XL	69.72	84.00	54.83	82.63	73.23	91.43	91.47	97.03	78.25	**65.51**	78.81
7B	Llama-2	60.77	49.36	48.23	56.26	51.46	61.66	78.63	88.91	59.54	49.29	60.41
	Falcon	54.91	49.36	48.23	55.00	50.51	51.16	64.74	77.94	53.72	35.49	54.11
	Llama-2-Ins	77.26	62.73	66.56	56.85	54.05	79.09	83.79	94.69	56.55	41.14	67.27
	Falcon-Ins	60.08	49.36	48.23	64.60	50.91	52.74	64.29	75.89	53.14	37.27	55.65
11B	T5	50.52	49.60	48.23	25.60	50.40	50.00	51.58	52.00	49.60	28.57	45.61
	Flan-T5-XXL	62.36	**87.01**	71.59	82.90	**75.81**	93.14	93.14	98.06	**79.55**	64.12	**80.77**
13B	Llama-2	72.14	50.64	64.08	63.89	55.90	81.37	83.22	94.18	63.33	45.34	67.41
	Llama-2-Ins	83.12	53.82	67.70	75.91	60.30	85.89	87.12	96.61	72.91	41.73	72.51
40B	Falcon	68.64	49.60	49.00	56.72	52.85	50.20	86.89	96.72	59.26	23.61	59.35
	Falcon-Ins	50.77	49.60	17.39	58.48	56.60	86.23	87.62	96.23	59.03	8.19	57.01
70B	Llama-2	**84.56**	63.19	78.56	71.52	71.31	**94.00**	**94.00**	**98.34**	76.84	55.91	78.82
	Llama-2-Ins	78.43	73.89	**82.02**	67.15	65.70	91.71	92.43	97.97	63.84	51.54	76.47
175B	GPT-3.5	80.20	58.18	62.77	78.13	71.01	55.50	93.03	97.94	73.05	48.86	71.87
-	MAX	84.56	87.01	82.02	83.74	75.81	94.00	94.00	98.34	79.55	65.51	84.45

Table 1: Results (accuracy) for all tasks on Implicature. The task numbers are as mentioned in Figure 1. 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. We present individual numbers in Appendix C and their respective prompts in Appendix D. Results considering Data leakage for vanilla LLMs can be found in Appendix C. **Max** denotes maximum accuracy across all models and **Avg** denotes average accuracy across the row. The bottom right value indicates average of maximum of all models.

4.1 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 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[1], 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 consis-

[1] https://github.com

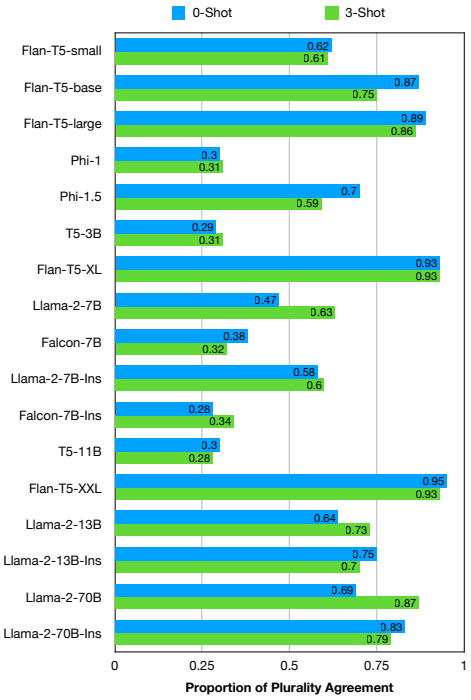

Figure 2: Comparison of PPA depicting the multiple choice symbol binding ability of different models. We see that for vanilla LLMs, few shot increases model consistency, and for instruction-tuned models adding more examples do not increase its consistency. The results are average across Task 4, 11, and 14 (one from each domain of pragmatics). Detailed results can be found in Appendix C.

