Failures of Scale in Large Language Models and Metric Pitfalls for Implicature Understanding

Anonymous ACL submission

Abstract

This paper presents a comprehensive zero-shot evaluation of LLMs to assess their comprehension of general-purpose implicature using the Pragmatic Understanding Benchmark and the Conversational Implicature dataset. [1] We focus on general implicature understanding, deliberately excluding specialized phenomena such as sarcasm and indirect response detection. Our findings reveal that LLMs with over 50 billion parameters, despite excelling at causal reasoning, struggle with implicature, often producing uninformative predictions similar to smaller models. In contrast, the 11 billion parameter FlanT5, fine-tuned on semantic tasks, outperformed larger models, suggesting benefits of semantic-pragmatic transfer learning.

By evaluating implicature as a generative task, we highlight inadequacies in current datasets and metrics for measuring deeper comprehension. Widely adopted metrics like METEOR and BERTScore struggle to assess deeper pragmatic meaning, while ROUGE offers more reliable evaluation for implicature tasks.

Our research reveals critical limitations in LLMs' understanding of implicature and highlighting the need for future improvements in task evaluation.

1 Introduction

Pragmatics (Grice, 1975), delves into the ways speakers convey meaning beyond the literal content of their words, focusing on how context, conversational maxims, and shared assumptions guide interpretation. Grice introduced the theory of implicature through his concept of cooperative principles, where meaning is implied rather than explicitly stated, relying on maxims such as relevance, quality, quantity, and manner. These implicatures play a critical role in human communication, as much of the intended meaning hinges on the listener's ability to infer what is left unsaid, guided by these maxims.

Recent advances in LLMs, such as Mixtral (Jiang et al., 2024), FlanT5 (Chung et al., 2022), LLaMA (Touvron et al., 2023a; Dubey et al., 2024b), and ChatGPT (Brown et al., 2020; OpenAI et al., 2024), have improved reasoning in natural language understanding, problem-solving, and creative writing. While these models excel in semantic tasks like natural language inference and causal reasoning, they struggle with new benchmarks evaluating indirect meaning and pragmatic understanding (White et al., 2024; Fourrier et al., 2024) as well as inferencing (Valmeekam et al., 2022), highlighting the need for further advancements in implicatures and pragmatic reasoning.

Implicature relies on context and indirect meaning, presents a significant challenge for LLMs focused on explicit content, as highlighted by the Pragmatic Understanding Benchmark (PUB) (Sravanthi et al., 2024a). Li (2022) observe that lexical characteristics correlate with conversational implicatures and not context while classification. While including narrow tasks like sarcasm and agreement detection, PUB highlights a performance gap between LLMs and humans in implicature among other pragmatic phenomena. Similarly, SOCKET (Choi et al., 2023) evaluates social understanding but limits its scope to classification, regression, and span detection, where models struggle with open-ended tasks. Motivated by these gaps, we study datasets and evaluation metrics that capture implicature understanding in detail.

Our main contributions are as follows:

- **Comprehensive Evaluation of Implicature Understanding in LLMs**: We evaluate state-of-the-art LLMs on three diverse implicature tasks namely Conversational Implicature dataset (George and Mamidi, 2020), Implicature Recovery and NLI Implicature (Sravanthi

[1] All code will be released.

et al., 2024b).

- **Generation Task Challenges and Evaluation Metrics**: By framing implicature understanding as a generation task, we reveal model shortcomings that traditional classification-based approaches overlooked, indicating the need for more sophisticated objectives for nuanced reasoning required in pragmatic inference. Additionally we highlight drawbacks of word and representation based evaluation metrics.

Our main findings are as follows:

- **Scale does not guarantee implicature Understanding**: Our analysis shows that larger model size does not guarantee better performance in implicature tasks.
- **Limitations of Existing Metrics in Evaluating Implicature Generation**: Existing metrics like ROUGE, METEOR and BERTScore fail to capture the nuanced reasoning in implicature, underscoring the need for better methods to evaluate implicature.
- **LLMs lack informativeness in implicatures**: Even with the classification objective (that leaks the implicature in the input), LLMs struggle to produce meaningful inferences.

We conclude that **implicature understanding remains an unresolved issue for LLMs**, as even best models like GPT-4 and LLaMA-3 demonstrate weaknesses when evaluated with generation.

2 Related Work

2.1 Implicature

Implicature refers to the phenomenon where an utterance conveys meaning beyond its literal or propositional content, meaning that is not explicitly encoded in the words themselves. This deeper meaning emerges from context and allows speakers to communicate more than what is directly said.

