A fine-grained comparison of pragmatic language understanding in humans and language models

Jennifer Hu[1], Sammy Floyd[1,2], Olessia Jouravlev[3], Evelina Fedorenko[1,4], Edward Gibson[1]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[2]Department of Psychology, Sarah Lawrence College
[3]Department of Cognitive Science, Carleton University
[4]McGovern Institute for Brain Research, Massachusetts Institute of Technology
{jennhu,samfloyd,evelina9,egibson}@mit.edu
olessiajouravlev@cunet.carleton.ca

Abstract

Pragmatics and non-literal language understanding are essential to human communication, and present a long-standing challenge for artificial language models. We perform a fine-grained comparison of language models and humans on seven pragmatic phenomena, using zero-shot prompting on an expert-curated set of English materials. We ask whether models (1) select pragmatic interpretations of speaker utterances, (2) make similar error patterns as humans, and (3) use similar linguistic cues as humans to solve the tasks. We find that the largest models achieve high accuracy and match human error patterns: within incorrect responses, models favor literal interpretations over heuristic-based distractors. We also find preliminary evidence that models and humans are sensitive to similar linguistic cues. Our results suggest that pragmatic behaviors can emerge in models without explicitly constructed representations of mental states. However, models tend to struggle with phenomena relying on social expectation violations.

1 Introduction

Non-literal language understanding is an essential part of communication. For example, in everyday conversations, humans readily comprehend the non-literal meanings of metaphors (*My new coworker is a block of ice*), polite deceits (*I love the gift*), indirect requests (*It's a bit cold in this room*), and irony (*Classy pajamas, dude!*). These phenomena fall under the broad label of **pragmatics**, which encompasses the aspects of meaning that go beyond the literal semantics of what is said (Horn, 1972; Grice, 1975; Yule, 1996; Levinson, 2000).

A long-standing challenge for NLP is to build models that capture human pragmatic behaviors. The remarkable abilities of modern language models (LMs) have triggered a recent effort to investigate whether such models capture pragmatic meaning, both through philosophical arguments (Bisk et al., 2020; Bender and Koller, 2020; Potts, 2020; Michael, 2020) and empirical evaluations (Jeretic et al., 2020; Zheng et al., 2021; Tong et al., 2021; Liu et al., 2022; Ruis et al., 2022; Stowe et al., 2022). However, prior empirical studies have primarily evaluated LMs based on a binary distinction between pragmatic and non-pragmatic responses, providing limited insights into models' weaknesses. A model could fail to reach the target pragmatic interpretation in multiple ways – for example, by preferring a literal interpretation, or by preferring a non-literal interpretation that violates certain social norms. Understanding these error patterns can suggest specific directions for improving the models, and foreshadow where pragmatics might go awry in user-facing settings (e.g., Saygin and Cicekli, 2002; Dombi et al., 2022; Kreiss et al., 2022).

From a cognitive perspective, understanding the pragmatic abilities of LMs could also offer insights into humans. Human pragmatic language comprehension involves a variety of mechanisms, such as basic language processing, knowledge of cultural and social norms (Trosborg, 2010), and reasoning about speakers' mental states (Brennan et al., 2010; Enrici et al., 2019; Rubio-Fernandez, 2021). However, it remains an open question when language understanding relies on explicit mentalizing – which may be cognitively effortful – versus lower-cost heuristics (e.g., Butterfill and Apperly, 2013; Heyes, 2014). Because LMs lack explicit, symbolic representations of mental states, they can serve as a tool for investigating whether pragmatic competence can arise without full-blown mentalizing (e.g., belief updates in the Rational Speech Act framework; Frank and Goodman, 2012).

Code and data: https://github.com/jennhu/lm-pragmatics

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4194–4213
July 9-14, 2023

In this paper, we perform a fine-grained comparison of humans and LMs on pragmatic language understanding tasks. Adopting the approach of targeted linguistic evaluation (e.g., Linzen et al., 2016; Futrell et al., 2019; Hu et al., 2020), our analysis serves two goals: assessing the pragmatic capabilities of modern LMs, and revealing whether pragmatic behaviors emerge without explicitly constructed mental representations. Our test materials are a set of English multiple-choice questions curated by expert researchers (Floyd et al., In prep), covering seven diverse pragmatic phenomena. We use zero-shot prompting to evaluate models with varying sizes and training objectives: GPT-2 (Radford et al., 2019), Tk-Instruct (Wang et al., 2022), Flan-T5 (Chung et al., 2022), and InstructGPT (Ouyang et al., 2022).

Through model analyses and human experiments, we investigate the following questions: (1) Do models recover the hypothesized pragmatic interpretation of speaker utterances? (2) When models do not select the target response, what errors do they make – and how do these error patterns compare to those of humans? (3) Do models and humans use similar cues to arrive at pragmatic interpretations? We find that Flan-T5 (XL) and OpenAI's text-davinci-002 achieve high accuracy and mirror the distribution of responses selected by humans. When these models are incorrect, they tend to select the incorrect literal (or straightforward) answer instead of distractors based on low-level heuristics. We also find preliminary evidence that models and humans are sensitive to similar linguistic cues. Our results suggest that some pragmatic behaviors emerge in models without explicitly constructed representations of agents' mental states. However, models perform poorly on humor, irony, and conversational maxims, suggesting a difficulty with social conventions and expectations.

