Beyond Semantics: Improving Pragmatic Understanding in Language Models Through Fine-Tuning and Data Extensions

Sravani Gunnu
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay
sravanigunnu@cse.iitb.ac.in

Abstract

Large Language Models (LLMs) excel in semantic understanding but face significant challenges in pragmatic reasoning, particularly in phenomena like *Implicature*. Benchmarks such as the Pragmatic Understanding Benchmark (PUB) and DiPlomat underscore these limitations, revealing substantial gaps in LLMs' ability to handle nuanced and context-dependent tasks. To address these challenges, this work focuses on creating reasoning-focused datasets tailored for preference-based fine-tuning. By augmenting existing Multiple-Choice Question Answering (MCQA) datasets like CIRCA, FLUTE, FigQA, IMPPRESS, and LUDWIG, a reasoning-enriched dataset of $72k$ annotated examples was generated. Additionally, a synthetic dataset of $37k$ high-quality examples was created using distinct instances from CIRCA and LUDWIG. These resources are designed to enable models to rank and refine reasoning outputs effectively, bridging the gap between semantic understanding and pragmatic reasoning. Experiments were conducted using *zero-shot prompting* and *supervised fine-tuning*, including LoRA-based fine-tuning of the *Gemma-2-2b-it* model on the CIRCA dataset, resulting in a **25%** improvement in accuracy. This work highlights the critical role of reasoning-enriched datasets and task-specific fine-tuning in addressing the limitations of LLMs and paves the way for advancing their pragmatic reasoning capabilities.

1 Introduction

1.1 Problem Statement

Despite the remarkable capabilities of Large Language Models (LLMs) in semantic understanding, they exhibit significant limitations in pragmatic reasoning. Pragmatics involves understanding implied meanings, context, and speaker intentions—critical aspects of human communication. Benchmarks like PUB (Sravanthi et al., 2024) and DiPlomat (Li et al., 2023) demonstrate that LLMs struggle to capture these nuanced linguistic phenomena, such as *Implicature, Presupposition, Reference, and Deixis*. The performance gap between LLMs and humans on these tasks underscores the need for targeted approaches to enhance their pragmatic reasoning capabilities. These challenges motivate efforts to develop and evaluate methods for improving LLMs' understanding of pragmatics.

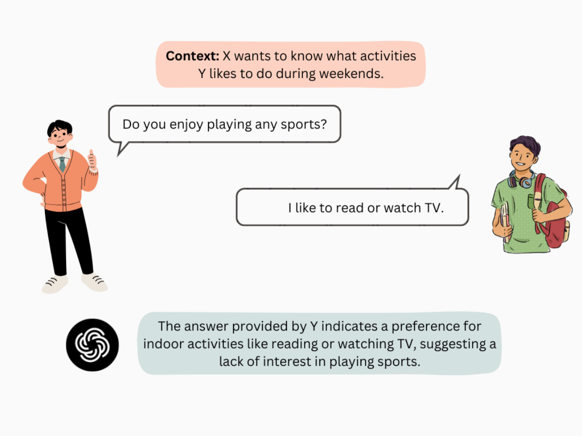

Context: X wants to know what activities Y likes to do during weekends.
Do you enjoy playing any sports?
I like to read or watch TV.
The answer provided by Y indicates a preference for indoor activities like reading or watching TV, suggesting a lack of interest in playing sports.

Figure 1: An example of implicature in a response where the reasoning provided by GPT.

1.1.1 Pragmatics in Language Processing

Pragmatics in language processing extends beyond the literal meaning of words, focusing on how context influences interpretation. As defined by Grice (Zheng et al., 2021), pragmatics examines phenomena like implicature (implied meaning), presupposition (implicit assumptions), and reference (pointing to entities, times, or locations). For example, the statement, *"It's chilly in here,"* could imply a request to close the window rather than merely stating the temperature. These context-dependent interpretations are essential for effective communication and represent a significant challenge for LLMs, which often lack the ability to infer nuanced, non-literal meanings (Jeretic et al., 2020).

Challenges Faced by LLMs in Understanding Pragmatics

LLMs excel in tasks that rely on semantic understanding but struggle with pragmatic reasoning due to their training objectives. Models like GPT-3 and LLaMA are optimized for predicting tokens based on massive text corpora, which often lack the context-rich examples needed to learn pragmatic phenomena (Sravanthi et al., 2024; Li et al., 2023). Key challenges include:

- **Context Dependence:** Pragmatic reasoning requires models to interpret language based on conversational context and speaker intentions.
- **Ambiguity:** Pragmatics often involves resolving ambiguous or incomplete information, a task where LLMs underperform (Zheng et al., 2021).
- **Implicature and Presupposition:** Understanding implied meanings and assumptions remains a significant challenge, as shown in benchmarks like IMPPRES (Jeretic et al., 2020).

These challenges highlight the need for specialized datasets and fine-tuning techniques to bridge the gap between LLMs and human-like understanding. Figure 2 illustrates the performance of various state-of-the-art models, including *Gemma-2-2b-it*, *Phi-2*, *Phi-3.5-mini-instruct*, and *GPT-4o-mini*, across pragmatic reasoning datasets such as FLUTE, FigQA, and IMPPRES. The results highlight significant variability in model performance, emphasizing the need for improved techniques and datasets to enhance their pragmatic understanding.

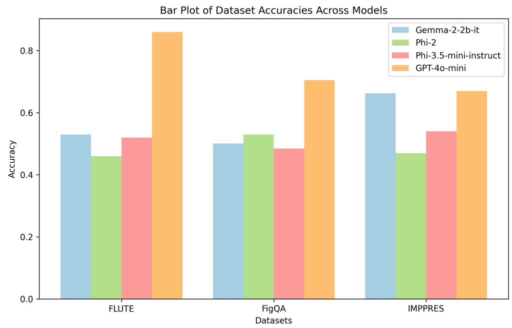


Figure 2: Performance of different models on pragmatic reasoning tasks across datasets. The bar plot highlights performance disparities, demonstrating the need for enhanced fine-tuning and reasoning capabilities in models.

1.2 Motivation

Pragmatics, a critical aspect of human language, enables individuals to derive meaning that goes beyond the literal interpretation of words by incorporating context, speaker intentions, and shared assumptions. For instance, interpreting the statement *Can you pass the salt?* not as a question about one's ability to pass the salt but as a polite request demonstrates the complexity and necessity of pragmatic reasoning in effective communication. This ability is fundamental for human interaction, where the meaning is often implied rather than explicitly stated (Sravanthi et al., 2024; Zheng et al., 2021).