Grice (1975), distinguishes between two types of implicature: conversational and conventional. Conversational implicatures arise from the cooperative principle, where speakers assume that communication is a collaborative effort, governed by four conversational maxims — quality, quantity, relevance, and manner. In contrast, conventional implicatures possess the same meanings regardless of context (e.g., "*but*" in "He is smart *but* lazy").

2.1.1 Conversational Implicature

Conversational implicatures are inherently context-sensitive i.e. interpretation relies on the conversation (unlike conventional implicatures, which have fixed meanings). Grice identified key types of conversational implicature. *Scalar implicatures* arise when a speaker uses a term on a scale (e.g., "some"), implying the negation of a stronger term (e.g., "all"). *Generalized implicatures* occur in most contexts with minimal background (e.g., "John started a new job" implying it was recent), whereas *particularized implicatures* depend heavily on the specific context (e.g., "The lights are off" implying the boss is not present). *Relevance implicatures* assume that utterances are contextually relevant, allowing for inferred meanings.

2.2 Datasets

This paper evaluates LLMs on three conversational implicature datasets (see Figure 1 for examples).

- **Conversational Implicature (CI)**: The Conversational Implicature Dataset (George and Mamidi, 2020) consists of **1000 dialogue instances**, formatted as triplets: `<context, utterance, implicature>`. Sourced from TOEFL listening comprehension tapescripts and movie scripts from the IMSDb database, a dialogue exchange is between **two speakers**, with **one turn** each (*context* and a *response* utterances). The **implicature** is a sentence that captures a meaning not explicitly stated in the conversation. The dataset focuses on indirect responses involving **figurative language** and **polar questions**.

  Srivastava et al. (2023) reduced CI Dataset (CID) from 1,000 to 563 samples and transformed it into a binary classification task to mitigate challenges of generation evaluation. However, we evaluate the generated outputs across all 1,000 samples.

  SOCKET (Choi et al., 2023) evaluates social understanding via classification, regression and span detection. Models perform worst on span detection because it is open-ended. Addressing this, our study marks the first evaluation with an objective to generate implicatures from conversational context, shifting from classification that leaks implicatures by embedding them in input.

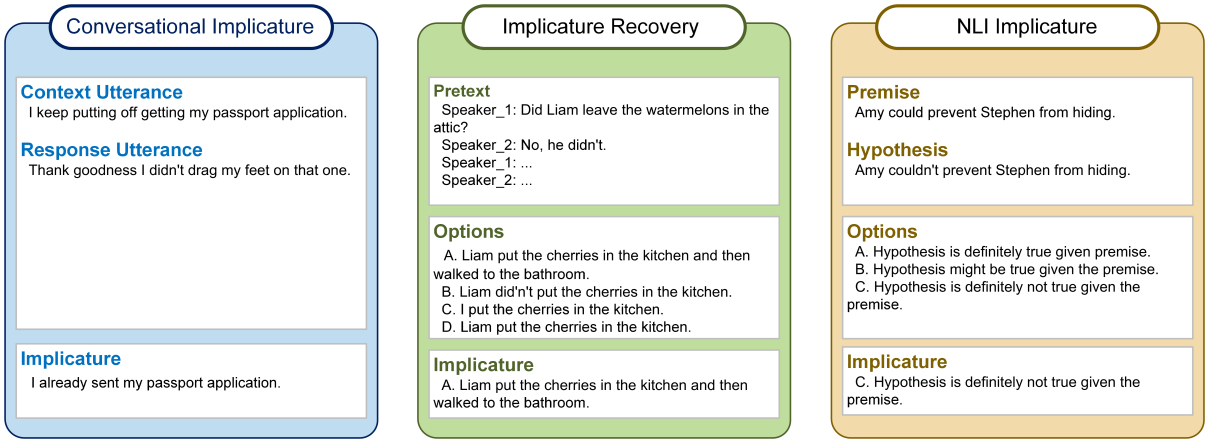


Figure 1: Examples illustrating instances from implicature datasets: Conversational Implicature (CI) Dataset (George and Mamidi, 2020), Implicature Recovery (IR), and NLI Implicature (NI) (Sravanthi et al., 2024a)

- **Implicature Recovery (IR)**: The Implicature Recovery Task (Task 4) from the PUB dataset (Sravanthi et al., 2024a), sourced from the GRICE dataset (Zheng et al., 2021), consists of **2000 dialogue instances**. These dialogues capture exchanges between two speakers discussing locations or actions associated with various objects in different settings. *Dialogues* are provided as *pretext*, followed by *multiple-choice options*, where the correct answer is determined through conversational implicature. These **multi-turn dialogues** evaluate situation modeling via tracking agents and understanding actions related to object movement.