2 Related work

Prior work has evaluated LMs' ability to recognize non-literal interpretations of linguistic input, such as scalar implicature (Jeretic et al., 2020; Schuster et al., 2020; Li et al., 2021) or figurative language (Tong et al., 2021; Liu et al., 2022; Gu et al., 2022; Stowe et al., 2022). In a broad-scale evaluation, Zheng et al. (2021) test five types of implicatures arising from Grice's (1975) conversational maxims, and evaluate their models after training on the task. In our work, we consider Gricean implicatures as one of seven phenomena, and we evaluate pre-trained LMs without fine-tuning on our tasks.

Similar to our work, Ruis et al. (2022) also use prompting to evaluate LMs on pragmatic interpretation tasks. They formulate implicature tests as sentences ending with "yes" or "no" (e.g., "Esther asked "Can you come to my party on Friday?" and Juan responded "I have to work", which means no."). A model is considered pragmatic if it assigns higher probability to the token that makes the sentence consistent with an implicature. In our work, models must select from multiple interpretations, enabling a detailed error analysis and comparison to humans. Ruis et al.'s materials also focus on indirect question answering as an implicature trigger, whereas we consider a broader range of pragmatic phenomena and utterance types.

Since pragmatic language understanding often draws upon knowledge of social relations, our tasks are conceptually related to benchmarks for evaluating social commonsense (e.g., Sap et al., 2019; Zadeh et al., 2019). These evaluations focus on the interpretation of actions and events, whereas we focus on the interpretation of speaker utterances. Another hypothesized component of pragmatics is Theory of Mind (ToM; Leslie et al., 2004; Apperly, 2011), or the ability to reason about others' mental states. Benchmarks for evaluating ToM in models (e.g., Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2022) primarily focus on false-belief tasks (Baron-Cohen et al., 1985), which assess whether a model can represent the beliefs of another agent that are factually incorrect but consistent with that agent's observations. LMs have been shown to succeed on some ToM tests (Kosinski, 2023) while failing on others (Sap et al., 2022; Ullman, 2023).

3 Evaluation materials

3.1 Overview of stimuli

Our evaluation materials are taken from Floyd et al.'s (In prep) experiments,[1] covering seven phenomena. Each item is a multiple choice question, with answer options representing different types of interpretation strategies. For most of the tasks, the question has three parts: a short story context (1-3 sentences), an utterance by one of the characters, and a question about what the character intended to convey.[2] Table 1 shows an example item for each

[1]Materials can be found at https://osf.io/6abgk/?view_only=42d448e3d0b14ecf8b87908b3a618672.

[2]The exceptions are Humor and Coherence.

Task	Example query	Example answer options
Deceits	Henry is sitting at his desk and watching TV, and reluctantly switches off the TV with the remote control and picks up a textbook. Shortly after, his mother comes in the room and asks, "What have you been doing up here?" Henry responds: "Reading." Why has Henry responded in such a way?	1. Correct He does not want to get into trouble for not studying. 2. Literal He has been reading for some time. 3. DistractorLexicalOverlap He does not want to offend his mom by not reading the books that she gave him. 4. DistractorSocialConvention He wants his mom to believe that he has been watching TV.
Indirect speech	Nate is about to leave the house. His wife points at a full bag of garbage and asks: "Are you going out?" What might she be trying to convey?	1. Correct She wants Nate to take the garbage out. 2. Literal She wants to know Nate's plans. 3. DistractorAssociative She wants Nate to bring his friends over. 4. DistractorLexicalOverlap She wants Nate to spend more time with the family.
Irony	It is a holiday. Stefan and Kim are sitting in the backseat of the car. They are fighting all the time. Their father says: "Oh, it is so pleasant here." What did the father want to convey?	1. Correct He does not want to listen to his kids' arguments. 2. Literal He enjoys listening to his kids fighting. 3. DistractorAssociative AC gives them some needed cool. 4. DistractorNonSequitur He remembers about his wife's birthday.
Maxims	Leslie and Jane are chatting at a coffee shop. Leslie asks, "Who was that man that I saw you with last night?" Jane responds, "The latte is unbelievable here." Why has Jane responded like this?	1. Correct She does not want to discuss the topic that Leslie has raised. 2. Literal She thinks that it is the best latte in the town. 3. DistractorAssociative The man who Leslie saw makes unbelievable lattes. 4. DistractorNonLiteral A coffee break is not a good time to discuss men.
Metaphor	Andrew and Bob were discussing the investment company where Andrew works. Bob said: "The investors are squirrels collecting nuts." What does Bob mean?	1. Correct They buy stocks hoping for future profit. 2. Literal Squirrels were hired to work in the company. 3. DistractorNonLiteral The investors dress and eat well. 4. DistractorNonSequitur Bob is allergic to nuts. 5. DistractorPlausibleLiteral The investors enjoy picking nuts as much as squirrels do.
Humor	Martha walked into a pastry shop. After surveying all the pastries, she decided on a chocolate pie. "I'll take that one," Martha said to the attendant, "the whole thing." "Shall I cut it into four or eight pieces?" the attendant asked.	1. Correct Martha said, "Four pieces, please; I'm on a diet." 2. Literal Martha said: "Well, there are five people for dessert tonight, so eight pieces will be about right." 3. DistractorAssociative Martha said, "You make the most delicious sweet rolls in town." 4. DistractorFunny Then the attendant squirted whipped cream in Martha's face. 5. DistractorNeutral Martha said, "My leg is hurting so much."
Coherence	Mary's exam was about to start. Her palms were sweaty.	1. Correct Coherent 2. Incorrect Incoherent

Table 1: Sample item from each task in our evaluation. All items are originally curated by Floyd et al. (In prep).

task, with annotated answer options. Green labels indicate the target pragmatic interpretation.[3] Blue labels indicate the literal interpretation. Red labels indicate incorrect non-literal interpretations, which are based on heuristics such as lexical similarity to the story, thus serving as distractor options.