While Large Language Models (LLMs) have shown exceptional progress in tasks involving semantic understanding, they continue to struggle with pragmatic phenomena, including implicature, presupposition, and contextual reasoning. These gaps are evident in recent benchmarks such as the Pragmatic Understanding Benchmark (PUB) and DiPlomat, which have exposed a significant disparity between human and machine performance on tasks requiring nuanced understanding of context and intention (Li et al., 2023). This shortcoming poses a critical challenge for applications like conversational AI, assistive technologies, and human-computer interaction systems, where an accurate interpretation of implied meanings is essential.

Addressing these challenges requires a shift from conventional supervised fine-tuning approaches, which are often limited by the lack of reasoning-focused datasets, to methods that leverage detailed reasoning annotations. Generating such datasets not only improves model fine-tuning but also enables evaluation frameworks that assess the ability of LLMs to handle open-ended, real-world language scenarios. For instance, datasets like CIRCA, IMPPRES, and PUB provide unique opportunities to evaluate and improve model performance on complex tasks such as resolving implicatures or understanding presuppositions (Sravanthi et al., 2024).

The goal of this research is to enhance LLMs' capabilities in pragmatic reasoning by generating high-quality reasoning annotations for existing datasets and creating synthetic datasets for fine-tuning. By providing detailed explanations for each Multiple Choice Question Answering (MCQA) option, this work establishes a foundation for preference-based fine-tuning, which aligns model predictions with human-like reasoning. This approach aims to close the performance gap between LLMs and humans on challenging pragmatic tasks without compromising the generalization capabilities of the models (Zheng et al., 2021).

This study not only addresses the immediate challenges in improving LLMs' understanding of pragmatics but also lays the groundwork for future research on open-ended language generation, contextual reasoning, and socially-aware AI systems. By bridging the gap between semantics and pragmatics, this work contributes to the development of LLMs that can engage in richer, more meaningful interactions, ultimately advancing the frontier of natural language understanding.

1.3 Contributions

My contributions to advancing pragmatic reasoning in LLMs are:

Generated detailed reasoning data for existing Multiple-Choice Question Answering (MCQA) datasets, such as CIRCA(Louis et al., 2020), FLUTE(Chakrabarty et al., 2022a), FigQA(Liu et al., 2022), IMPPRES(Jeretic et al., 2020), and LUDWIG, resulting in 72,000 annotated examples.

Synthethic Data Generation: Created a synthetic dataset with 37,000 unique data points derived from distinct examples in CIRCA and LUDWIG.

Conducted Zero-Shot Experiments: Conducted zero-shot experiments on Gemma2-2b-it and Phi-2 models across the FLUTE, CIRCA, FigQA, and IMPPRES datasets to evaluate their initial performance. Additionally, fine-tuned the Gemma2-2b-it model on the CIRCA dataset, achieving a significant accuracy improvement of 25%.

Preliminary observations indicate the following:

- Models fine-tuned with reasoning data demonstrate significant performance improvements on pragmatic reasoning tasks, especially *Implicature*.

- Preference-based fine-tuning allows for better ranking of reasoning outputs, enhancing interpretability and accuracy (Zheng et al., 2021; Jeretic et al., 2020).

- Synthetic datasets generated from CIRCA and LUDWIG offer diverse and challenging examples, bridging gaps in existing datasets (Li et al., 2023).

These findings highlight the value of reasoning-focused datasets and fine-tuning techniques in advancing LLMs' capabilities.

2 Background

2.1 Overview of Language Models (LLMs)

Language models (LLMs) represent a cornerstone in modern natural language processing (NLP), powering advancements in tasks such as machine translation, summarization, and conversational AI. Traditional approaches relied on Recurrent Neural Networks (RNNs) and their enhanced versions, such as Long Short-Term Memory (LSTM) networks, which were limited by their inability to efficiently capture long-range dependencies in text. This limitation was addressed with the introduction of the Transformer architecture by (Vaswani, 2017), which revolutionized NLP by employing self-attention mechanisms. This allowed models to process sequences in parallel, capturing global dependencies while dramatically improving scalability.

LLMs like GPT-3, GPT-4, and LLaMA-2 exemplify the progress achieved in this domain. These models are pre-trained on massive corpora using objectives such as next-token prediction, enabling them to learn patterns in language from diverse datasets. By leveraging this knowledge, they can perform a wide range of tasks in zero-shot, one-shot, or few-shot settings without task-specific training. Their versatility and ability to generalize make them indispensable tools for advancing AI capabilities (Radford et al., 2019; Hu et al., 2021).

The evolution of LLMs has also paved the way for multi-task learning, where a single model is designed to handle various tasks. Such models rely on a unified architecture that generalizes well across different domains, moving beyond narrow task-specific systems to more robust generalists (Radford et al., 2019).

2.1.1 LLM Architectures and Their Training Paradigms

Modern large language models (LLMs) are based on the Transformer architecture, a breakthrough that introduced self-attention mechanisms for efficiently capturing long-range dependencies in sequential data. Transformers are composed of encoder-decoder layers, where encoder layers process input sequences and decoder layers generate contextually relevant outputs. This architecture uses multi-head self-attention and feedforward networks, making it highly parallelizable and scalable for training on large datasets (Vaswani, 2017).

Training Paradigms LLMs are trained using various paradigms:
Autoregressive Models: Models like GPT-3 predict the next token in a sequence based on previous tokens. This paradigm is effective for tasks such as text generation and code completion (Radford et al., 2019).
Masked Language Models: Models like BERT predict masked tokens within a sequence, leveraging bidirectional context to excel in comprehension tasks such as question answering (**?**).
Encoder-Decoder Architectures: Models like T5 combine the strengths of both paradigms by encoding input sequences and decoding them for specific tasks, enabling performance on both generative and extractive tasks (Raffel et al., 2020).

Fine-tuning Methods Fine-tuning LLMs for downstream tasks is essential to adapt pre-trained models to specific domains. While full fine-tuning updates all model parameters, recent advancements have introduced parameter-efficient methods to reduce computational costs while maintaining performance.

Full-Model Fine-Tuning (FMT): In full-model fine-tuning, all parameters of the pre-trained model are updated. While effective, this method is computationally intensive and memory-demanding, especially for models with billions of parameters like GPT-3 (175B) (Ouyang et al., 2022), T5-XXL (11B) (Raffel et al., 2020), and PaLM (540B) (**?**). Despite its high computational cost, FMT has been widely adopted in tasks such as machine translation, summarization, and natural language inference, where domain-specific adaptation is crucial.