#Params	Model name	Task Number				
		11	12	13	14	Avg
-	Human	69.70	85.67	67.50	80.00	75.72
-	Baseline	43.72	84.12	64.30	27.72	54.97
60M	Flan-T5-small	54.00	56.40	84.35	40.61	58.84
220M	Flan-T5-base	48.50	64.29	83.89	56.55	63.31
770M	Flan-T5-large	52.28	77.60	83.61	63.10	69.15
1.3B	Phi-1	42.77	64.59	84.38	17.47	52.30
	Phi-1.5	33.72	65.10	84.38	47.95	57.79
3.5B	T5	29.52	64.39	60.40	37.12	47.86
	Flan-T5-XL	57.22	**83.30**	77.49	72.27	72.57
7B	Llama-2	44.72	64.59	**84.39**	63.93	64.41
	Falcon	40.85	64.59	**84.39**	27.29	54.28
	Llama-2-Ins	39.67	64.59	**84.39**	62.45	62.78
	Falcon-Ins	36.03	64.59	**84.39**	27.73	53.19
11B	T5	27.27	63.10	60.40	39.30	47.52
	Flan-T5-XXL	**61.83**	83.06	76.71	67.03	72.16
13B	Llama-2	42.99	64.59	**84.39**	73.97	66.49
	Llama-2-Ins	47.61	64.59	**84.39**	73.29	67.47
40B	Falcon	38.00	64.59	56.84	31.51	47.74
	Falcon-Ins	43.66	64.59	60.47	31.96	50.17
70B	Llama-2	53.20	73.88	84.38	**85.39**	74.21
	Llama-2-Ins	53.39	65.31	**84.39**	74.20	69.32
175B	GPT-3.5	50.67	65.71	45.10	73.97	58.86
-	MAX	61.83	83.30	84.39	85.39	78.73

Table 2: Results (accuracy) for Tasks on Presupposition, Reference, and Deixis. The task numbers are as mentioned in Figure 1. 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. We present individual numbers in Appendix C and their respective prompts in Appendix D. Results considering Data leakage for vanilla LLMs can be found in Table Appendix C. **Max** denotes maximum accuracy across all models and **Avg** denotes average accuracy across the row. The bottom right value indicates average of maximum of all models.

tently on new data, and we do not notice a surge in numbers for a particular model on these tasks. Secondly, similar to Robinson & Wingate (2023), we see that shuffling candidate answers does not cause a dip in PPA performance (Appendix C), 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 & Wingate (2023).

5 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 & Ward (2004). Some of the basic subfields of pragmatics include implicature, presupposition, speech acts, reference, deixis, and 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) use indirect answers in polar questions, Zheng et al. (2021) utilize hierarchical grammar models to understand implicature and deictic reference in simple conversations, Jeretic et al. (2020) utilize NLI to understand scalar implicatures, Deng et al. (2014) make use of implicature rules to optimize sentiment detection, Lahiri (2015) create 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 & Cheung (2022) also reveal 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 understand humor, irony, and conversational maxims Grice (1975). These approaches have offered only a restricted understanding of the shortcomings exhibited by these models by either evaluating only a single phenomenon or with a smaller number of samples to make it quantifiable. Other existing work includes Deng et al. (2014) the PragmEval framework Sileo et al. (2022), which does not cover important aspects of pragmatics that can be used to evaluate LLMs and PragmaticQA Qi et al. (2023), open-domain question answering dataset, consisting of 6873 QA pairs, designed to delve into pragmatic reasoning within conversations spanning various topics, but so far, it has not been released openly.

Nevertheless, previous empirical investigations have predominantly assessed language models on their abilities on tasks like language modeling (Marcus et al., 1994), translation (Edunov et al., 2018), common sense reasoning (Srivastava et al., 2022), comprehension, summarisation (Fabbri et al., 2019), language understanding (Wang et al., 2019b), etc. Any practical implementation of language models that necessitates interaction with humans will rely on the models' capacity for pragmatic communication skills. This crucial ability for effective communication in applications involving humans isn't fully considered by the benchmarks used to assess how well language models align with this requirement. To the best of our knowledge, we are the first ones to combine major aspects of pragmatics to create a quantifiable benchmark.

