Abstract

Humans often express their communicative intents indirectly or non-literally, which requires their interlocutors—human or AI—to understand beyond the literal meaning of words. While most existing work has focused on discriminative evaluations, we present a new approach to generatively evaluate large language models’ (LLMs’) intention understanding by examining their responses to non-literal utterances. Ideally, an LLM should respond in line with the true intention of a non-literal utterance, not its literal interpretation. Our findings show that LLMs struggle to generate pragmatically relevant responses to non-literal language, achieving only 50-55% accuracy on average. While explicitly providing oracle intentions significantly improves performance (e.g., 75% for Mistral-Instruct), this still indicates challenges in leveraging given intentions to produce appropriate responses. Using chain-of-thought to make models spell out intentions yields much smaller gains (60% for Mistral-Instruct). These findings suggest that LLMs are not yet effective pragmatic interlocutors, highlighting the need for better approaches for modeling intentions and utilizing them for pragmatic generation.1

1 Introduction

Humans possess the ability to communicate and understand each other even through non-literal utterances and conversational implicatures (Roberts and Kreuz, 1994; Dews and Winner, 1999; Glucksberg and McGlone, 2001; Dews and Winner, 1999). This is attributed to their ability to make pragmatic inferences arising from contextual factors and conventions in conversation, rather than specific words or phrases (Grice, 1975; Davis and Davis, 2016). Since humans often use non-literal language in communication, large language models (LLMs) must also develop pragmatic understanding to facilitate effective and nuanced human-AI interactions.

In this work, we introduce a new generative evaluation framework designed to evaluate the ability of LLMs to understand and resolve intentions through pragmatic response generation. In Figure 1, Kelly uses hyperbole to express her desire to read numerous books. A contextually appropriate response would be to ideally echo sentiments like “That sounds like a great plan” rather than interpreting “a million” literally, as seen in responses like “That’s quite an ambitious reading list”. Our framework uses this intuition to compare LLMs’ responses to human-like expectations, enabling a nuanced assessment of their pragmatic understanding and response accuracy.

Our primary focus on pragmatic response generation marks a departure from prior work (Zheng et al., 2021; Hu et al., 2022; Srivastava et al., 2023; Ruis et al., 2023), which has predominantly measured intention understanding through a discriminative contrastive multiple-choice classification. We show that this setting does not necessarily reflect LLMs’ abilities in generating pragmatic responses, nor does it correspond to the use of LLMs as conversational agents (West et al., 2023).

We evaluate the pragmatic understanding of several state-of-the-art open-source LLMs on various types of non-literal language from Hu et al. (2022). We observe that LLMs often struggle with generating contextually appropriate responses and tend to interpret non-literal language literally, with an accuracy of 50-55%. Furthermore, we find that LLMs’ ability in detecting intentions does not translate to their pragmatic response generation, highlighting a key distinction between merely detecting intentions and pragmatically acting on them in a generative setting. Finally, we explored approaches to improve LLMs’ pragmatic response abilities. Using chain-of-thought prompting to make models explicitly spell out intentions before generation has mini-
Annie and Kelly are discussing their plans for summer. Annie asks Kelly: “How many books do you plan to read this summer?”

```
I've compiled a list of a million books I plan to read, Annie.

Kelly

That sounds like a great plan, Kelly!
```

1) measures LLMs’ pragmatic response generation

We introduce a new framework to evaluate pragmatic response generation. Specifically, it requires:

- **Context C**: A short narrative involving 2 or more characters.
- **Non-literal Utterance** $U_L^N$: A speaker-generated utterance using non-literal language.
- **True Intention** $I_T$: The actual intended meaning of the speaker.
- **Incorrect Literal Intention** $I_L$: An incorrect literal interpretation of the speaker’s intention.
- **Reference Dialog Chains based on $I_T$ and $I_L$**: Speaker alternatively uses direct language to convey intentions $I_T$ as $U_T^T$ and $I_L$ as $U_L^L$. The listener responds accordingly to $U_L^T$ and $U_L^L$, with $U_L^T$ and $U_L^L$ respectively. See Figure 1.