Each task has 20-40 items, which were manually curated by expert researchers to cover a broad range of non-literal phenomena and elicit individual differences among humans. The stimuli were not specifically designed to require Theory of Mind reasoning (ToM). However, behavioral and neural evidence suggests that many of the tested phenomena rely on mentalizing processes. In Section 3.2, we briefly describe the role of ToM for each tested phenomenon, and how LMs' training corpora may provide linguistic cues to perform the tasks.

3.2 Tested phenomena

Deceits. Humans produce polite deceits ("white lies") in the service of social and personal relationships (e.g., Camden et al., 1984). Behavioral studies in young children suggest that understanding white lies requires interpretive ToM, or the ability to allow different minds to interpret the same information in different ways (Hsu and Cheung, 2013). Furthermore, the tendency to produce white

[3]We refer to these answer options as "Correct" throughout the paper. However, these answers are only "correct" in the sense of a normative evaluation. We acknowledge the wide variation in individual humans' abilities and tendencies to use non-literal language, which is not captured in our analyses. We thank an anonymous reviewer for highlighting this point.

lies is linked to emotional understanding abilities, (Demedardi et al., 2021), and moral judgments about white lies are linked to second-order false-belief understanding (Vendetti et al., 2019).

The Deceits task presents a story with a white lie, and asks why the speaker has used this utterance. The underlying intentions behind polite deceits are rarely explicitly explained in text. As a result, it is unlikely that LMs learn a direct connection between the utterance and the speaker's intention during training on static texts. However, instances of polite deceits in text corpora may be accompanied by descriptions of characters' emotional states, which may indicate that speakers' intentions differ from what is literally conveyed by their utterance. This highlights the importance of context in interpreting deceits, which we return to in Section 5.3.1.

Indirect speech. Humans often use language in a performative sense, such as indirectly requesting an action from other individuals (e.g., Austin, 1975; Searle, 1975). Indirect or polite speech comprehension has been captured by Rational Speech Act (RSA; Frank and Goodman, 2012) models, which characterize listeners as performing Bayesian inference about a speaker who chooses utterances based on a tradeoff between epistemic and social utility (Brown and Levinson, 1987; Yoon et al., 2016, 2020; Lumer and Buschmeier, 2022).

The IndirectSpeech task presents a story with an indirect request, and asks what the speaker intends to convey. Like deceits, it's unlikely that indirect speech acts are explained in text data. However, indirect requests may be followed by descriptions of the completion of the implied request – for example, that someone closed a window after hearing the utterance "It's cold in here". Therefore, models may learn relationships between the utterances and desired outcomes through linguistic experience.

Irony. Humans use irony to convey the opposite of the semantic content of their utterance (Booth, 1974; Wilson and Sperber, 1992; Attardo, 2000; Wilson and Sperber, 2012). As such, irony has long been hypothesized to rely on social reasoning and perspective-taking (e.g., Happé, 1993; Andrés-Roqueta and Katsos, 2017). Indeed, human irony comprehension behaviors are captured by Bayesian reasoning models that take into account speakers' affective goals (Kao and Goodman, 2014). In addition, neuroimaging studies suggest that irony interpretation relies on brain regions that are implicated in classic ToM tasks (Spotorno et al., 2012).

The Irony task presents a story with an ironic statement, and asks what the character intends to convey. While ironic statements are also rarely explained in text, models could leverage accompanying cues such as descriptions of characters' emotional states or a mismatch in sentiment.

Maxims of conversation. Grice (1975) proposes that communication follows a set of *maxims*: be truthful; be relevant; be clear, brief, and orderly; and say as much as needed, and no more. A prevailing theory is that listeners derive implicatures by expecting speakers to be cooperative (i.e., abide by the maxims) and reasoning about speakers' beliefs and goals. Indeed, there is extensive evidence for RSA models capturing these implicatures, such as those arising from the maxims of *quantity* (Potts et al., 2016; Frank et al., 2018; Degen, 2023) and *manner* (Bergen et al., 2016; Franke and Jäger, 2016; Tessler and Franke, 2018).

The Maxims task presents a story with a character flouting one of Grice's maxims, and asks why the character has responded in such a way. Based on linguistic input, it may be easy for LMs to recognize when a speaker is flouting a maxim – for example, if an utterance is particularly long, features an uncommon syntactic construction, or diverges semantically from the context. However, it is unclear whether LMs will be able to recover the speaker's underlying intentions.

Metaphor. Metaphors (Lakoff and Johnson, 1980) are used to draw comparisons between entities in a non-literal sense. Metaphor understanding has been hypothesized to require mentalizing (Happé, 1993), and fine-grained metaphor comprehension behaviors are captured by RSA models where listeners and speakers reason about each others' beliefs and goals (Kao et al., 2014).