Robinson & Wingate (2023) show that MCSB is an important ability for language models to be consistent in their answers. Following Rae et al. (2021), Chinchilla (Hoffmann et al., 2022), InstructGPT (Chan, 2023), Robinson & Wingate (2023) also demonstrate that MCP can significantly enhance LLM accuracy and reduce unpredictability in evaluations. Since CP has also been a reliable evaluation scheme in prompting followed by Brown et al. (2020), Holtzman et al. (2021), Zhao et al. (2021), and Izacard et al. (2022), we utilize both CP and MCP for evaluations since models with lower MCSB abilities can utilize Cloze prompts.

6 Conclusion

Our research provides a unified benchmark for evaluating pragmatic understanding in LLMs, combining existing and newly annotated datasets. We offer curated test sets with MCQA prompts to assess LLMs' capabilities in different pragmatic domains. LLMs excel in semantic understanding but struggle with pragmatic comprehension. Our benchmark evaluation across fourteen tasks in four pragmatics domains shows that vanilla LLMs struggle to capture context and perform slightly better than random baseline. LLMs perform well on tasks with positive hints but face challenges with contrastive hints and new tasks like reference resolution in metonymic language. The evaluation indicates their difficulty in understanding figurative language and the importance of context. LLMs also struggle with tasks related to agreement detection, where humans outperform them. On the contrary, we observe instruction fine-tuning helps models achieve near-human performance on many tasks.

In conclusion, while LLMs have made strides in semantic understanding, our research highlights the need for further development in pragmatic comprehension. Addressing the identified limitations will lead to more contextually aware and human-like language models, benefiting applications like chatbots, information retrieval, and dialogue systems. By pushing model development and evaluation boundaries, we can strive for more sophisticated and nuanced language models aligned with human pragmatic capabilities.

References

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agar-

wal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: figurative language understanding through textual explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 7139–7159. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.481. URL https://doi.org/10.18653/v1/2022.emnlp-main.481.

Anastasia Chan. GPT-3 and instructgpt: technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry. *AI Ethics*, 3(1):53–64, 2023. doi: 10.1007/s43681-022-00148-6. URL https://doi.org/10.1007/s43681-022-00148-6.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Z. Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. Convokit: A toolkit for the analysis of conversations. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (eds.), *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pp. 57–60. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.sigdial-1.8/.

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. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. URL https://arxiv.org/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. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. doi: 10.48550/arXiv.2204.02311. URL https://doi.org/10.48550/arXiv.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. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/arXiv.2210.11416. URL https://doi.org/10.48550/arXiv.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. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Mark Davies. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464, 2010.

Lingjia Deng, Janyce Wiebe, and Yoonjung Choi. Joint inference and disambiguation of implicit sentiments via implicature constraints. In Jan Hajic and Junichi Tsujii (eds.), *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 79–88. ACL, 2014. URL https://aclanthology.org/C14-1009/.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 489–500. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1045. URL https://doi.org/10.18653/v1/d18-1045.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 1074–1084. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1102. URL https://doi.org/10.18653/v1/p19-1102.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://doi.org/10.1162/tacl_a_00370.

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

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. *CoRR*, abs/2306.11644, 2023. doi: 10.48550/arXiv.2306.11644. URL https://doi.org/10.48550/arXiv.2306.11644.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7038–7051. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://doi.org/10.18653/v1/2021.emnlp-main.564.