  The dataset is simplistic, as Zheng et al. (2021) construct pretexts from **manually defined templates**, restricting conversational context, subtopics, utterance types, speech templates, and named entities to a static set. This may account for the **high human accuracy** reported by Sravanthi et al. (2024b), despite being atypically different from natural conversations.

- **NLI Implicature (NI)**: The NLI Implicature Task (Task 10) from the PUB dataset (Sravanthi et al., 2024b), sourced from the IMPRESS dataset (Jeretic et al., 2020), comprises **2100 instances** of **natural language inference (NLI) tasks**. Each instance includes a *premise* and a *hypothesis*, along with a set of *multiple-choice options* designed to assess the logical relationship between the premise and the hypothesis. The objective is to determine whether the hypothesis is true, false, or indeterminate based on the given premise. This task evaluates a model's capability to handle **implicature recovery** within the context of NLI.

3 Methodology

3.1 Dataset Evaluation

In lieu with Sravanthi et al. (2024b), we assess IR and NI as zero-shot classification. Recognizing the challenges of evaluating social understanding tasks through span detection (Choi et al., 2023), we evaluate CI with a text-to-text generation objective.

- **Conversational Implicature (CI)**: We evaluate implicature understanding with conversations from CID with a previously unstudied text-to-text objective. The models were prompted to generate implicatures based on two-turn conversations (context and response utterances). The generated outputs are compared to gold-standard implicatures in the dataset using evaluation metrics reported in Section 3.2.1.

- **Implicature Recovery (IR)**: Following Sravanthi et al. (2024b), we evaluate implicature understanding in a zero-shot setting. The model selects the correct label from four conversational implicatures (*options*) based on a multi-turn conversation (*pretext*). As the benchmark did not provide prompts, we present our prompts in Appendix A.1. The accuracy of LLaMA-2 and ChatGPT-3.5 are reported from Sravanthi et al. (2024b).

- **NLI Implicature (NI)**: Similar to Sravanthi et al. (2024b), we evaluate the ability of LLMs

to derive implicature within the framework of Natural Language Inference (NLI). The task involves determining whether a given hypothesis is entailed by, contradicts, or is neutral with respect to a premise. Similar to IR, we conduct zero-shot evaluation (Appendix A.1 provides the prompts). We report the accuracy of LLaMA-2 and ChatGPT-3.5 from Sravanthi et al. (2024b).

3.2 Models

Our study seeks to evaluate the performance of current state-of-the-art LLMs that have demonstrated reasoning capabilities.GPT-4 model has been tested on variety of datasets which cover contextual understanding in long texts, natural language inferencing and a diverse range of problems. (Liu et al., 2020, 2021; Saha et al., 2020; Yanaka et al., 2019b,a; Joshi et al., 2020; Williams et al., 2018). Building on the findings of Liu et al. (2023a) and Petruzzellis et al. (2024) on GPTs' and LLaMAs' excellent performance in reasoning ability and NLI tasks, we evaluate implicature understanding which involves reasoning, logical and inferencing. A brief outline of mode sizes and pre-training tasks can be found in Table 1. The models selected for this study are outlined below:

- **LLaMA-2**: LLaMA-2 (Touvron et al., 2023a) improves upon the previous generation with more diverse training dataset and architectural enhancements for better generalization and reasoning. It is trained on public dataset - CommonCrawl, C4, GitHub, arXiv, StackExchange, etc. (Touvron et al., 2023b).

- **LLaMA-3**: LLama-3.1 (Dubey et al., 2024b) improves on earlier models with enhanced scale, multimodal capabilities, and better contextual understanding for more complex language tasks. It builds on the core datasets as LLaMA-2, with additional training on multimodal and multilingual datasets - LAION and mC4 (Dubey et al., 2024a).

- **GPT-3.5**: GPT-3.5 (Brown et al., 2020) enhances its predecessors with improved language understanding and generation, offering more accurate and contextually aware responses.

- **GPT-4**: GPT-4 (OpenAI et al., 2024) excels in zero-shot complex problem-solving and has been evaluated on academic and professional benchmarks along with reasoning datasets like HellaSwag, DROP, etc. These are closely related to implicatures involving narrative coherence, common sense and discrete reasoning, extracting information from long texts.

- **Flan-T5-XXL**: Sravanthi et al. (2024a) report that Flan-T5-XXL outperforms models on pragmatic tasks. It builds on the T5 architecture by focusing on instruction-tuning, which enables task generalization with natural language instructions (Chung et al., 2022).