The Metaphor task presents a story with a metaphor, and asks what the speaker intends to convey. For models, the challenges of metaphor comprehension include accessing world knowledge and forming abstract relationships between domains. However, it is possible that the relevant properties of the entities under comparison could emerge through linguistic experience.

Humor. Humor is one of the most distinctive aspects of human conversation, reflecting communicative goals with complex social function (Veatch, 1998; Martin and Ford, 2018). Neu-

roimaging studies suggest that joke understanding is supported by regions in the ToM brain network (Kline Struhl et al., 2018). Behavioral tests also reveal associations between ToM and humor abilities (Aykan and Nalçacı, 2018; Bischetti et al., 2019).

The Humor task presents a joke and asks which punchline makes the joke the funniest.[4] Some theories argue that humor is triggered by linguistic incongruency effects (e.g., Deckers and Kizer, 1975), which might be straightforward for LMs to detect. Recent work has also shown that LMs can explain certain jokes (Chowdhery et al., 2022). However, some of Floyd et al.'s Humor items require complex world knowledge – for example, that slicing a pie into four versus eight pieces does not change the total amount of pie (see Table 1). As such, selecting the funniest punchline is a nontrivial task.

Coherence inferences. Humans also make pragmatic inferences beyond the sentence level – for example, by assuming that consecutive sentences form a logical or sequential relationship. Moss and Schunn (2015) and Jacoby and Fedorenko (2020) find that constructing these discourse relationships loads on regions of the ToM brain network, suggesting a role of ToM in coherence inferences.

The Coherence task presents a pair of sentences, and asks whether the pair forms a coherent story.[5] We assume that LMs' training data, which consists of naturalistic text, is primarily coherent. Therefore, we expect LMs to be able to distinguish between coherent and incoherent sentence pairs (for an in-depth study, see Beyer et al., 2021).

4 Experiments

4.1 Evaluation paradigm

Our evaluation paradigm uses *zero-shot prompting*. Prompting can easily be adapted to all of our seven tasks, allowing us to compare performance across tasks within a model. Prompting also allows us to present models with inputs that are nearly identical to the stimuli seen by humans in Floyd et al.'s experiments, whereas other methods would require converting the stimuli into task-specific formats. We choose zero-shot prompts in order to evaluate the knowledge that emerges through training, and not through in-context adaptation to the task.

[4]Unlike the other tasks, there is no speaker utterance.

[5]This task differs from the others in that there is no speaker utterance, and the answer options are identical across items ("Coherent" or "Incoherent").

Model	# parameters	Training
GPT-2	117M	Autoregressive LM
Tk-Instruct (3B)	3B	Multitask
Tk-Instruct (11B)	11B	Multitask
Flan-T5 (base)	250M	Multitask
Flan-T5 (XL)	3B	Multitask
InstructGPT-3 (ada)	350M (est.)	Multitask, human feedback
text-davinci-002	Unknown	FeedME

Table 2: Models tested in our experiments.

Prompt structure. Each prompt consists of two parts: task instructions, and a query. The instructions are nearly identical to the instructions presented to humans in Floyd et al.'s experiments, prepended with the keyword "Task:". The only other modification is that the original instructions had a final sentence of "Please answer as quickly as possible", which we replaced with a sentence like "The answer options are 1, 2, 3, or 4".[6]

For all tasks except Humor, the query consists of the scenario (prepended with keyword "Scenario:") and question, and then the numbered answer options (prepended with "Options:").[7] The prompt concludes with the keyword "Answer:". Full example prompts are given in Appendix A.

Evaluation. To evaluate a model on a given item, we feed the prompt to the model, and measure the model's probability distribution over tokens conditioned on the prompt. We compare the probabilities of each answer token (e.g., "1", "2", "3", or "4") under this distribution. The model is considered correct on a given item if it assigns highest probability to the correct answer token, among all the possible answer tokens for that item.

We generated 5 versions of each item by randomizing the order of answer options. This was done to control for the base probabilities of the answer tokens. Since we do not analyze generated text, the model results themselves are deterministic.

4.2 Models

We test seven models across four model families, summarized in Table 2.[8] As a baseline, we first test a base **GPT-2** model (117M parameters; Radford et al., 2019), which is trained on an autoregressive language modeling objective.

Second, we test a set of models which are based on T5 (Raffel et al., 2020) and instruction-finetuned

[6]The exact answer options changed according to the task.

[7]For the Humor task, the joke is prepended with "Joke:", and the answer options are prepended with "Punchlines:".

[8]All non-OpenAI models were accessed via Huggingface (Wolf et al., 2020) and run on a single NVIDIA A100 GPU.

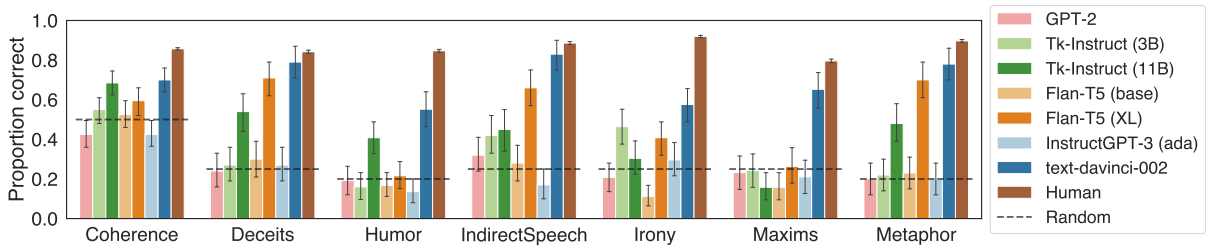

Figure 1: Accuracy for each task. Error bars denote 95% CI. Dashed line indicates task-specific random baseline.