Parameter-Efficient Fine-Tuning (PEFT): PEFT methods optimize a subset of parameters, significantly reducing computational overhead. Popular approaches include:

- **Low-Rank Adaptation (LoRA):** LoRA injects trainable rank-decomposition matrices into the Transformer layers' dense weight matrices, allowing the pre-trained weights to remain frozen. This approach significantly reduces the number of trainable parameters and computational requirements. LoRA achieves comparable or superior performance to full fine-tuning, with reduced memory usage and inference latency (Hu et al., 2021).

- **Prefix Tuning:** This method adds trainable "soft prompts" to the input sequence, optimizing only the new parameters introduced for specific tasks. It is particularly effective for tasks with limited labeled data but may face instability as the number of trainable tokens increases (Li and Liang, 2021).

- **Adapters:** Adapters insert small trainable layers between the frozen pre-trained model layers. These layers enable task-specific adaptations while preserving the base model's generalization capabilities. However, adapters can introduce inference latency in some setups (Houlsby et al., 2019).

Reinforcement Learning from Human Feedback (RLHF): RLHF fine-tunes LLMs by aligning their outputs with human preferences using reward signals. This method enhances the interpretability and alignment of model outputs, making it ideal for sensitive applications like conversational AI. It involves three stages: supervised fine-tuning, reward model training, and policy optimization using techniques like Proximal Policy Optimization (PPO) (Ouyang et al., 2022).

Comparative Insights While full fine-tuning offers the highest flexibility, its computational cost is prohibitive for large models. Parameter-efficient methods like LoRA and prefix tuning strike a balance between efficiency and performance, making them suitable for large-scale deployment. RLHF further enhances user alignment, bridging the gap between human preferences and model outputs. The choice of fine-tuning method depends on factors such as task complexity, data availability, and computational resources (Hu et al., 2021; Ivison et al., 2024).

2.2 Existing Benchmarks and Datasets

Benchmark datasets have been a cornerstone of evaluating the capabilities of language models, serving as a standard for measuring performance and guiding improvements in model design. Early benchmarks, such as GLUE and SuperGLUE, assessed tasks like sentence similarity, textual entailment, and natural language inference, focusing primarily on semantic understanding. Similarly, datasets like MMLU broadened the scope to reasoning and knowledge-based tasks, but they did not address the deeper challenges of pragmatic reasoning, where context, speaker intent, and implied meanings are crucial (Radford et al., 2019; Ivison et al., 2024).

Recognizing this gap, specialized benchmarks such as the Pragmatics Understanding Benchmark (PUB) were developed to evaluate models' pragmatic reasoning abilities. PUB includes tasks that assess phenomena like implicature, presupposition, deixis, and reference, presented in a Multiple Choice Question Answering (MCQA) format. These tasks test models' ability to interpret implied meanings and context-dependent information rather than relying solely on explicit textual cues. PUB provides a comprehensive evaluation framework for advancing pragmatic understanding in language models (Li et al., 2023; Ivison et al., 2024).

Beyond PUB, datasets like CIRCA evaluate conversational implicatures and ambiguity resolution by testing how models handle indirect speech acts. CIRCA highlights the importance of aligning model outputs with human dialogue norms, advancing conversational AI systems (Li et al., 2023). Similarly, FLUTE provides tasks that challenge models on figurative language, including metaphors and idioms, requiring deeper linguistic reasoning (Chakrabarty et al., 2022b). IMPPRES, on the other hand, focuses on scalar implicatures and presuppositions, offering curated examples that probe a model's ability to infer unstated meanings (Jeretic et al., 2020).

DiPlomat introduces situated reasoning within dialogues, requiring models to consider conversational context, environmental factors, and social cues. This makes it particularly valuable for interactive virtual assistants and multi-modal AI systems (Li et al., 2023). The GRICE dataset evaluates adherence to conversational maxims like relevance and quantity, while Social-IQA incorporates pragmatic reasoning in social scenarios, bridging pragmatics and social inference (Houlsby et al., 2019; Radford et al., 2019).

Collectively, these benchmarks emphasize the importance of pragmatics, addressing nuances of human communication often overlooked by earlier datasets. They enable researchers to design systems capable of navigating real-world communication challenges, paving the way for sophisticated and context-aware language technologies. Despite these advancements, there remains a need for unified benchmarks that integrate major domains of pragmatics for comprehensive evaluation (Li et al., 2023; Ivison et al., 2024).

3 Literature Survey

3.1 Key Studies on Pragmatics in NLP

Pragmatics, as a field within natural language processing (NLP), deals with understanding how context influences the meaning of language. Over the years, researchers have explored various pragmatic phenomena, including implicature, presupposition, deixis, and reference, to bridge the gap between machine understanding and human communication. Early studies mostly focused on surface-level semantics, leaving deeper, context-driven aspects of language relatively untouched. However, with the growing complexity of NLP applications, such as conversational agents and virtual assistants, pragmatics has become essential.

Several key studies have provided valuable insights into this domain. For example, indirect answers in polar questions have been used to study implicatures, revealing how models interpret implicit meanings (Louis et al., 2020). Hierarchical grammar models have explored how conversational implicatures and deictic references are processed (Zheng et al., 2021). Natural Language Inference (NLI) frameworks have been applied to scalar implicatures, and corpus-level annotations have advanced the study of sentence-level implicature ratings (Lahiri, 2015). Similarly, studies on presuppositions have analyzed search engine queries with questionable assumptions and examined how Transformers exploit structural and lexical cues to interpret implicit information (Kabbara and Cheung, 2022).

Despite these advancements, earlier research often focused on isolated phenomena or relied on limited datasets, making it difficult to generalize findings. These challenges motivated the development of more comprehensive benchmarks to systematically evaluate pragmatic reasoning.

3.2 Summary of Studies Addressing LLM Capabilities in Pragmatics

Recognizing the need for comprehensive evaluations, recent work has introduced datasets and benchmarks that focus specifically on pragmatics. Two prominent resources in this area are the Pragmatics Understanding Benchmark (PUB) and the

DiPlomat dataset.

The PUB Dataset: PUB is a large-scale benchmark designed to assess LLMs across a variety of pragmatic phenomena, such as implicature, presupposition, deixis, and reference. It includes 28,000 data points, with 6,100 newly annotated examples, and uses a Multiple Choice Question Answering (MCQA) format. PUB's evaluation of LLMs has led to some important observations:

- Smaller models show significant improvements in pragmatic reasoning when fine-tuned on conversational tasks.
- Larger models, such as GPT-3 and LLaMA-2, perform competitively even without specific fine-tuning, indicating the benefits of scale.
- There is considerable variability in model performance across tasks, with implicature and deixis proving the most challenging.
- Human evaluators still outperform even the best-performing models, highlighting the gap between machine and human understanding of pragmatics (Ivison et al., 2024).