3.2.1 Metrics

For evaluating implicature, we select metrics suited for paraphrasing and summarization, where exact matches between generated and reference texts are unnecessary. **ROUGE** (Lin, 2004) measures n-gram overlap with a focus on recall, while **METEOR** (Lavie and Agarwal, 2007) incorporates synonyms and stemming, balancing precision, recall, and alignment. **BERTScore** (Zhang et al., 2019) goes beyond surface-level overlap by assessing semantic similarity using BERT embeddings. Zhang et al. (2020) employ ROUGE for open-ended intent generation, a task akin to implicature, Fabbri et al. (2021) suggest pre-trained model-based metrics are ideal for summarization, with Shen et al. (2021) noting METEOR's superiority over ROUGE-L and BLEU.

3.3 Experimental Setup

Our experiments were conducted on multiple platforms - Google Colaboratory, Kaggle Notebooks, and a local system. On Google Colab, we utilized a T4 GPU with 15 GB of GPU memory and 12.7 GB of RAM. Kaggle Notebooks were configured with two T4 GPUs, each with 15 GiB of memory, 29 GiB CPU and 57.6 GiB of disk space. For local setup, we used Nvidia GeForce RTX 4050 GPU with 6 GB memory, 16 GB RAM, 512 GB SSD and 12th Gen Intel Core i7 processor.

For LLaMA 3 models, we utilized Groq AI [2] platform's API. For experiments with ChatGPT, we utilized the GPT-API to make calls to GPT-4 with temperature settings at 0.7 and a batch request of 10-20 queries per minute with a sleep time of 60 seconds. [3] For FlanT5, we used two cores of

[2] http://groq.com

[3] All experiments were conducted from 15th August 2024 to 10th September 2024.

Model	#params	Categories of Pretraining Tasks
llama-3-8b	8B	General knowledge, Mathematics and Reasoning, Code, Multilingual
flan-t5-xxl	11B	NLP tasks, Muffin, CoT, Natural Instructions
llama-3-70b	70B	General knowledge, Mathematics and Reasoning, Code, Multilingual
ChatGPT-4	1.8T	General Knowledge, Conversational Understanding, Reasoning, Multimodal tasks

Table 1: Overview of model scales and pre-training data

NVIDIA GeForce RTX 4090. The prompts used for models and task under consideration have been provided in Appendix A.1.

4 Results

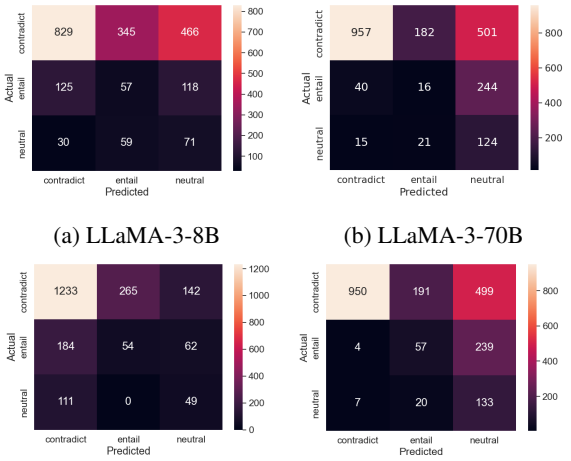
Table 3 benchmarks LLMs against humans. The results reveal a significant gap in IR performance, with FlanT5-XXL achieving an accuracy of 84.65% compared to 93.67% for humans. In contrast for NI, FlanT5-XXL scored 63.61%, while human performance was lower at 57.91%, corroborating the results of Sravanthi et al. (2024a). This suggests that NI may be a solved problem, unlike IR.

This is surprising given the simplistic nature of IR sentences. Zheng et al. (2021) create conversation pretexts using manual speech templates, limiting conversational context, subtopics, utterance types, speech templates and named entities to a static set. This may explain the high human accuracy in IR. While humans often disagree in making inferences (Pavlick and Kwiatkowski, 2019), the unexpectedly excellent human implicature for IR suggests that the dataset lacks complexity. In the future, we aim to investigate more intricate samples that better reflect real-world human conversations, allowing for a more thorough and nuanced study of implicature.

LLaMA-3 underperforming LLaMA-2 at NI, is likely due to prompt modifications or training data that may not have adequately prioritized inferential reasoning tasks. Overall, LLAMA models perform poorer than other models. A factor contributing to this may be pre-training that does not prioritize context retention and multi-turn dialogue understanding as effectively as FlanT5.

Table 1 outlines the model pre-training data and relative sizes. FlanT5 has 4 main categories of pre-training tasks, of which two are **Muffin** (covering mathematical and reasoning problems, dialog data (Byrne et al., 2019; Anantha et al., 2021; Dai et al., 2022) and program synthesis (Yasunaga and Liang, 2020; Li et al., 2022)) and **CoT** Wei et al. (2023) (focused on arithmetic reasoning, common sense reasoning, explanations, sentence composition, and implicit reasoning). Approximately 15.44% of pre-training datasets are inferential reasoning, suggesting that inference based pre-training enhances implicature understanding. Furthermore, the shared NLI input format between FlanT5's pre-training tasks and NI facilitates semantic-pragmatic transfer.