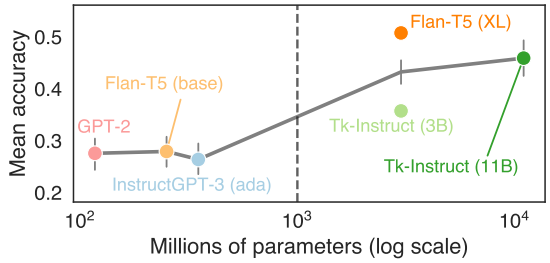

Figure 2: Mean accuracy vs. millions of parameters. Vertical dashed line indicates 1 billion parameters. text-davinci-002 was excluded from this analysis, as the number of parameters is unknown.

on a diverse collection of tasks (Wei et al., 2022). This set of models consists of two **Tk-Instruct** models (3B and 11B; Wang et al., 2022), which were fine-tuned on 1.6K tasks, and two **Flan-T5** models (base: 250M parameters; XL: 3B parameters; Chung et al., 2022), which were fine-tuned on 1.8K tasks. The fine-tuning tasks cover a wide range of categories, such as commonsense reasoning, translation, mathematics, and programming.

Finally, we test two **InstructGPT**-based models (Ouyang et al., 2022) via the OpenAI API: text-ada-001 (350M parameters), which we refer to as InstructGPT-3 (ada); and text-davinci-002, which comes from the GPT-3.5 family of models.[9,10] These models are fine-tuned to follow instructions and align with human feedback.

We compare models to a baseline from 374 humans, collected by Floyd et al. (In prep). Their experiments presented multiple choice questions to humans in nearly identical format to our prompts.

5 Results

We now return to the three questions posed in the Introduction, in each of the following subsections.

5.1 Do models choose the target pragmatic interpretation?

Figure 1 shows the proportion of trials where models and humans select the pragmatic answer. The smallest models (GPT-2, Flan-T5 (base), InstructGPT-3 (ada)) fail to perform above chance. The largest models (Tk-Instruct (11B), Flan-T5 (XL), text-davinci-002) perform above chance on all tasks (except Tk-Instruct (11B) on Maxims), and in some cases near human-level. Overall, models perform worst at the Humor, Irony, and Maxims tasks. Interestingly, these phenomena involve speakers violating listeners' expectations in some way: producing a funny punchline to a mundane story (Humor), stating the direct opposite of the speaker's belief (Irony), or disobeying one of the assumed rules of conversation (Maxims). It may be that models fail to represent certain social expectations that are maintained by human listeners.

Next, we investigated the relationship between model size and accuracy. Figure 2 shows the mean accuracy achieved by each model (averaged across tasks) vs. millions of parameters. The line and error bars denote the mean and 95% CIs, while points represent individual models. We find a coarse effect of model size: there is a stark jump in accuracy after 1B parameters (dashed line). However, model size does not fully explain variance in accuracy: all models with <1B parameters achieve similar accuracy, and Flan-T5 (XL) outperforms Tk-Instruct (3B), despite both having 3B parameters.

5.2 Do models and humans make similar types of errors?

Recall from Section 3 that each item has a set of answer options that correspond to different strategies (Table 1).[11] In addition to the target pragmatic answer (Correct), each item also has a plausible but unlikely literal answer (Literal), as well as distrac-

[9]Parameter estimates come from https://blog.eleuther.ai/gpt3-model-sizes/. Although the size of text-davinci-002 is unknown, we assume that it is larger than InstructGPT-3 (ada).

[10]The OpenAI model results might not be reproducible, but timestamps of API calls can be found in Appendix B.

[11]The exception is Coherence, which is excluded here.