**Evaluating Pragmatic Understanding**

Our framework evaluates the extent to which LLMs’ generated responses reflect an understanding of the underlying speaker’s intention. We operationalize this into an automatic metric by using similarity measurements. Ideally, if LLMs can accurately infer and use the intent to generate cooperative responses using direct language, they should respond as if the non-literal utterance was instead communicated literally. Thus, if an LLM generates pragmatic cooperative responses, the response should be closer in similarity to response generated under the true intention than to one based on the literal interpretation i.e., the relation $\text{sim}(U_T^T, U_T^N) > \text{sim}(U_L^T, U_N^T)$ should hold under the context $C$.

**Data**

Hu et al. (2022) evaluate intention detection with a context $C$, a single non-literal utterance $U_L^N$, and verbalized intents that include a literal intent $I_L$ and true intent $I_T$. To instantiate our framework, we augment this data with dialog chains $(U_T^T, U_T^N)$ conditioned on the literal intent $I_L$ and $(U_T^T, U_T^L)$ conditioned on the true intent $I_T$. We use GPT-4 to get reference chains (See Appendix A.2).

We consider four non-literal language phenomena from Hu et al. (2022):

1. **INDIRECT SPEECH**: Speakers phrase requests indirectly, such as questions (“Can you pass the salt?”) or statements (“It is cold in here”).

---

2Hu et al. (2022) have other tasks but we do not include them (e.g., Deceits is too non-cooperative).
2. **Irony.** Speakers use irony to mean the opposite of what they say. Irony is not explicitly defined in the context $C$, but $C$ may include information about characters’ emotional states.

3. **Maxims of Conversation.** In this task, speakers flout one of Grice’s maxims.

4. **Metaphor.** In this task, the speaker uses metaphors to draw comparisons between entities in a non-literal sense.

**Models** We evaluate five state-of-the-art LLMs: Llama2-7B-chat, Llama2-13B-chat, Llama2-70B-chat, Mistral-7B-Instruct-v0.2 and Zephyr-7B-β instruction finetuned models. We generate candidate listener responses $U_2^N$ using these models, given the preceding context $C$ and the speaker’s non-literal utterance $U_1^N$. We exclude closed-source API models (GPT-3.5/4/variants) from our evaluation suite, since we follow (Hu et al., 2022)’s discriminative setup which requires access to models’ input token probabilities. Please refer to Appendix A.3 for generation details.

**Evaluators**

**Human Evaluation** Since LLM responses are intended for human conversational partners, we solicit human judgments to check whether understanding of the true intent is reflected in the generated response. We employ 9 students from our institution to evaluate whether Mistral-Instruct responses successfully capture the true intended intention $I_T$ behind the speaker’s non-literal utterance $U_1^N$, within the given context $C$. We choose Mistral-Instruct arbitrarily, since it is reported to surpass Llama-2-13B-chat model (Jiang et al., 2023) and is similar in performance to Llama-2-70B-chat (Zheng et al., 2023). We find that our annotators have a good agreement.3

**GPT-4 Contextual Similarity** Separately, we tasked GPT-4 with a contextual similarity evaluation (cf. Section 2): Given the context $C$, the speaker’s true intended meaning $I_T$, and the Mistral-Instruct generated response $U_2^N$, GPT-4 uses all the information to identify whether $U_2^N$ is more similar to the reference response conveying the true intention ($U_2^T$) or the one with the incorrect literal intention ($U_2^L$). We find that GPT-4 agrees well with human annotators.4

**Non-Contextual Embedding Similarity with Llama-3-8B-Instruct** We also measure the non-contextual cosine similarity of $U_2^N$ embeddings with reference response conveying the true intention ($U_2^T$) versus the incorrect literal intention ($U_2^L$). Using LLM2Vec (BehnamGhader et al., 2024), we obtain text embeddings from Llama-3-8B-Instruct. The similarity measured using Llama-3 embeddings generally aligns with human annotations, though it agrees less than GPT-4’s contextual similarity evaluation.5 Additionally, we experiment with contextual embedding similarity variations (Yerukola et al., 2023), where the context $C'$ can be $I_T$, $I_L$, or turn-1 responses $U_1^T$ or $U_1^L$. However, this setting performed worse. We hypothesize

---

3pairwise agreement $= 0.8$, Krippendorff’s $\alpha = 0.6$
4We average across individual pairwise agreements of each annotator with GPT-4 (pairwise agreement $= 0.77$, $\sigma = 0.05$; Krippendorff’s $\alpha = 0.54$, $\sigma = 0.1$)
5Similar to GPT-4, we average across individual pairwise agreements of each annotator with Llama-3-embeddings (pairwise agreement $= 0.74$, $\sigma = 0.005$; Krippendorff’s $\alpha = 0.46$, $\sigma = 0.01$)
that non-literal language nuances are harder to be captured by embeddings alone.