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

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*

2023, Toronto, Canada, July 9-14, 2023, pp. 4194–4213. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.230. URL https://doi.org/10.18653/v1/2023.acl-long.230.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *CoRR*, abs/2208.03299, 2022. doi: 10.48550/arXiv.2208.03299. URL https://doi.org/10.48550/arXiv.2208.03299.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models imppressive? learning implicature and presupposition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8690–8705. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.768. URL https://doi.org/10.18653/v1/2020.acl-main.768.

Jad Kabbara and Jackie Chi Kit Cheung. Investigating the performance of transformer-based NLI models on presuppositional inferences. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pp. 779–785. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.65.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. $(qa)^2$: Question answering with questionable assumptions. *CoRR*, abs/2212.10003, 2022. doi: 10.48550/arXiv.2212.10003. URL https://doi.org/10.48550/arXiv.2212.10003.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In Greg Kondrak and Taro Watanabe (eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pp. 986–995. Asian Federation of Natural Language Processing, 2017. URL https://aclanthology.org/I17-1099/.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need II: phi-1.5 technical report. *CoRR*, abs/2309.05463, 2023. doi: 10.48550/arXiv.2309.05463. URL https://doi.org/10.48550/arXiv.2309.05463.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *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*, pp. 4437–4452. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.330. URL https://doi.org/10.18653/v1/2022.naacl-main.330.

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

Mitchell P. Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: Annotating predicate argument structure. In *Human Language Technology, Proceedings of a Workshop held at Plainsboro, New*

Jerey, USA, March 8-11, 1994. Morgan Kaufmann, 1994. URL https://aclanthology.org/H94-1020/.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1260. URL https://doi.org/10.18653/v1/d18-1260.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. NOPE: A corpus of naturally-occurring presuppositions in english. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pp. 349–366. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.conll-1.28. URL https://doi.org/10.18653/v1/2021.conll-1.28.

Peng Qi, Nina Du, Christopher D. Manning, and Jing Huang. Pragmaticqa: A dataset for pragmatic question answering in conversations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 6175–6191. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.385. URL https://doi.org/10.18653/v1/2023.findings-acl.385.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL https://arxiv.org/abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 784–789. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Joshua Robinson and David Wingate. 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, 2023. URL https://openreview.net/pdf?id=yKbprarjc5B.

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. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/arXiv.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. A pragmatics-centered evaluation framework for natural language understanding. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pp. 2382–2394. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.255.

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. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615, 2022. doi: 10.48550/arXiv.2206.04615. URL https://doi.org/10.48550/arXiv.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. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/arXiv.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 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, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

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. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5085–5109. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.340. URL https://doi.org/10.18653/v1/2022.emnlp-main.340.

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. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *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)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1101. URL https://doi.org/10.18653/v1/n18-1101.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhao21c.html.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. GRICE: A grammar-based dataset for recovering implicature and conversational reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pp. 2074–2085. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.182. URL https://doi.org/10.18653/v1/2021.findings-acl.182.

A SAMPLING

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.

B ANNOTATION DETAILS

For Task 14, we have created a new dataset for Reference via metonymy. This dataset is curated by four recent graduates of Literature from a reputable university. The annotators are given basic examples from Wikipedia and a list of metonymic words as references. We encourage 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.

For Task 13, we selected all the questions and corresponding conversations from the GRICE dataset that have Yes/No answers. These questions were then filtered using a manually curated list of deictic

terms. The filtered questions were randomly sampled and manually verified to ensure that they test the phenomenon of deixis.

For Task 12, 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. Regular feedback was provided to the annotators to maintain an almost equal number of samples for both valid and invalid presuppositions.

Task 10 and 11 are directly taken from the Impress dataset without any modifications in the statements and options. These tasks are presented as simple Natural Language inference tasks.

Task 7, 8, and 9 are formulated based on the FLUTE dataset. 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 responder must choose between an entailed sentence or a contradictory sentence as the meaning of the premise. In Task 8, the responder 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 a correct meaning of the figurative sentence. This is considered a contrastive hint. Through these tasks, we aim to test if the models are able to answer correctly based on good semantic overlap with the positive hint, or if they actually understand the tasks. The poor performance of the models on the task with contrastive hints validates this, as most models tend to choose the option with high semantic overlap with the given hint.