As of August 2024, LiveBench (White et al., 2024) reports ChatGPT-o1-preview [4] as the state-of-the-art across a range of reasoning tasks, scoring 72.66% and 68.00% on complex linguistic tasks unsolved by earlier models (Jiang et al., 2023). Huggingface's OpenLLM Leaderboard 2 (Fourrier et al., 2024) showcases models with superior reasoning, yet our findings indicate that even well-performing models like GPT-4 and LLaMA-3 cannot deduce implicatures. While NI may be handled well, the complexity of IR and CI remains an open problem (Table 3). Valmeekam et al. (2022) show that models like GPT-3 (davinci), Instruct-GPT3 and BLOOM have limited reasoning and planning capabilities, which we also extend to implicature tasks.



(a) LLaMA-3-8B (b) LLaMA-3-70B

(c) Flan-T5 XXL (d) GPT-4

Figure 2: Comparison of class-wise NI predictions

[4] https://openai.com/index/introducing-openai-o1-preview

Models	ROUGE (1,2,L)			METEOR	BERTScore
llama-3-8B	.058	.015	.053	.173	.883
flan-t5-xxl	**0.408**	**0.279**	**0.399**	**0.376**	**.907**
llama-3-70B	.133	.033	.128	.236	.889
ChatGPT-4	.080	.028	.073	.172	.887

Table 2: Average scores for ROUGE(1, 2, L), METEOR and BERTScore for CID (George and Mamidi, 2020)

Accuracy	IR	NI
Majority	25.85	57.91
llama-2-7b	*56.26	*49.29
llama-3-8b	69.85	45.57
flan-t5-xxl	**84.65**	**63.61**
llama-2-70b	*75.91	*55.09
llama-3-70b	78.85	52.24
ChatGPT-3.5	*78.13	*48.86
ChatGPT-4	76.5	54.28
Human	*93.67	*57.91

Table 3: Results for various zero shot models for the PUB implicature tasks. Numbers indicated with an asterisk are taken from Sravanthi et al. (2024b)

4.1 Analysis

4.1.1 NI Predictions

While Table 3 indicates that for NI, FlanT5-XXL outperforms human participants, note that NI comprises simplistic statements from straightforward conversations. Furthermore, the majority baseline is quite high at 57.91, which helps put the results in context.

Figure 2 shows the classwise performance of FlanT5-XXL and the latest LLMs for NI. One major observation is the dataset skew, where the *Contradiction* (1640) class significantly outnumbers the *Entailment* (300) and *Neutral* (160) classes. The skew towards the contradictory class is apparent in the majority predictor's high accuracy in Table 3.

The **LLaMA-2-70B** model shows a tendency to produce many **False Negatives** for the Neutral class. This behavior suggests the model struggles to generate strong, informative implication judgments.

LLaMA-3-8B handles entailment implicatures slightly better than LLaMA-2 70B but still struggles with contradictory and neutral implicatures resulting in a drop in accuracy.

Both the **LLaMA-3 8B** and **LLaMA-2 70B** models exhibit overprediction of the neutral class, leading to lower overall performance which are **uninformative**.

Interestingly, **LLaMA-3 8B** performs worse than **LLaMA-2 70B**, likely due to differences in prompt structures. We observed minor changes to the prompt structure lead to great different results which is also noted by Salinas and Morstatter (2024). Sravanthi et al. (2024a) also point out sensitivity to prompt templates. Future research could explore prompt engineering to better understand this discrepancy.

4.1.2 Shortcomings of Evaluation Metrics

Existing metrics face challenges in evaluating tasks without direct content overlap, such as summarization and paraphrasing, in contrast to machine translation. While Moramarco et al. (2022) find metrics like Levenshtein distance, METEOR, and BERTScore useful for specific clinical post-editing errors, Novikova et al. (2017) emphasize that metric performance varies by dataset and system, often diverging from human ratings. Corpus-similarity metrics, including BLEU and ROUGE, correlate poorly with human evaluations, particularly when texts prioritize variety and readability, as noted by Belz and Reiter (2006). Although embedding-based metrics like BERTScore and BLEURT show improved alignment with human ratings, their Pearson correlation remains low, indicating significant room for improvement (Clinciu et al., 2021).