Thus, we use the better performing GPT-4 contextual similarity evaluation as a proxy for our evaluation paradigm in all our subsequent experiments.

3 Results on Pragmatic Response Generation

In this section, we analyze how well LLMs can generate contextually relevant responses. We compare our proposed generative approach, which evaluates implicit understanding in responses to $U_1^N$, against a discriminative multiple-choice setup as in Hu et al. (2022), which evaluates intention detection in $U_1^N$ utterances.

Results Figure 2 indicates that LLMs exhibit better performance in responding to INDIRECT SPEECH among various non-literal language types, potentially due to conventionalization of responses, or explicit descriptions of requests completed seen during training (Hu et al., 2022). Models perform the worst at responding to flouted MAXIMS, performing worse than chance. For instance, models fail to detect the attempt to change the subject in “Oh, it’s such a pleasant day today” amidst a discussion about a “bad date”. Llama-2 models exhibit marginally better metaphorical language understanding (METAPHORS) compared to Mistral and Zephyr models. In the Llama-2 family, we see that models perform better with increasing size. In aggregate, we see that LLMs perform at or near chance in generating an appropriate response that reflects having inferred the true intent.

Comparison against Discriminative Intention Detection We follow the multiple-choice setup as in Hu et al. (2022) (details in Appendix B). In Figure 2, we consistently see that models find it easier to detect true intentions in social situations that involve flouting conversational norms (MAXIMS) in a multiple-choice setup. However, they struggle with using this potentially inferred understanding in pragmatic response generation.

We see that trends do not remain consistent across different models and phenomena, and that on average, models struggle more in the generative setting. We hypothesize that in a discriminative setup, the model can access all options, thus it knows the answer form in advance and has the ability to evaluate the answers contrastively. However, in a generative setup, the model’s generation is free-form, requiring consistency and minimal compounding errors. This underscores the importance of evaluating model performance in both discriminative and generative settings to obtain a better understanding of LLMs’ pragmatic understanding.

4 Chain-of-Thought Prompting for Pragmatic Response Generation

Motivated by the ability of LLMs to detect intentions in some phenomena, we explore ways to improve their understanding of implicit intentions and, thereby enhancing their capability to generate pragmatic responses using chain-of-thought prompting (CoT) (Camburu et al., 2018; Wei et al., 2022).

Experiments using Chain-of-Thought In our experiments with CoT, we first generate an inferred intention and then a response (unless otherwise specified). We examine how response generation performance is affected by introducing varying levels of oracle cues at the inferred intention generation step, organized by increasing amounts of “hand-holding”:

(0) No oracle information (Naive)
(1) Counterfactual reasoning to clarify the non-literal utterances (no inferred intention here)
(2) Questioning a specific phenomenon (e.g., ‘is Kelly being ironic’)
(3) Merely indicating non-literal language use
(4) Identifying the phenomenon (e.g., ‘Kelly is being ironic’)
(5) Providing the true intention as CoT (no model-generated inferred intention here)
(6) Providing true intention and phenomenon information (e.g., “Kelly wants to read a lot and is using irony to convey it”)

Results Figure 3 illustrates that specifying the type of non-literal language used along with the speaker’s true intent (Prompt 6) significantly improves the model’s ability to generate appropriate responses, with top-performing Mistral-Instruct achieving 75% accuracy. Even providing subsets of this, such as just the true intention (Prompt 5), generally improves performance. In these cases, the task essentially becomes leveraging the provided oracle true intention in response generation. However, despite this simplification, there is still room for significant improvement in pragmatic response generation.
Intuitively, if models can accurately infer these intention cues themselves, they could generate pragmatic responses. We observe a slight improvement in performance (on average) when no oracle information is provided (Prompt 0) or when prompted for counterfactual reasoning regarding the non-literal expression (Prompt 1). Providing explicit cues about the phenomenon (e.g., ‘Kelly is being ironic’ vs. ‘Is Kelly being ironic?’) help slightly (Prompts 2-4), although not as significantly as providing the true intention.