The Tasks 5 and 6 were constructed using the FigQA dataset. This dataset consists of sentences with figurative language and their corresponding meanings. Each sentence also has a negative version, which has the opposite meaning compared to the original sentence. In Task 5, the figurative sentence is spoken by the first person in the conversation, while the corresponding meaning or the opposite meaning is spoken by the second person. The responder is asked to determine whether the first speaker agrees with the second speaker or not. In Task 6, the first speaker speaks a simple sentence, while the second person speaks a figurative sentence that either has the same meaning or the opposite meaning. To make the second sentence more conversational, certain words like 'Yeah', 'Yes', 'True', and 'Of course' were added. If the figurative sentence has the same meaning as the first sentence, then both speakers agree with each other. If the figurative sentence has the opposite meaning, then the second speaker is being sarcastic with the first speaker.

Task 4 is directly taken from the GRICE dataset without any modification. The GRICE dataset includes, for each turn of every conversation, a question about the implied meaning of the response in that turn, along with four options. We present the entire conversation and select the last turn to provide the corresponding options for the implied meaning of the last response in the conversation.

The CIRCA dataset is utilized for tasks 1, 2, and 3. It consists of pairs of YES/NO questions and indirect answers, along with annotations for interpreting the indirect answers. Task 1 involves using the YES/NO and indirect answers to determine if a response to a question is direct or indirect. In task 2, the responder classifies the response as Yes, No, Yes upon some condition, etc. The prompt example provides all the options for this task. The questions in task 2 were annotated by 2 expert annotators to capture the implied meaning of the response given to the question. The implied meaning explains the response given to the asked question. In task 3, the responder is provided with both the implied meaning and the question from task 2.

C Prompting Experiments

In this section we show our results for all prompting and PPA experiments. We also note that if we avoid models with potential data leakage and consider vanilla LLMs then we can see that Llama-2 performs consistently well, even better than GPT-3.5 at pragmatic tasks. We notice a gap between the understanding of these LLMs and humans.

#Params	Model name	Task Number									
		1	2	3	4	5	6	7	8	9	10
-	Human	90.85	74.00	79.67	93.67	95.00	97.00	92.33	96.33	91.33	57.91
-	Baseline	50.68	48.54	48.00	25.85	50.00	50.00	50.45	50.45	50.45	78.09
60M	Flan-T5-small	51.2	69.88	56.98	69.7	56.75	51.95	77.8	88.25	58.25	57.19
220M	Flan-T5-base	51.2	69.88	56.98	69.7	56.75	51.95	77.8	88.25	58.25	57.19
770M	Flan-T5-large	72.24	52.08	50.28	79.3	62.05	61.45	87.57	95.71	71.13	61.76
1.3B	Phi-1	0	0	0	0	0	0	0	0	0	0
	Phi-1.5	49.4	25.6	47.64	34.45	41.75	49.95	60.79	81.07	52.43	14.28
3.5B	T5	61.72	4.16	1.16	24.05	41.2	44.9	40.4	14.58	14.52	14.29
	Flan-T5-XL	60	84	54.83	81.6	71.1	59.05	91.47	96.78	78.25	64.33
7B	Llama-2	60.77	49.36	48.23	56.26	51.46	61.66	78.63	88.91	59.54	49.29
	Falcon	50.68	48	47.64	0.00	0	0	0.00	0	53.72	-
	Llama-2-Ins	77.26	62.73	66.56	56.85	54.05	79.09	83.79	94.69	56.55	41.14
	Falcon-Ins	54.16	33.91	22.42	25.45	50	50.05	52.37	56.55	50.4	13.57
11B	T5	50.52	0.07	0.28	0	5.75	0.15	18.08	4.97	4.07	27.95
	Flan-T5-XXL	62.36	85.27	71.59	82.9	75	61.7	92.66	97.23	79.55	63.05
13B	Llama-2	51.76	50.64	51.08	46.9	55.9	52.15	83.22	94.18	63.33	14.24
	Llama-2-Ins	83.12	27.85	54.15	58.45	60.3	57	87.12	96.61	58.53	12.67
40B	Falcon	68.64	10.97	49	-	52.85	50	86.89	96.72	59.26	23.61
	Falcon-Ins	49.48	10.01	17.39	-	56.6	50	87.62	95.64	59.03	8.19
70B	Llama-2	62.84	57.9	71.71	66.9	70.95	51.05	92.77	96.84	76.84	53.38
	Llama-2-Ins	77.28	66.16	80.29	67.15	65.7	50.35	92.43	97.97	63.84	50.86
175B	GPT-3.5	80.20	58.18	62.77	76.55	70.25	55.50	92.88	96.84	73.05	48.86