4.1.3 Evaluation metrics for Conversational Implicatures

Table 2 shows that across all NLG metrics, LLaMA-3-70B outperforms ChatGPT-4 on CID. Larger models like ChatGPT-4 and LLaMA-3-70B score 0.172 and 0.236 on METEOR, respectively, which are still low. In contrast, FlanT5 scores 0.376, reflecting a notable improvement while still being low. These low scores indicate that implicature remains a challenge when posed as generation, while classification tasks often leak answers directly within the input, leading to higher accuracy.

A notable discrepancy exists between METEOR and other metrics. METEOR erroneously ranks LLaMA-3-8B higher than ChatGPT-4, while clas-

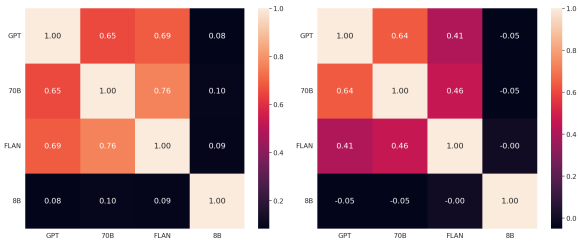
sification (IR & NI) consistently indicates that LLaMA-3-70B is the poorest performer. Surprisingly, n-gram-based ROUGE-1, ROUGE-2, and ROUGE-L metrics perform better in capturing these differences than METEOR, designed to account linguistic transformations.

The ROUGE and METEOR scores for CI and accuracy for IR and NI reveal a significant gap between LLaMA-3-70B and ChatGPT-4, a gap not reflected in BERTScore. It's narrow range of similar scores across all models reported for CI. We note that it often rates high scores for incorrect responses, due to the overlapping content and implicatures.

These findings underscore the necessity for a metric that can accurately evaluate the understanding of conversational implicature.

4.1.4 Correlations for Implicature as Classification

We compute the Matthews Correlation Coefficient (MCC) between the predictions for various models to analyze the correlations of their implicature (IR and NI) distributions. MCC ranges from -1 to 1, where 1 indicates perfect agreement, 0 signifies random prediction, and -1 denotes total disagreement.



(a) IR (b) NI

Figure 3: MCC across tested models for Implicature (IR & NI)

In IR (Fig 3a), correlations among the models are generally stronger. The highest correlation is observed between FlanT5 and LLaMA-3-70B, at 0.76, which points to a strong alignment in their performance on the IR task. Similarly, ChatGPT-4 and FlanT5 exhibit a strong positive correlation of 0.69, showing that these models share more similarities in IR compared to their behavior in NI tasks. Although llama 8B still demonstrates lower correlations with the other models, its relationships with ChatGPT-4, LLaMA-3-70B, and FlanT5 show slight improvement (e.g., 0.08 between ChatGPT-4 and LLaMA-3-8B), suggesting more alignment on the IR task than on NI.

In NI (Fig 3b), correlations among the models vary notably. ChatGPT-4 and LLaMA-3-70B exhibit a moderate positive correlation of 0.64, indicating some overlap in their performance on NI tasks. However, the correlation between ChatGPT-4 and FlanT5 is lower, at 0.41, suggesting that these two models perform more differently on NI.

Strikingly, LLaMA-3-8B shows very low or slightly negative correlations with ChatGPT-4 and LLaMA-3-70B (both at -0.05), indicating that LLaMA-3-8B behaves in a distinct manner for both tasks.

As noted in Section 4, models outperform humans in NI but not in IR. This may be due to the difficulty of tracking longer, repetitive conversations with shifting state information, which makes its conversations atypical of the conversational data used for pre-training.

5 Conclusion

Our key contributions and findings are as follows:

- **Comprehensive Evaluation of Implicature Understanding in LLMs**

  We conduct a thorough evaluation of recent models (FlanT5-XXL, LLaMA-3-8B, LLaMA-3-70B and GPT-4) against previously reported models (LLaMA-2-7B and LLaMA-2-70B) on three broad implicature tasks: Conversational Implicature, Implicature Recovery and NLI Implicature.

- **Broadening the Scope of Implicature Datasets**

  We bring together a more diverse set of datasets, including those from the Pragmatic Understanding Benchmark (Sravanthi et al., 2024a) and the Conversational Implicature Dataset (George and Mamidi, 2020), moving beyond narrow, binary definitions of implicature tasks like sarcasm detection. Our findings (Table 2 & 3) show that models perform poorly on implicature, except for the simplistic NI task.

- **Generation Task Challenges and Evaluation Metrics**

  We frame implicature as a generation task, contrasting with earlier works that posed it as a classification (Srivastava et al., 2023; Sravanthi et al., 2024b). While Srivastava et al.