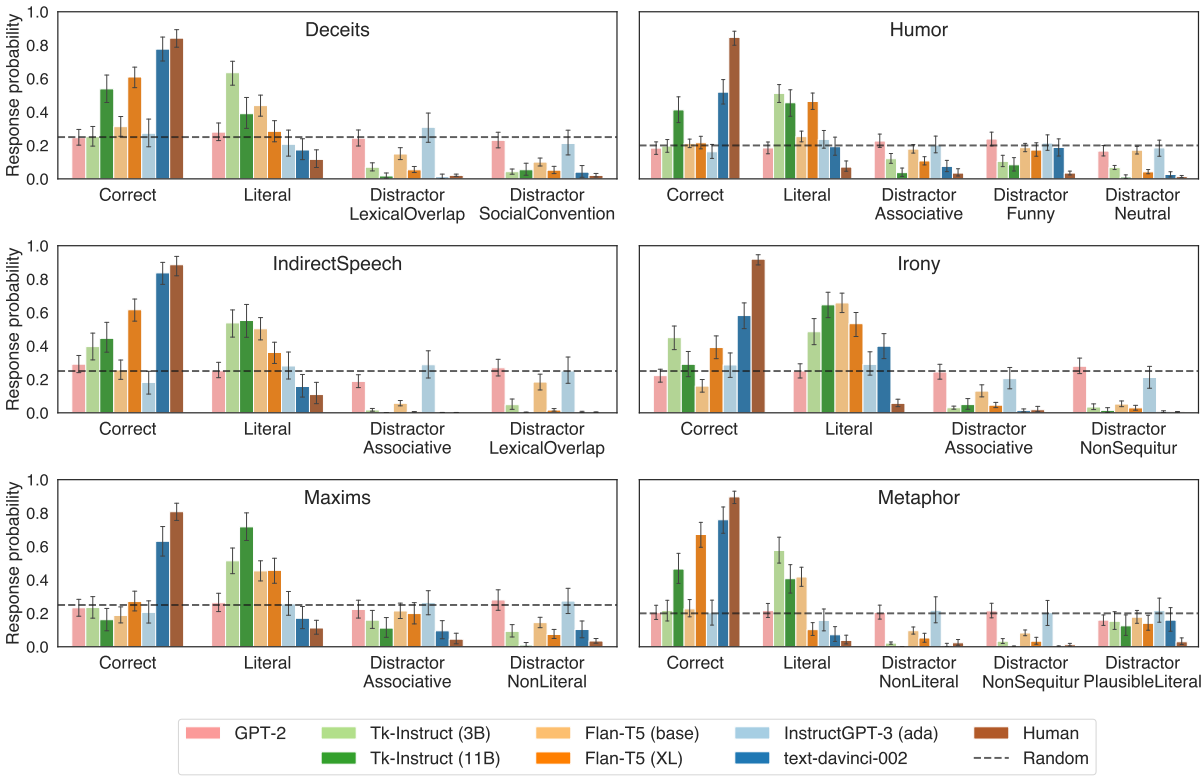

Figure 3: Response distributions across models and humans. Answer options for each task are shown on the x-axis. For models, y-axis denotes probability assigned to each answer option. For humans, y-axis denotes empirical frequency of each answer option being selected. Error bars denote 95% CI. Dashed line indicates random baseline.

tors based on lexical overlap or semantic associations (Distractor*). For each item, we computed the human empirical distribution over answer choices, and compared it to models' probability assigned to the answer tokens (e.g., "1", "2", "3", and "4").

Figure 3 shows the answer distributions for each task. Across tasks, humans primarily select the Correct option, occasionally the Literal option, and rarely the distractors. We find a similar pattern for text-davinci-002, although the model is more likely to select the Literal option in general. The other large models (Tk-Instruct (11B), Flan-T5 (XL)) also generally assign highest probability to the Correct and Literal options, although the distribution looks less human-like. The next-largest models (Tk-Instruct (3B), Flan-T5 (base)) prefer the Literal option, and the remaining models (GPT-2, InstructGPT-3 (ada)) are at chance. These results show that larger models consistently identify the literal interpretation of an utterance, suggesting that their pragmatic failures are unlikely to be explained by a failure to represent basic semantic meaning (for our test materials).

However, even high-performing models occasionally do select the distractor answers, revealing interesting behaviors. For example, in the Metaphor task, text-davinci-002 and Flan-T5 (XL) prefer the DistractorPlausibleLiteral option – which is a figurative reading of the utterance – over the Literal option – which is completely non-figurative. Similarly, in the Humor task, text-davinci-002 is much more likely to select the DistractorFunny option over the other (non-humorous) distractors. This suggests a coarse sensitivity to humor, even if the model selects the human-preferred punchline only 55% of the time (see Figure 1). We take this analysis to illustrate the value of looking beyond binary pragmatic/non-pragmatic response distinctions, and using controlled distractor items to evaluate models' abilities (e.g., McCoy et al., 2019).

5.3 Are models and humans sensitive to similar linguistic cues?

Having found qualitatively similar response patterns between humans and models, we now ask *how* models and humans arrive at pragmatic interpretations, and whether they use similar types of information. We begin with a broad evaluation of the extent to which models and humans rely on linguistic context (Section 5.3.1). We then take a

more granular approach and ask whether model and human performance is correlated at the item level – i.e., if models and humans exhibit similar sensitivity to the cues that make a non-literal interpretation more or less likely (Section 5.3.3).

5.3.1 The role of context

Many cues for enriched language understanding come from the context in which the speaker makes their utterance. However, some aspects of non-literal comprehension might arise given the utterance in isolation, while others are highly sensitive to specific contextual details (e.g., Levinson, 2000). Therefore, we expect that the degree to which humans rely on context to select non-literal interpretations will vary across the tested tasks.

To investigate this variation, we created a new set of stimuli by removing the context stories, leaving only the speaker utterance and final question (e.g., *Dan says, "The dog knocked it over." Why has Dan responded in such a way?*).[12] We re-ran the human experiment on 30 participants, following the protocols of Floyd et al. (In prep)'s original experiment using the no-context modified materials.[13] We also re-ran the three models that achieved highest accuracy on the original items: T*k*-Instruct (11B), Flan-T5 (XL), and text-davinci-002.