Table 3: Results (accuracy) for all tasks on 0 shot MCQA for Implicature. The task numbers are as mentioned in Figure 1.

#Params	Model name	Task Number									
		1	2	3	4	5	6	7	8	9	10
-	Human	90.85	74.00	79.67	93.67	95.00	97.00	92.33	96.33	91.33	57.91
-	Baseline	50.68	48.54	48.00	25.85	50.00	50.00	50.45	50.45	50.45	78.09
60M	Flan-T5-small	48.60	37.40	37.99	47.55	49.85	50.00	50.85	57.01	48.87	14.29
220M	Flan-T5-base	48.68	37.76	53.83	56.70	56.00	50.25	50.85	54.63	47.68	14.29
770M	Flan-T5-large	49.28	2.20	3.27	63.15	65.50	60.75	53.84	61.13	46.84	14.29
1.3B	Phi-1	49.36	49.36	48.00	39.55	50.00	50.00	53.33	69.27	40.56	9.23
	Phi-1.5	49.32	49.36	48.00	50.85	50.00	50.00	69.49	82.43	45.31	28.90
3.5B	T5	23.36	40.52	46.97	28.45	50.00	49.45	51.69	53.11	47.23	14.29
	Flan-T5-XL	55.28	2.20	3.27	69.55	68.00	62.85	53.28	61.02	46.61	14.29
7B	Llama-2	49.32	49.36	48.00	53.05	50.00	50.00	77.46	86.78	47.06	49.29
	Falcon	49.32	49.36	48.00	55.00	50.40	50.00	62.66	76.27	40.28	14.29
	Llama-2-Ins	49.32	49.36	48.00	56.85	50.05	50.00	76.38	87.68	49.94	41.14
	Falcon-Ins	49.32	49.36	48.00	56.05	49.95	50.00	64.29	74.63	45.93	14.29
11B	T5	44.20	49.36	48.08	25.60	49.60	49.85	51.58	51.64	49.60	14.29
	Flan-T5-XXL	50.68	2.20	3.51	73.30	59.85	62.45	58.19	67.97	50.56	14.29
13B	Llama-2	49.32	49.36	48.00	54.10	46.75	50.00	80.06	88.36	48.25	17.43
	Llama-2-Ins	49.32	49.36	48.00	55.35	44.45	50.00	80.40	88.98	52.54	21.81
40B	Falcon	49.32	49.36	-	48.70	49.95	50.00	68.87	81.02	43.90	-
	Falcon-Ins	49.32	49.36	-	53.65	50.00	50.00	69.77	81.41	48.08	-
70B	Llama-2	49.32	49.36	48.00	55.90	47.55	50.00	80.51	90.28	47.80	17.95
	Llama-2-Ins	49.32	49.36	48.00	50.30	47.90	50.00	82.71	90.45	49.94	16.60
175B	GPT-3.5	80.20	58.18	62.77	76.55	70.25	55.50	92.88	96.84	73.05	48.86

Table 4: Results (accuracy) for all tasks on 0 shot Cloze for Implicature. The task numbers are as mentioned in Figure 1.