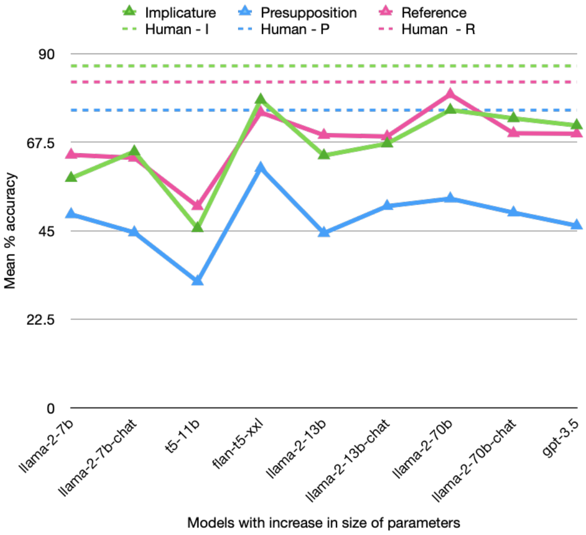


Figure 3: Performance of LLMs on the PUB dataset across various pragmatics tasks, showcasing the performance gap between models and human evaluators.

The DiPlomat Dataset: DiPlomat shifts the focus to situated pragmatic reasoning within dialogues. Unlike PUB, it requires models to interpret not just conversational context but also physical and social cues. Key findings from DiPlomat include:

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

Table 1: Comparison of Matthew's correlation coefficient (ϕ) for Human-GT and Human-LLM (llama-2-base-70b) across 300 examples. Tasks 1-10 examine Implicature, Tasks 11-12 assess Presupposition, and Task 13 focuses on Reference and Deixis. Red text indicates correlation differences between Human-GT and Human-LLM for each task.This table is adapted from the Pragmatics Understanding Benchmark (PUB) paper.

- Models with multi-modal capabilities perform better in tasks involving physical context but still struggle with nuanced conversational implicatures.
- While fine-tuning leads to modest improvements, models still lag far behind human performance in understanding conversational subtleties.
- DiPlomat highlights the limitations of current LLMs in integrating multi-modal and pragmatic reasoning (Li et al., 2023).

3.3 Progress in Fine-Tuning and Instructions

Fine-tuning has been a key method for adapting LLMs to specific tasks. Traditional fine-tuning involves updating all model parameters, which, while effective, is computationally intensive for large models. Recently, techniques like LoRA and prefix tuning have emerged as parameter-efficient alternatives, allowing models to adapt to tasks with fewer trainable parameters and reduced computational overhead (Hu et al., 2021; **?**).

Instruction-following approaches, such as InstructGPT, have also shown promise. By training models on human-provided instructions and examples, these methods align model outputs more closely with user expectations. Reinforcement learning from human feedback (RLHF) further refines this alignment, making models more interpretable and context-aware. However, these methods still struggle with nuanced aspects of pragmatics, such as humor, irony, and conversational implicature (Ouyang et al., 2022).

4 Experiments

4.1 Datasets Extension

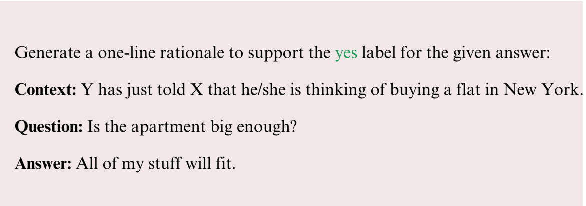
CIRCA (Contextual Interpretation of Reference and Conversational Ambiguities)

The CIRCA dataset(Louis et al., 2020), comprising approximately 29,000 data points, was originally developed to assist machine learning systems in interpreting indirect answers to polar (yes/no) questions. It contains pairs of yes/no questions and indirect responses, annotated to reflect the implied meaning of the answer. The dataset spans ten distinct social conversational contexts, such as food preferences, activities, and hypothetical scenarios. Examples of CIRCA questions include:

Data Extension: To align with the goals of fine-tuning and preference-based optimization, the CIRCA dataset was extended in several ways:

Rationale Generation for Correct Labels: Using GPT-4o mini, rationales were generated for each correct label (e.g., Yes/No) based on the context, question, and answer. **Figure 4** shows the prompt used for generating these rationales.

Rationale Generation for Incorrect Labels: For preference-based optimization, rationales for incorrect labels were also needed. These were generated using a set of 50 pre-designed templates, inspired by related works. Figure **??** provides examples of the templates used for generating incorrect label rationales.



Generate a one-line rationale to support the yes label for the given answer:

Context: Y has just told X that he/she is thinking of buying a flat in New York.

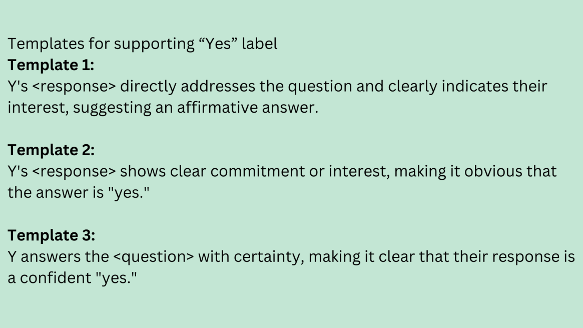
Question: Is the apartment big enough?

Answer: All of my stuff will fit.

Figure 4: Prompt used to generate rationales for correct labels using GPT-4o mini.

Templates for Rationale Generation: To generate rationales for incorrect labels as part of preference-based optimization, a set of 50 paraphrased templates was created. Below are two examples of template sets for supporting "Yes" and "No" labels:

Expanded Dataset: The resulting dataset includes rationale explanations for both correct and incorrect labels, enriching the dataset for fine-tuning and preference-based tasks.



Templates for supporting “Yes” label

Template 1:

Y's <response> directly addresses the question and clearly indicates their interest, suggesting an affirmative answer.

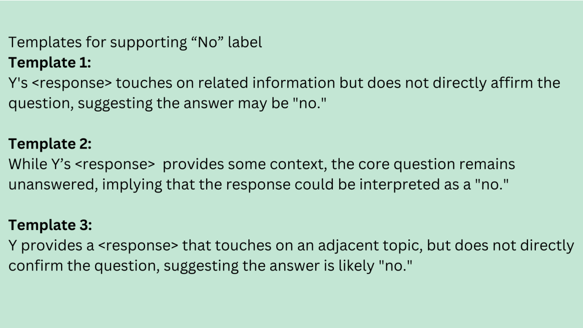
Template 2:

Y's <response> shows clear commitment or interest, making it obvious that the answer is "yes."