Figure 4 shows the mean accuracy difference on the original versus no-context versions of each item.[14] We find that models and humans exhibit a similar qualitative pattern: removing the story leads to the largest degradation for Irony, followed by Deceits and Maxims. This aligns with our intuitions, because in these cases, speakers' utterances can be interpreted either literally or as the complete opposite, based on the specific social situation (e.g., "It is so pleasant here"). In contrast, there are smaller degradations for IndirectSpeech and Metaphor. This suggests that some indirect requests are conventionalized (e.g., "I am getting cold"), although their interpretations may be facilitated by context (e.g., Gibbs, 1979). Similarly, this suggests that metaphor interpretation may draw more upon global knowledge than local context.

5.3.2 Scrambling

Next, we tested whether models rely on syntactic and discourse-level information from the context, or whether they can perform the tasks when ordering cues are removed. We constructed two scrambled versions of each item by randomizing the order of sentences and words. In both versions, the instructions, final question (e.g., *Why has Dan responded in such a way?*), and answer options were unmodified and remained in their original positions. Again, we only tested the best-performing models on these items.

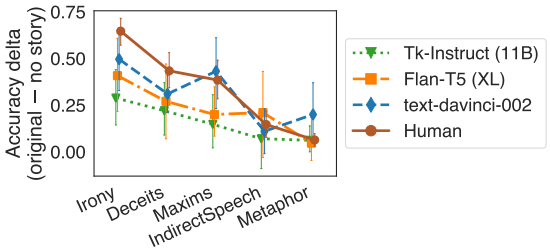

Figure 4: Mean by-item difference in accuracy once story context was removed.

We found that models maintain reasonable performance for most tasks, with the notable exception of Metaphor (Figure 7; Appendix D). This robustness to scrambling accords with prior evidence that models often rely on lexical information without human-like compositionality (e.g., Dasgupta et al., 2018; Nie et al., 2019; McCoy et al., 2019). We expect that scrambling, especially at the word-level, would likely disrupt human performance, but this remains an open empirical question. We leave an investigation of human performance to future work.

5.3.3 Item-level alignment

Up to this point, we analyzed differences across phenomena by averaging over items. However, there is also variance *within* each phenomenon in the types of cues that suggest how the utterances should be interpreted. For example, some items contain explicit descriptions of characters' emotional states (e.g., "Sarah becomes angry"). If models and humans leverage these cues in similar ways, then we would expect to see correlations between model and human performance at the item level.

For each task and model, we compute the Pearson correlation between by-item mean accuracy achieved by humans and by-item mean probability that models assigned to the correct answer (Figure 5). In general, the larger models (T*k*-Instruct (11B), Flan-T5 (XL), text-davinci-002) are better aligned with humans, and the strongest correlations occur for IndirectSpeech, Irony, Maxims, and Metaphor. This suggests that for those tasks, mod-

[12] This manipulation is not compatible with the Humor and Coherence tasks, so they are excluded from this analysis.

[13] Details can be found in Appendix C.1.

[14] See Figure 6 in Appendix C.2 for comparison of raw accuracy scores on the original and no-context items.

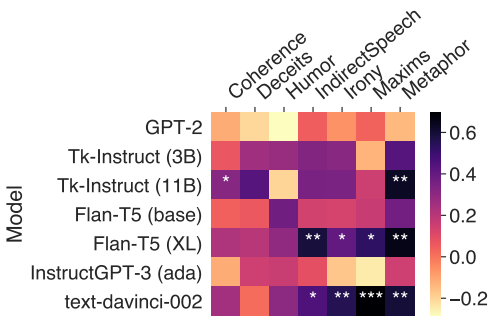

Figure 5: Pearson correlation coefficients between by-item human accuracy and model probability of the correct answer. Cells are marked with significance codes.

els and humans are similarly sensitive to cues that make a non-literal interpretation likely.

6 Discussion

We used an expert-curated set of materials (Floyd et al., In prep) to compare LMs and humans on seven pragmatic phenomena. We found that Flan-T5 (XL) and text-davinci-002 achieve high accuracy and match human error patterns: within incorrect responses, these models tend to select the literal interpretation of an utterance over heuristic-based distractors. We also found preliminary evidence that LMs and humans are sensitive to similar linguistic cues: model and human accuracy scores correlate at the item level for several tasks, and degrade in similar ways when context is removed.

Our results suggest that language models can consistently select the pragmatic interpretation of a speaker's utterance – but how? The models tested in our experiments reflect a variety of learning processes through which pragmatic knowledge could emerge. GPT-2 is trained to learn the distribution of linguistic forms; the Tk-Instruct and Flan-T5 models are pre-trained on a denoising task and fine-tuned on thousands of instruction-based tasks; and the OpenAI models receive signal from human feedback. Our experiments are not designed to tease apart the contributions of these training procedures to models' behaviors. Therefore, we do not intend to make strong claims about the mechanisms by which models learn pragmatics.

A shared feature of our tested models is the lack of explicitly constructed mental state representations. In this sense, our results are potentially compatible with two hypotheses. One possibility is that the models do not have an ability that can be considered an analog of Theory of Mind (ToM). This view is supported by evidence that language models perform poorly on social commonsense and false-belief tasks (Sap et al., 2022), and are remarkably brittle to small perturbations of classic tests (Ullman, 2023). If models truly lack ToM, then their pragmatic behaviors might be explained by inferences based on low-level linguistic cues. Taken a step further, this finding could potentially suggest that certain human pragmatic behaviors arise through inferences based on language statistics, with no need for mental state representations.