(2023) noted that binary framing aids evaluation, our generation objective, free from input leakage, reveals poor metric performance that classification objectives —despite their high accuracy — fail to capture. Traditional metrics like ROUGE and METEOR, which focus on surface-level word overlap, are inadequate for nuanced logical reasoning. BERTScore also falls short, yielding a narrow range. Thus, both n-gram-based and word embedding-based metrics prove inconsistent and insufficient for assessing content that involves reasoning and paraphrasing.

- **Scale vs. Instruction Fine-tuning**

  Scale alone does not guarantee performance in implicature tasks. FlanT5-XXL outperforms larger models, a trend also observed in pragmatic tasks (Sravanthi et al., 2024b; Choi et al., 2023). Instruction-based fine-tuning, facilitates better transfer of semantic understanding to pragmatic reasoning tasks, suggesting further investigation into why smaller models excel in these areas. Additionally, LLaMA-3 models struggle with generating strong implicatures (e.g., non-neutral ones) and often produce less informative outputs.

- **Informativeness and Quality of Implicature Predictions**

  The distributions of the model-generated implicatures, reveal their informativeness and quality. Our evaluation highlights twhen LLMs are poor at implicature they are not making inferences but are merely generating neutral, uninformative outputs.

- **Implicature is Not a Solved Problem**

  Implicature understanding remains unsolved, as the generation objective uncovers poor metric performance overlooked by classification tasks, despite their high accuracy. Classification prediction's deep dive reveals lack of informativeness of model predictions.

  The performance of state-of-the-art models - GPT-4 and LLaMA-3, while strong in many other reasoning tasks, are fundamentally weak in implicature.

6 Limitations and Future Work

We focused on zero-shot performance however, few-shot or fine-tuning could offer insights into semantic and pragmatic transfer for implicature, which we plan to investigate in future work.

The model-human gap in IR, highlights that simplistic datasets due lengthy sentences and the atypical conversations do not work well for model evaluation, underscoring the need for more natural datasets which we plan to do in the future.

This study focuses on evaluating the impact of scale across publicly accessible foundational models without comparing chat models (e.g., LLaMA-3-8B-Chat) with their base versions. LLaMA-2 models were excluded from the detailed analysis.

G-Eval (Liu et al., 2023b), an advanced evaluation metric, was not used due to cost constraints but could be valuable to assess conversational nuances.

The lack of human predictions for IR and NI released by Sravanthi et al. (2024b) limited our ability to compare human and model implicatures and analyze informativeness and distributional variations. We find this interesting because differences in human and model-generated implicature could yield valuable insights.

7 Challenges

The absence of prompts from prior work necessitated significant effort in prompt design, as a single prompt across all models proved impractical. Minor prompt changes caused considerable variations in responses, particularly for smaller models as noted by Sravanthi et al. (2024a). LLaMA-3-7B required highly specific instructions for accurate outputs, while prompts effective for larger models often resulted in poor recall or nonsensical language for this model. Additionally, the lack of public APIs for the larger LLaMA-3-350B limited our ability to conduct thorough evaluations across various model scales.

8 Ethical Considerations

This work evaluates pragmatic reasoning through zero-shot experiments on conversational implicature. We utilized public datasets focusing on generic implicature. No new data was collected or generated. We do not verify or report the datasets for bias. LLMs may produce biased or harmful outputs due to their training data. Before applying LLMs to real-time or domain-specific data, a more extensive analysis is necessary.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

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 Zhao, Yanping Huang, Andrew 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. *arXiv preprint*.

Miruna Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. *arXiv preprint arXiv:2103.08545*.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

A Dubey et al. 2024a. Llama-3: Training datasets and benchmarks. `https://ai.meta.com/llama3`. Accessed: 2024-09-26.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon

Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open

llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323.

HP Grice. 1975. Logic and conversation. *Cole and Morgarn*.

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*, pages 8690–8705, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore. Association for Computational Linguistics.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231.

Xianbo Li. 2022. Text classification of conversational implicatures based on lexical features. *Applied Artificial Intelligence*, 36(1):2127598.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context - investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13388–13396.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023a. Evaluating the logical reasoning ability of chatgpt and gpt-4.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *Preprint*, arXiv:2007.08124.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik

Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Flavio Petruzzellis, Alberto Testolin, and Alessandro Sperduti. 2024. Assessing the emergent symbolic reasoning abilities of llama large language models.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. Conjnli: Natural language inference over conjunctive sentences. *Preprint*, arXiv:2010.10418.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729*.

Juanjuan Shen, Yu Zhou, Yongchao Wang, Xiang Chen, Tingting Han, and Taolue Chen. 2021. Evaluating code summarization with improved correlation with human assessment. In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, pages 990–1001.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024a. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *Preprint*, arXiv:2401.07078.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024b. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *arXiv preprint arXiv:2401.07078*.