#Params	Model name	Task Number									
		1	2	3	4	5	6	7	8	9	10
-	Human	90.85	74.00	79.67	93.67	95.00	97.00	92.33	96.33	91.33	57.91
-	Baseline	50.68	48.54	48.00	25.85	50.00	50.00	50.45	50.45	50.45	78.09
60M	Flan-T5-small	50.69	37.77	31.34	83.74	50.00	50.46	50.46	50.57	76.23	13.45
220M	Flan-T5-base	50.93	64.36	56.64	76.97	54.14	62.69	62.69	70.97	60.46	54.32
770M	Flan-T5-large	69.44	54.22	50.20	40.66	64.04	73.43	73.43	77.09	49.89	62.97
1.3B	Phi-1	0.00	0.00	0.00	31.77	0.00	0.00	0.00	0.00	52.51	0.91
	Phi-1.5	50.69	40.20	40.78	0.00	50.00	53.09	53.09	55.83	15.14	14.12
3.5B	T5	47.38	0.76	0.60	15.51	13.99	27.20	27.20	18.06	14.91	39.77
	Flan-T5-XL	57.86	83.19	52.05	82.63	72.98	91.43	91.43	97.03	74.97	65.51
7B	Llama-2	60.77	25.91	33.02	37.02	50.00	61.66	61.66	70.40	59.54	13.50
	Falcon	54.91	14.15	15.40	27.83	49.19	49.54	18.51	32.86	52.63	18.29
	Llama-2-Ins	77.26	45.17	55.31	0.00	51.21	79.09	79.09	90.17	0.00	10.57
	Falcon-Ins	60.08	28.56	36.12	64.60	49.49	52.74	52.74	54.46	53.14	10.71
11B	T5	44.48	0.00	0.08	15.51	1.67	14.57	14.57	14.97	14.91	2.31
	Flan-T5-XXL	62.02	87.01	70.19	57.27	75.66	93.14	93.14	98.06	61.60	64.12
13B	Llama-2	72.14	39.34	64.08	63.89	53.13	81.37	81.37	92.17	59.89	23.87
	Llama-2-Ins	75.44	53.82	67.70	75.91	58.13	85.89	85.89	95.66	72.91	33.67
40B	Falcon	-	-	-	-	-	-	-	-	-	-
	Falcon-Ins	50.77	2.45	11.50	1.72	49.85	86.23	86.23	96.23	32.06	-
70B	Llama-2	84.56	63.19	78.56	71.52	71.31	94.00	94.00	98.34	67.89	54.32
	Llama-2-Ins	78.43	73.89	82.02	36.46	65.10	91.71	91.71	97.37	50.40	51.54
175B	GPT-3.5	73.87	43.81	53.02	78.13	71.01	54.85	93.03	97.94	71.43	32.52

Table 5: Results (accuracy) for all tasks on 3 shot MCQA for Implicature. The task numbers are as mentioned in Figure 1.