Template 3:

Y answers the <question> with certainty, making it clear that their response is a confident "yes."

Figure 5: Examples of templates used for generating rationales to support the "Yes" label. These templates provide a variety of interpretations to align responses with the correct label.



Templates for supporting “No” label

Template 1:

Y's <response> touches on related information but does not directly affirm the question, suggesting the answer may be "no."

Template 2:

While Y’s <response> provides some context, the core question remains unanswered, implying that the response could be interpreted as a "no."

Template 3:

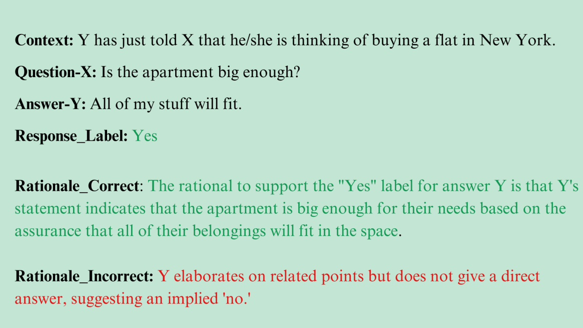
Y provides a <response> that touches on an adjacent topic, but does not directly confirm the question, suggesting the answer is likely "no."

Figure 6: Examples of templates used for generating rationales to support the "No" label. These templates highlight common patterns that justify the negative response.

Significance of Extension: With nearly 29,000 data points, the extended CIRCA dataset serves as a robust resource for evaluating and fine-tuning language models on tasks involving conversational implicature. The addition of rationales enhances its applicability for preference-based learning, enabling models to not only predict labels but also provide interpretable reasoning for their choices.

FLUTE (Figurative Language Understanding Task Evaluation)

The FLUTE dataset (Chakrabarty et al., 2022b) addresses the challenges of understanding figurative language within the framework of natural language inference (NLI). It consists of 9,000 instances across four categories: Sarcasm, Similes, Metaphors, and Idioms. Each instance contains a premise, hypothesis, and an explanation for the correct entailment or contradiction label. The



Context: Y has just told X that he/she is thinking of buying a flat in New York.

Question-X: Is the apartment big enough?

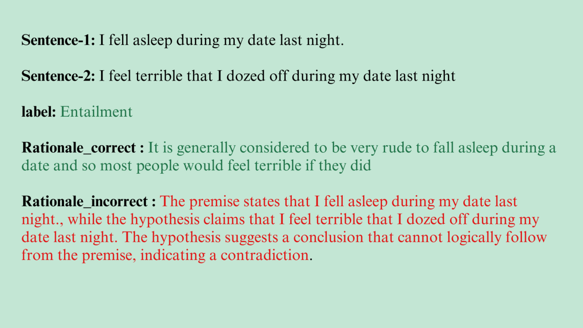
Answer-Y: All of my stuff will fit.

Response_Label: Yes

Rationale_Correct: The rational to support the "Yes" label for answer Y is that Y's statement indicates that the apartment is big enough for their needs based on the assurance that all of their belongings will fit in the space.

Rationale_Incorrect: Y elaborates on related points but does not give a direct answer, suggesting an implied 'no.'

Figure 7: Examples of CIRCA data, showcasing indirect responses to polar questions along with rationale annotations for both correct and incorrect labels.



Sentence-1: I fell asleep during my date last night.

Sentence-2: I feel terrible that I dozed off during my date last night

label: Entailment

Rationale_correct : It is generally considered to be very rude to fall asleep during a date and so most people would feel terrible if they did

Rationale_incorrect : The premise states that I fell asleep during my date last night., while the hypothesis claims that I feel terrible that I dozed off during my date last night. The hypothesis suggests a conclusion that cannot logically follow from the premise, indicating a contradiction.

Figure 8: Example of rationale generation for FLUTE dataset. The correct rationale supports the entailment label, while the incorrect rationale highlights a contradiction scenario.

dataset was developed using a Human-AI collaboration framework involving GPT-3, crowd workers, and expert annotators, demonstrating how scalable pipelines can support the creation of high-quality datasets for complex linguistic tasks. While the original dataset provides detailed rationales for correct labels, explanations for incorrect labels were not included, limiting its applicability for preference-based fine-tuning.

Data Extension: To enhance the FLUTE dataset's utility for advanced evaluation and preference-based learning tasks, rationales for incorrect labels were generated as part of this work. The extension process involved three key steps:
(1) Cosine Similarity Filtering: Explanations for correct and incorrect labels were compared using cosine similarity. Sentences with a similarity score of zero were directly included, as they represented completely distinct reasoning.
(2) Template Mapping: For cases where the cosine similarity was non-zero, a set of carefully designed templates was used to generate incorrect label rationales. These templates were paraphrased to ensure logical and linguistic diversity.
(3) Manual Validation: To ensure quality, a manual review of all generated rationales was conducted. This step guaranteed the logical consistency and alignment of explanations with the premise and hypothesis.

The resulting extended dataset comprises 7,534 entries, including explanations for both correct and incorrect labels. Each entry includes the premise, hypothesis, label, original explanation, and newly generated explanation for incorrect labels.

Significance of Extension: By including explanations for incorrect labels, the extended FLUTE dataset provides a richer resource for evaluating language models on tasks involving figurative language. This work enables preference-based learning, allowing models to differentiate between valid and invalid reasoning more effectively. The dataset now serves as a robust benchmark for improving textual reasoning capabilities in language models, advancing the field of explainable NLP.

FigQA (Figurative Question Answering)

The FigQA dataset (Rakshit and Flanigan, 2022) addresses the challenges of understanding figurative language in NLP, which often requires commonsense reasoning and flexibility in word meaning inference. The dataset emphasizes creative expressions such as "She thinks of herself as a particle of sand in the desert," which challenge models that predominantly rely on literal interpretations. Correct inference for such cases involves contextual reasoning, creativity, and an understanding of figurative language. FigQA is formatted as a Winograd schema, where each entry consists of a shared ‘startphrase‘ and two possible endings with opposite meanings. This structure reduces shortcut learning and ensures reliance on contextual understanding.

Transformation into NLI Format: To enhance the dataset's utility for natural language inference (NLI) tasks, the original FigQA entries were split into two rows: - One marking the correct ending as **Entailment**. - Another marking the incorrect ending as **Contradiction**.