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, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew 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 Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng,

Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, 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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Hugo Touvron et al. 2023b. Llama 2 training datasets and benchmarks. https://ai.facebook.com/blog/large-language-model-llama-2. Accessed: 2024-09-26.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Adina Williams, Nikita Nangia, and Samuel 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, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Michihiro Yasunaga and Percy Liang. 2020. Graph-based, self-supervised program repair from diagnostic feedback. *Preprint*, arXiv:2005.10636.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Query understanding via intent description generation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1823–1832.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

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*, pages 2074–2085, Online. Association for Computational Linguistics.

A Appendix

Minitron papers: 1. https://arxiv.org/pdf/2408.11796 1a. https://www.arxiv.org/pdf/2407.14679

A.1 Prompts For Tasks and Models

Models/Dataset	Prompt Structure
LLaMA-3-8B/CID	"Given the sentences are a part of a conversation: Sentence 1: <Sentence 1> , Sentence 2: <Sentence 2> . In context to the first sentence, what implicature is given by second sentence. In the response only write the implicature and avoid any extra information or repeating the sentences from input prompt"
LLaMA-3-70B/CID	"Given the sentences are a part of a conversation: Sentence 1: <Sentence 1> , Sentence 2: <Sentence 2> . In context to the first sentence, what implicature is given by second sentence. In the response only write the implicature and avoid any extra information or repeating the sentences from input prompt"
GPT-4/CID	"Given are set of 20 different conversations each having a context utterance and a response utterance. For each conversation given below, in context to the context utterance, what implicature is given by the response utterance ? In the response only write the implicature and avoid any extra information or repeating the sentences from input prompt. Keep the response short (atmost 2 lines). Number each response. Always put asterik (*) at end of every conversation response. Conversations go below : Conversation <Number> Context Utterance : <Context> Response Utterance : <Response> ....."
Flan-T5 XXL/CID	"Given the sentences are a part of a conversation: Sentence 1: <Sentence 1> , Sentence 2: <Sentence 2> . In context to the first sentence, what implicature is given by second sentence. In the response only write the implicature and avoid any extra information or repeating the sentences from input prompt"
LLaMA-3-8B/IR	"Given the sentences are a part of a conversation: Pretext: <Pretext> According to the pretext, which of the four options is correct. <4 Options> In the response only write the correct option letter (A, B, C or D) and avoid any extra information."
LLaMA-3-70B/IR	"Given the sentences are a part of a conversation: Pretext: <Pretext> According to the pretext, which of the four options is correct. <4 Options> In the response only write the correct option letter (A, B, C or D) and avoid any extra information."
GPT-4/IR	"Given below are 10 different conversations. Each conversation is followed by 4 options with only one option being correct. Choose the correct option letter only (A, B, C or D) and avoid any extra information. Do not write the complete correct option. Mention only the correct option letter. Don't skip any conversation. Answer all conversations and number them as well. Conversations go as below : Conversation <Number> <Pretext> <4 Options> ....."
Flan-T5 XXL/IR	"Given the sentences are a part of a conversation: Pretext: <Pretext> According to the pretext, which of the four options is correct. Options: <4 Options> In the response only write the correct option letter (A, B, C or D) and avoid any extra information."
LLaMA-3-8B/NI	"Given the premise : <Premise> and the hypothesis. Which of the following options is correct? In the response only write the correct option letter (A, B or C) and avoid any extra information. Options: <3 Options labelled as A, B and C)>
LLaMA-3-70B/NI	"Given the premise : <Premise> and the hypothesis. Which of the following options is correct? In the response only write the correct option letter (A, B or C) and avoid any extra information. Options: <3 Options labelled as A, B and C)>"
GPT-4/NI	"Given are 20 sets of premises and corresponding presented hypothesis. For each given pair of premise and hypothesis, 3 options have been given (A, B and C). For each pair of premise and hypothesis, choose the correct option letter (A, B or C) that correctly comments on trueness of hypothesis given the premise. Number them properly and do not write the correct option completely. Skip any unecessary information. Do not skip any given pair of input. Below goes all pairs : <Pretext Number> <Premise>: <Hypothesis (3 Options labelled A, B and C)> .. .. (For 20 pairs)"
Flan-T5 XXL/NI	"Given the premise : <Premise> and the hypothesis. Which of the following options is correct? In the response only write the correct option letter (A, B or C) and avoid any extra information. Options: <3 Options labelled as A, B and C)>

Table 4: Prompts for generating implicatures for various Models and Datasets (prompts were changed as per individual models and tasks to create meaningful responses, notably for the smaller models)