#Params	Model name	Task Number									
		1	2	3	4	5	6	7	8	9	10
-	Human	90.85	74.00	79.67	93.67	95.00	97.00	92.33	96.33	91.33	57.91
-	Baseline	50.68	48.54	48.00	25.85	50.00	50.00	50.45	50.45	50.45	78.09
60M	Flan-T5-small	50.69	37.77	31.34	83.74	50.00	50.46	50.46	50.57	76.23	13.45
220M	Flan-T5-base	50.93	64.36	56.64	76.97	54.14	62.69	62.69	70.97	60.46	54.32
770M	Flan-T5-large	69.44	54.22	50.20	40.66	64.04	73.43	73.43	77.09	49.89	62.97
1.3B	Phi-1	0.00	0.00	0.00	31.77	0.00	0.00	0.00	0.00	52.51	0.91
	Phi-1.5	50.69	40.20	40.78	0.00	50.00	53.09	53.09	55.83	15.14	14.12
3.5B	T5	47.38	0.76	0.60	15.51	13.99	27.20	27.20	18.06	14.91	39.77
	Flan-T5-XL	57.86	83.19	52.05	82.63	72.98	91.43	91.43	97.03	74.97	65.51
7B	Llama-2	60.77	25.91	33.02	37.02	50.00	61.66	61.66	70.40	59.54	13.50
	Falcon	54.91	14.15	15.40	27.83	49.19	49.54	18.51	32.86	52.63	18.29
	Llama-2-Ins	77.26	45.17	55.31	0.00	51.21	79.09	79.09	90.17	0.00	10.57
	Falcon-Ins	60.08	28.56	36.12	64.60	49.49	52.74	52.74	54.46	53.14	10.71
11B	T5	44.48	0.00	0.08	15.51	1.67	14.57	14.57	14.97	14.91	2.31
	Flan-T5-XXL	62.02	87.01	70.19	57.27	75.66	93.14	93.14	98.06	61.60	64.12
13B	Llama-2	72.14	39.34	64.08	63.89	53.13	81.37	81.37	92.17	59.89	23.87
	Llama-2-Ins	75.44	53.82	67.70	75.91	58.13	85.89	85.89	95.66	72.91	33.67
40B	Falcon	-	-	-	-	-	-	-	-	-	-
	Falcon-Ins	50.77	2.45	11.50	1.72	49.85	86.23	86.23	96.23	32.06	-
70B	Llama-2	84.56	63.19	78.56	71.52	71.31	94.00	94.00	98.34	67.89	54.32
	Llama-2-Ins	78.43	73.89	82.02	36.46	65.10	91.71	91.71	97.37	50.40	51.54
175B	GPT-3.5	73.87	43.81	53.02	78.13	71.01	54.85	93.03	97.94	71.43	32.52

Table 6: Results (accuracy) for all tasks on 3 shot Cloze for Implicature. The task numbers are as mentioned in Figure 1.

Models	14 - 0-shot	14 - 3-shot	3 - 0-shot	3 - 3-shot	10 - 0-shot	10 - 3-shot
Flan-T5-small	0.79	0.79	0.62	0.71	0.45	0.35
Flan-T5-base	0.92	0.86	0.89	0.81	0.82	0.59
Flan-T5-large	0.94	0.79	0.9	0.92	0.85	0.87
Phi-1	0.29	0.31	0.29	0.28	0.33	0.36
Phi-1.5	0.68	0.79	0.64	0.5	0.8	0.5
T5-3B	0.26	0.31	0.29	0.27	0.34	0.36
Flan-T5-XL	0.96	0.96	0.93	0.93	0.92	0.92
Llama-2-7B	0.51	0.79	0.26	0.58	0.66	0.53
Llama-2-7B-Ins	0.74	0.81	0.58	0.55	0.44	0.45
Falcon-7B-Ins	0.25	0.29	0.25	0.36	0.33	0.36
Falcon-7B	0.3	0.3	0.38	0.32	0.45	0.34
Llama-2-70B	0.84	0.94	0.75	0.86	0.48	0.83
Llama-2-70B-Ins	0.88	0.87	0.73	0.78	0.89	0.74
Llama-2-13B	0.83	0.83	0.57	0.68	0.54	0.7
Llama-2-13B-Ins	0.84	0.82	0.69	0.68	0.74	0.62
Flan-T5-XXL	0.98	0.92	0.94	0.95	0.95	0.93
T5-11B	0.27	0.26	0.25	0.25	0.39	0.34

Table 7: PPA across 3 tasks 0-shot and 3-shot. The task numbers are as mentioned in Figure 1.

D 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 3: 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 4: 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 5: 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 6: 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 7: 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 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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 14: 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 15: 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 16: Prompt example for Task 15