This transformation doubled the number of entries, resulting in a dataset of **20,511** instances. Each example now consists of two sentences (‘sentence1‘ and ‘sentence2‘), a label (entailment or contradiction), and a rationale explaining the relationship between the sentences.

Rationale Generation and Validation: To ensure the dataset's robustness, rationales for both entailment and contradiction labels were generated and validated. The process involved:

Rationale Generation: GPT-4o mini was used to generate rationales for both correct and incorrect labels. These rationales were grounded in the relationship between ‘sentence1‘ and ‘sentence2‘.
Cosine Similarity Filtering: Incorrect label rationales were compared with correct label rationales using cosine similarity. Cases with zero similarity were directly included, while non-zero similarity cases were refined using paraphrased templates to ensure diversity.
Manual Validation: All generated rationales were manually reviewed to ensure their alignment with the dataset's figurative context and logical consistency.

Significance of Extension: The transformation of FigQA into an NLI dataset and the addition of rationales for both entailment and contradiction labels significantly enhance its usability. With 20,511 entries, the extended dataset provides a robust benchmark for evaluating and fine-tuning language models on tasks requiring figurative reasoning. The meticulous rationale validation process ensures that the dataset maintains its high quality, making it a valuable resource for advancing explainable NLP.

IMPPRES (Implicature and Presupposition)

The IMPPRES dataset (Jeretic et al., 2020) was designed to evaluate pragmatic reasoning in language models, focusing specifically on implicature and presupposition. These pragmatic phenomena go beyond literal meaning, requiring an understanding of implied meanings and assumptions inherent in statements. The dataset is organized into two top-level folders, "implicature" and "presupposition," each containing multiple sub-datasets. The "implicature" portion includes seven sub-datasets in JSON Lines format.These sub-datasets collectively cover a wide range of pragmatic reasoning scenarios, providing a robust foundation for evaluating language models.

Dataset Extension Process: The original dataset, consisting of 10,200 entries, was extended by generating rationales for each entry using GPT-4o mini. The labels used for rationale generation were based on the `gold_label_prag` column, which reflects the pragmatic interpretation. The extension process followed these steps:

Rationale Generation: GPT-4o mini was used to generate rationales for both correct and incorrect labels. The rationales were tailored to explain the specific pragmatic phenomena at play, such as scalar implicature or presupposition.
Template Usage for Incorrect Labels: Paraphrased templates were employed to generate diverse and plausible rationales for incorrect labels, similar to the methods used for FLUTE (Chakrabarty et al., 2022b) and FigQA (Rakshit and Flanigan, 2022).
Validation and Quality Control: All generated rationales were manually validated to ensure logical consistency and alignment with the dataset's linguistic phenomena.

Example from Extended Dataset:

Sentence1: *Mary had 3 of the 5 cookies.*
Sentence2: *Mary didn't have all the cookies.*
Label: *Neutral (Pragmatic)*
Rationale for Correct Label: *The premise "Mary had 3 of the 5 cookies" suggests that Mary did not have all the cookies, but it does not definitively confirm this, making the statement neutral from a pragmatic standpoint.*
Rationale for Incorrect Label: *The hypothesis "Mary didn't have all the cookies" is not directly entailed by the premise, as it is possible Mary had all the cookies; thus, the label entailment is incorrect.*

Significance of Extension: The extension of the IMPPRES dataset enhances its utility for evaluating and fine-tuning language models on tasks involving pragmatic reasoning. With the addition of rationales for both correct and incorrect labels, the dataset now provides a comprehensive resource for preference-based learning and interpretability studies. By focusing on both implicature and presupposition, this extended dataset bridges critical gaps in pragmatic evaluation.

4.2 Synthetic Data Generation

To enhance the diversity and robustness of the dataset for pragmatic reasoning tasks, synthetic data was generated using the CIRCA dataset as a base. This process involved filtering, generating

Original data point:

Startphrase: *The girl had the flightiness of a sparrow.*

Ending1: *The girl was very stable. (Incorrect)*
Ending2: *The girl was very fickle. (Correct)*

Transformed data:

Sentence1: *The girl had the flightiness of a sparrow.*
Sentence2: *The girl was very fickle.*
Label: *Entailment*
Reasoning for Correct Label: *The phrase "the flightiness of a sparrow" implies instability, aligning with the description of the girl as fickle.*

Sentence1: *The girl had the flightiness of a sparrow.*
Sentence2: *The girl was very stable.*
Label: *Contradiction*
Reasoning for Incorrect Label: *The phrase "the flightiness of a sparrow" implies instability, directly contradicting the description of the girl as stable.*

Figure 9: An example from the transformed dataset illustrating entailment and contradiction reasoning.

new data points, validating their relevance, and ensuring they met the criteria for indirect responses.

Data Preparation: The CIRCA dataset, consisting of unique question-answer pairs, was filtered to identify 3,345 unique questions. For each question, five indirect responses supporting the label "Yes" and five indirect responses supporting the label "No" were generated. This resulted in a broad range of indirect answers for each question, enabling comprehensive coverage of potential response patterns.

Synthetic Data Generation Process: The synthetic responses were generated using GPT-4o-mini, leveraging few-shot prompting to provide the necessary contextual examples for response generation. The process included:

- **Temperature Diversity:** Five different temperatures—0, 0.5, 0.9, 1.2, and 1.5—were used during generation to introduce randomness and variety in the responses.
- **Few-shot Prompting:** Between 3 to 6 examples were randomly selected from a pre-curated list of 50 examples to serve as contextual prompts for GPT-4o-mini during the response generation.
- **Duplicate Removal:** After generation, the data was processed to remove duplicate entries, leading to the deletion of 2,356 responses.

Validation of Synthetic Data: To ensure that the generated responses adhered to the desired criteria of being indirect, a BERT-based classifier (`bert-base-uncased`) was employed. The classifier was trained on the original CIRCA dataset to predict whether a response to a question was direct or indirect. The validation results are summarized in Table 2.

The confusion matrix for the classification results is shown in Table 3.

Final Dataset: After validation and further cleaning, including the removal of the 5 misclassified responses, the final synthetic dataset consists of 31,089 entries. Each entry aligns with the goal of providing indirect responses, validated through the

Metric	Value
Accuracy	0.9999
Precision	1.0000
Recall	0.9999
F1 Score	0.9999

Table 2: Performance Metrics for Synthetic Data Validation

	Direct	Indirect
Direct	0	0
Indirect	5	33,435

Table 3: Confusion Matrix for Synthetic Data Validation

classifier's performance.

Dataset Statistics: Table 4 provides a summary of all datasets used, including the synthetic data.