These findings highlight the importance of explicitly modeling intention in LLMs, indicating that response accuracy to non-literal language can improve with such approaches. Overall, there is a clear need for: (a) better learning mechanisms to help models effectively disentangle the linguistic strategies used and communicative intent (e.g., recognizing how exaggeration can create irony to highlight disagreement), and (b) effective utilization of learned intentions during response generation.

5 Related Work

Non-literal language understanding in LLMs

Recent work has proposed several ways to evaluate LLMs’ ability to interpret non-literal language, including implicature (Ruis et al., 2023; Kim et al., 2023b), figurative language use (Liu et al., 2022a; Chakrabarty et al., 2022b; Gu et al., 2022b; Chakrabarty et al., 2022a; Wachowiak and Gromann, 2023; Lai and Nissim, 2024), detecting profundity (Herrera-Berg et al., 2023), broader benchmarks for social language understanding (Choi et al., 2023) and various pragmatic phenomena (Li et al., 2017a; Zheng et al., 2021; Hu et al., 2022). Kim et al. (2023b) also find that chain-of-thought helps improve a model’s ability to interpret the use of implicatures. These tasks have focused on evaluating models’ ability to interpret the true intent underlying an utterance, but not respond to it as we do in this work. Another line of work has considered LLMs’ mentalizing abilities using false belief tasks (Shapira et al., 2023) or question answering (Le et al., 2019; Kim et al., 2023a). Zhou et al. (2023a) consider a task that evaluates how models respond using knowledge of other agents’ mental states.

Generating responses based on inferred intents

Some work has presented resources for intent or emotion-conditioned response generation, where a conversational agent must respond conditioned on a particular intent or emotion. Li et al. (2017b) and Rashkin et al. (2019) present datasets of dialogues annotated with discrete emotion or intent labels. Zhang and Zhang (2019) and Chen et al. (2022) present approaches to modeling intent explicitly. Gu et al. (2022a) generate explicit scene elaborations to improve figurative language understanding. While these works consider conditioning on intent, they do not explicitly focus on generating or evaluating responses to non-literal language use.

6 Summary

We propose a new framework to evaluate how well LLMs understand intentions and respond to non-literal language, moving beyond previously employed multiple-choice settings. Our results show that LLMs often struggle to generate contextually relevant responses. While chain-of-thought prompting to spell out inferred intentions offers marginal improvements, explicitly providing oracle intentions and cues, such as for irony, significantly enhances performance. These findings highlight the current limitations of LLMs in pragmatic understanding, suggesting that improved learning mechanisms to explicitly model intentions and linguistic strategies could significantly enhance conversational abilities.

7 Limitations & Ethical Considerations

Despite taking the first step towards proposing a new generative framework for evaluating intention resolution in LLMs, there are several limitations and ethical concerns, which we list below.

Limited Context Scope In this study, our primary focus is the evaluation of intention under-
standing and using it in pragmatic response generation. Future work should explore introducing other forms of context into the pragmatic generation pipeline, such as richer social and power dynamics (Antoniak et al., 2023), emotional states (Zhou et al., 2023b), and external knowledge (Ghazvininejad et al., 2018), all of which can significantly contribute to varied levels of pragmatic understanding.

**Amount of context** In our experiments, we opted to include short 1-3 sentence stories. Future work can explore longer stories and include more preceding dialog turns. We hypothesize that more context will make this task more challenging, and we would need nuanced ways of understanding intentions at different turns.

**Limited number of non-literal phenomenon** We explore the evaluation of only four phenomena: INDIRECT SPEECH, IRONY, MAXIMS, and METAPHORS. Future work should consider other types of figurative language, such as cultural metaphors (Kabra et al., 2023), visual metaphors (Liu et al., 2022b), idioms, proverbs, etc. Expanding the scope to include these elements would provide a more comprehensive understanding of LLMs’ capabilities in interpreting nuanced language.

**Potentially Inconsistent Human Evaluation** In our work, we employ only 9 expert human annotators and assume human judgments as the gold standard. Concurrent work has shown that human evaluation might not always be consistent (Clark et al., 2021; Karpinska et al., 2021); however human judgments continue to be the gold standard for evaluating open-ended text generation.

**Potential effects on Factuality** In our work, we show that LLMs struggle with responding pragmatically to non-literal language. Training approaches which might help with better intention modeling to handle non-literal language may potentially affect faithfulness or factuality of LLMs responses.