Dataset	Total Entries
CIRCA	29,461
FLUTE	7,534
FigQA	20,511
IMPPRES	10,200
Synthetic (CIRCA)	31,089
Sythetic(LUDWIG)	7,079

Table 4: Dataset Statistics

Significance of Synthetic Data: The synthetic data, generated with diverse temperatures and validated rigorously, enhances the dataset's coverage of indirect response patterns. By leveraging few-shot prompting and a high-performing BERT classifier for validation, the synthetic data contributes significantly to creating a richer and more comprehensive resource for training and evaluating language models on pragmatic reasoning tasks.

4.3 Evaluation Methodology

To evaluate the performance of language models on pragmatic reasoning tasks, two methodologies were employed: zero-shot prompting and fine-tuning. These evaluation techniques were chosen to test the baseline generalization capabilities of models and the impact of task-specific adaptation through training.

Zero-shot prompting serves as a baseline for assessing a model's ability to perform tasks without any explicit task-specific fine-tuning. This approach is particularly useful for evaluating large pre-trained language models, which are designed to generalize across tasks by relying solely on their pre-training. In this study, system prompts were carefully crafted for each dataset to align with the task's requirements. For example, in the CIRCA dataset, the prompts were designed to classify indirect responses in conversations, while the prompts for FLUTE focused on identifying and explaining figurative language. The models evaluated in this setting included *Gemma-2-2b-it*, *Phi-2*, and*GPT-4o-mini* and their performances were measured in terms of accuracy, precision, and recall across the CIRCA, FLUTE, FigQA, and IMPPRES datasets.

Fine-tuning, on the other hand, was employed to improve the model's performance on specific tasks. A LoRA-based fine-tuning methodology was adopted, which involves injecting trainable low-rank matrices into the weight structure of pre-trained Transformer models. This technique allows for parameter-efficient tuning while keeping most of the model's pre-trained weights frozen, significantly reducing the computational and memory costs. Fine-tuning experiments were conducted on the *Gemma-2-2b-it* model using the CIRCA dataset, which focuses on indirect response classification. The fine-tuning process included training with a reduced learning rate and gradient checkpointing to optimize memory usage. The results revealed a substantial improvement in the model's performance, with the accuracy increasing by **25%** from 56% to 70%. This demonstrates the effectiveness of task-specific training in addressing the nuanced challenges of pragmatic reasoning tasks.

Table 5: Performance Metrics of *Gemma-2-2b-it* Model Before and After Fine-Tuning on CIRCA Dataset

Metric	Before Fine-Tuning	After Fine-Tuning
Accuracy	0.56	0.70
Precision	0.625	0.857
Recall	0.666	0.600

Table 6: Performance Metrics of *Gemma-2-2b-it* Model Before and After Fine-Tuning on CIRCA Dataset

5 Results and Analysis

5.1 Quantitative Results

The results of both zero-shot evaluations and fine-tuning experiments are presented in detail through tables and visualizations. The performance metrics, including accuracy, precision, and recall, for all

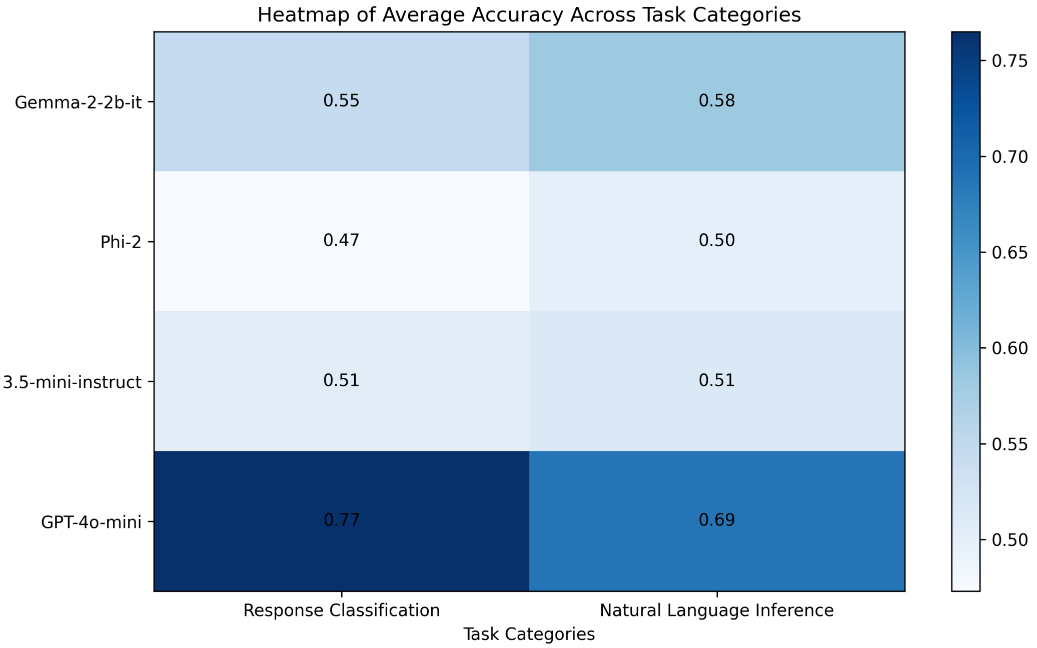


Figure 10: Heatmap of model performance across datasets and tasks. The heatmap visualizes the accuracy scores for each model and dataset.

models and datasets, are tabulated below in Table 7. The table highlights the zero-shot performance for each dataset and model. In addition, the average metrics across tasks are summarized in Table 8, which provides a comparative view of the overall performance of each model.

5.2 Qualitative Analysis

Qualitative analysis highlights the improvements in reasoning and decision-making after fine-tuning. Before fine-tuning, *Gemma-2-2b-it* struggled with nuanced classifications, such as identifying polite refusals in CIRCA as contradictions. Fine-tuning enabled the model to correctly classify these cases, demonstrating enhanced contextual understanding.

5.3 Comparative Analysis

Among the models, *GPT-4o-mini* consistently performed better in zero-shot evaluations, while fine-tuning allowed *Gemma-2-2b-it* to achieve comparable results in the CIRCA dataset. This indicates that fine-tuning is crucial for smaller models to bridge the performance gap with larger, pre-trained counterparts. The average accuracy across tasks shows that *GPT-4o-mini* maintains the highest overall performance, especially in natural language inference tasks.

The experiments reveal that fine-tuning significantly improves the performance of models on pragmatic reasoning tasks. *Gemma-2-2b-it* improved its accuracy by 25% on CIRCA after fine-tuning. These results demonstrate the importance of task-specific training in enhancing the capabilities of language models. Future work will extend these experiments to additional datasets and explore preference-based fine-tuning techniques.

.

6 Summary, Conclusion, and Future Work

6.1 Summary

This study explored methods to enhance pragmatic reasoning in Large Language Models (LLMs) through both zero-shot evaluation and fine-tuning techniques. The focus was on leveraging reasoning-focused datasets such as CIRCA, FLUTE, FigQA, and IMPPRES to assess and improve models' understanding of indirect responses, figurative language, and implicature. The dataset extension process involved generating reasoning rationales for both correct and incorrect labels using advanced prompting strategies and template-based generation. This ensured the availability of high-quality training data for preference-based fine-tuning.

Task Category	Dataset	Model	Accuracy	Precision	Recall
Response Classification	CIRCA	gemma-2-2b-it	0.560	0.625	0.666
		phi-2	0.486	0.460	0.530
Natural Language Inference	FLUTE	gemma-2-2b-it	0.530	0.392	0.168
		phi-2	0.460	0.4123	0.5341
		Phi-3.5-mini-instruct	0.520	0.4592	0.511
		gpt-4o-mini	0.860	0.819	0.870
	FigQA	gemma-2-2b-it	0.501	0.511	0.181
		phi-2	0.530	0.508	0.604
		Phi-3.5-mini-instruct	0.485	0.463	0.468
		gpt-4o-mini	0.705	0.740	0.593
	IMPPRES	gemma-2-2b-it	0.6624	0.2037	0.2340
		phi-2	0.470	0.208	0.615
		Phi-3.5-mini-instruct	0.540	0.202	0.461
		gpt-4o-mini	0.670	0.200	0.230

Table 7: Zero-shot evaluation metrics (Accuracy, Precision, and Recall) for different models (Gemma-2-2b-it, Phi-2, Phi-3.5-mini-instruct, and GPT-4o-mini) across multiple datasets (CIRCA, FLUTE, FigQA, and IMPPRES). This table highlights the performance of each model on tasks involving response classification and natural language inference.

Task Category	Gemma-2-2b-it	Phi-2	Phi-3.5-mini-instruct	GPT-4o-mini
Response Classification	54.50%	47.30%	50.50%	76.50%
Natural Language Inference	58.20%	50.00%	51.30%	68.80%

Table 8: Average accuracy for each model across task categories (Response Classification and Natural Language Inference). Values above 70% are marked in green, while those below 50% are marked in red.

In the experimental phase, zero-shot evaluations highlighted the generalization capabilities of pre-trained models like *GPT-4o-mini* while revealing gaps in the reasoning capabilities of smaller models such as *Gemma-2-2b-it* and *Phi-2*. Fine-tuning experiments, particularly on the CIRCA dataset using LoRA-based techniques, demonstrated significant performance improvements, with *Gemma-2-2b-it* achieving a 25% increase in accuracy. Quantitative results were complemented by qualitative analyses, showcasing enhanced reasoning and decision-making capabilities in fine-tuned models.

The results were comprehensively presented through tables, including metrics and average accuracies across tasks, and visualized using bar plots, line graphs, and heatmaps. These findings underscore the importance of task-specific training in addressing the challenges of pragmatic reasoning.

6.2 Conclusion

The findings of this research underline the critical role of reasoning-focused datasets and fine-tuning methodologies in advancing the pragmatic understanding of LLMs. Zero-shot prompting proved effective for larger, pre-trained models, but smaller models required fine-tuning to achieve competitive performance. The experiments highlighted the ability of fine-tuned models to better align with human-like reasoning, especially in tasks requiring nuanced contextual interpretation.

Fine-tuning using LoRA not only enhanced accuracy but also demonstrated computational efficiency, making it a viable option for adapting large models to specific domains. The improvements in *Gemma-2-2b-it* validate the effectiveness of preference-based training, which utilizes reasoning data for both correct and incorrect labels to refine decision-making processes. The significant gains achieved on datasets like CIRCA emphasize the importance of creating diverse, high-quality datasets for future advancements.

This work contributes to the growing body of research on computational pragmatics by providing a systematic approach to dataset extension, zero-shot evaluation, and fine-tuning. It sets a strong foundation for further exploration of pragmatic phenomena in language models, bridging the gap between semantic understanding and real-world communication complexities.

6.3 Future Work

While this study achieved notable progress, several areas remain open for further exploration. Future work will focus on extending the fine-tuning experiments to additional datasets, such as FLUTE and FigQA, to evaluate the generalizability of the proposed methods across diverse pragmatic tasks. A key direction will be the implementation of preference-based fine-tuning techniques. By leveraging reasoning data for both correct and incorrect labels, preference-based fine-tuning can further refine models' decision-making processes, enabling them to rank responses effectively and improve reasoning consistency.

Another critical part involves experimenting with advanced reinforcement learning methodologies, such as Reinforcement Learning from Human Feedback (RLHF), utilizing the generated reasoning data. RLHF has the potential to align models more closely with human-like reasoning by incorporating feedback loops into the fine-tuning process. Investigating different variants of RLHF, including techniques like Direct Preference Optimization (DPO) or Proximal Policy Optimization (PPO), will provide deeper insights into optimizing models' pragmatic reasoning abilities.

Additionally, future work can extend evaluations to multilingual pragmatic reasoning, addressing the need for models that can handle diverse linguistic contexts. Current datasets are primarily focused on English, but pragmatic reasoning is inherently language-dependent. Expanding the scope to include low-resource languages will lead to more inclusive and versatile models capable of understanding cultural and linguistic nuances.

Incorporating external knowledge bases into training workflows presents another promising direction. Pragmatic reasoning often requires understanding unstated implications and cultural context, which can be enhanced by integrating commonsense reasoning frameworks or domain-specific knowledge into the training process.

Finally, these methods can be applied to evaluate the impact of fine-tuning on other high-level linguistic phenomena, such as humor, irony, emotional reasoning, and conversational adaptability. By continuing to refine these approaches, researchers can create LLMs that are not only semantically accurate but also contextually aware and pragmatically sophisticated, paving the way for more advanced and human-like AI systems.

References

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

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *arXiv preprint arXiv:2406.09279*.

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

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

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

Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2023. Diplomat: a dialogue dataset for situated pragmatic reasoning. *Advances in Neural Information Processing Systems*, 36:46856–46884.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

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

Geetanjali Rakshit and Jeffrey Flanigan. 2022. FigurativeQA: A test benchmark for figurativeness comprehension for question answering. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 160–166, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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