8 Acknowledgements

We would like to thank our student annotators for helping us with intention resolution annotations. We thank OpenAI for providing researcher credits to access GPT-4. This project is funded in part by DSO National Laboratories and an Amazon Research Award, Spring 2023 CFP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of Amazon.

References

Maria Antoniak, Anjalie Field, Ji Min Mun, Melanie Walsh, Lauren F. Klein, and Maarten Sap. 2023. Riveter: Measuring power and social dynamics between entities. In ACL demonstrations.


A Pragmatic Response Generation

A.1 Data
We consider four non-literal language phenomenon from Hu et al. (2022):
1. INDIRECT SPEECH - 20 examples
2. IRONY - 25 examples
3. MAXIMS OF CONVERSATION 20 examples
4. METAPHOR - 20 examples
These examples were manually curated by expert researchers to cover a broad range of non-literal phenomena and elicit individual differences among humans.

A.2 Gold Dialog Chains based on IT and IL
In our experiments, we use GPT-4 to generate counterfactual gold dialog chains based on IT and IL. We first generate UIL_1 by providing GPT-4 with only the incorrect literal intention IL to prevent any possible contextual inference from C from being reflected in UIL_1. Next, we generate UIT_1 by providing GPT-4 with the context C and true intention IT, mimicking an ideal direct cooperative utterance generation setting. We follow the same setting to generate UIL_2 and UIT_2 from UIL_1 and UIT_1, respectively.

A.3 Models
We use the following prompt to generate responses using LLMs:
Generate a short, concise single sentence response.
Context: {context}
{speaker’s non-literal utterance}
Generate a co-operative response without any non-literal language as {listener character name}.

{listener character name} replies,
We use temperature sampling with temperatures of 0.3 and 0.5 to generate responses, averaging the results from both settings. We find that a lower temperature=0 results in incoherent, uninteresting responses, while a temperature > 0.5 leads to very verbose responses that digress from the main context and extrapolate excessively.

A.4 Evaluators
For GPT-4 as an evaluator, we provide the following prompt:
Task: You will read a short story. The story will be followed by a question. Your task is to decide which response option is closest to the 'Generated Response'. The answer options are 1 or 2.
Scenario: {Context}
{speaker’s non-literal utterance U_1^N}
Intention: {True intention}
Generated Response:
{Person2} replies, {U_2^N from model}

Compare the below utterances to the Generated Response. Which of the below utterances is closest to the above Generated Response?
Options:
1. {U_2^T from true intention dialog chain}
2. {U_2^L from incorrect literal intention dialog chain}

Answer (option number 1 or 2): {number}
We randomly shuffle intentions provided as options 1 and 2. We set the temperature=0 to ensure the stability of the evaluation.

For human evaluation, we provide the following prompt:
Task: You will be provided a short story, an utterance by one of the characters in the story (person1). Person1 uses non-literal language (like irony). Person2 from the story responds to person1’s utterance. The task is to identify if the "true intention" (provided) is resolved/understood in person2’s response or not.
Make a binary yes/no choice.
We employ 9 students from our institution – 6 women, 3 men (20-30 age group) living in the United States of America.

B Discriminative Setup
We follow setup in Hu et al. (2022) for our discriminative setup comparison. They use a multiple-choice setup. They compute the probability of answer options – true intention IT and literal misinterpretation IL – given the context C, the speaker’s non-literal utterance U_1^N, and task instructions. We
measure accuracy as assigning the highest probability to the correct answer token (e.g., “1”, “2”). We follow the same prompt template as Hu et al. (2022):

Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer. Your task is to decide what the character in the story is trying to convey. The answer options are 1 or 2.

Scenario: {context} {dialog}. What might {person1} be trying to convey?
Options:
1) {option1}
2) {option2}
Answer:

C Chain-of-thought Prompting

Please refer to for the chain-of-thought prompting templates used for all the models

C.1 Inferred Intention vs Response Accuracy

We evaluate similarity of CoT generated intents with the true intent and the incorrect literal intent using GPT-4. We follow a similar prompt as GPT-4 evaluator in Appendix A.4. We observe in Figure 4 that a model that is able to correctly infer the underlying true intention is also better at generating contextually relevant responses, corroborating our finding from PROMPT 5-6 in Section 4.
Figure 5: Chain-of-thought Prompting templates used in Section 4. Orange highlighted text is the explicitly provided oracle information.