A second possibility is that models do have a heuristic version of ToM, which is not explicitly engineered but instead emerges as a by-product of optimizing for other objectives (such as linguistic prediction). Since language contains many descriptions of agents' beliefs, emotions, and desires, it may be beneficial – perhaps even necessary – to induce representations of these mental states in order to learn a generative model of linguistic forms. Indeed, Andreas (2022) argues that whereas language models have no explicit representation of communicative intents, they can infer approximate representations of the mental states of the agents that produce a given linguistic context. If this hypothesis is true, however, it would still remain unclear whether ToM is *necessary* to support the pragmatic behaviors tested in our evaluation materials.

Our experiments do not differentiate between these two hypotheses. However, fine-grained behavioral evaluations – such as those presented in this work – are important for revealing models' capabilities and weaknesses, and offer a first step toward understanding how pragmatic behaviors can be supported. A promising direction for future work is to test models with a wider range of training objectives, or even new architectures, such as distinct language and social reasoning modules (see Mahowald et al., 2023). In addition, although there is evidence for the role of mentalizing in our tested pragmatic phenomena (see Section 3.1), one limitation of our stimuli is that they were not specifically designed to require ToM. New datasets that perform targeted manipulations of ToM alongside tests of language comprehension could help reveal how linguistic experience and ToM jointly support pragmatic behaviors.

Acknowledgments

We would like to thank the anonymous reviewers as well as Roger Levy, Christopher Potts, and

Josh Tenenbaum for their constructive feedback. We also thank Quinn Langford for help with coding details of the stimuli. This work was in part supported by a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT. J.H. is supported by an NSF Graduate Research Fellowship (#1745302) and an NSF Doctoral Dissertation Research Improvement Grant (BCS-2116918). S.F. is funded by the NSF SPRF (#2105136). E.F. was additionally supported by NIH award R01-DC016950 and by research funds from the McGovern Institute for Brain Research and the Department of Brain and Cognitive Sciences.

Limitations

We note several methodological limitations with our experiments. First, since the evaluation materials were manually crafted, there is a rather small number of items (compared to the size of automatically generated NLP benchmarks). Small evaluation sets can introduce issues of statistical power (Card et al., 2020) and introduce bias based on lexical items. We feel this is not a major concern, because (1) our materials are validated by expert researchers; (2) models can be directly compared to humans in Floyd et al.'s experiments; and (3) in practice, there is enough signal to distinguish between the tested models.

Second, we only evaluate models on English-language materials, and some of the tasks were designed based on norms of communication and social interaction in Western cultures. As pragmatics can vary widely across language and cultures (Li, 2012; Rubio-Fernandez and Jara-Ettinger, 2020; Floyd, 2021; Brown et al., 2021; Dideriksen et al., 2022), an important direction for future work is to evaluate pragmatics beyond English (Ameka and Terkourafi, 2019; Blasi et al., 2022).

Third, aside from the OpenAI API models, we were only able to test models with $\leq$11B parameters due to limited computational resources. Models with parameter sizes between 11B and the size of text-davinci-002 could exhibit qualitatively different behaviors.

Finally, we emphasize that it is impossible to predict how models will respond to an arbitrary input. Therefore, we caution against extrapolating from our results and expecting that models will behave "pragmatically" in downstream applications. This is especially true for models behind the OpenAI API, and text-davinci-002 in particular, for which very little is publicly known about the training protocol.

Ethics statement

Language technologies have the potential to cause harm at the individual and societal levels. Large language models (LLMs), which are typically trained on vast amounts of internet text, have been shown to perpetuate stereotypes based on gender, race, and sexual orientation. Applications using LLMs could reinforce systematic discrimination and amplify existing socioeconomic inequities. For example, LLMs could perpetuate social biases by assisting with hiring decisions or legal rulings.

The remarkable fluency of LLM-generated text also poses risks for the general public. LLMs have long been used to generate text that is difficult to distinguish from human-written text, raising concerns about detecting fake news and misinformation. Recently, LLMs have been used to synthesize knowledge – for example, by answering scientific questions (Taylor et al., 2022) or acting as search engines (Shah and Bender, 2022). Using LLMs as knowledge-providers could tremendously impact the nature of human collaboration and work, raising the need for model transparency and explainability.

References

Felix K. Ameka and Marina Terkourafi. 2019. What if…? Imagining non-Western perspectives on pragmatic theory and practice. *Journal of Pragmatics*, 145:72–82.

Jacob Andreas. 2022. Language Models as Agent Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Clara Andrés-Roqueta and Napoleon Katsos. 2017. The Contribution of Grammar, Vocabulary and Theory of Mind in Pragmatic Language Competence in Children with Autistic Spectrum Disorders. *Frontiers in Psychology*, 8.

Ian Apperly. 2011. *Mindreaders: The cognitive basis of "Theory of Mind"*. Psychology Press, New York.

Salvatore Attardo. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics*, 32(6):793–826.

John L. Austin. 1975. *How to do things with words*.

Simge Aykan and Erhan Nalçacı. 2018. Assessing Theory of Mind by Humor: The Humor Comprehension and Appreciation Test (ToM-HCAT). *Frontiers in Psychology*, 9.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Leon Bergen, Roger Levy, and Noah D. Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Luca Bischetti, Irene Ceccato, Serena Lecce, Elena Cavallini, and Valentina Bambini. 2019. Pragmatics and theory of mind in older adults' humor comprehension. *Current Psychology*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170. Publisher: Elsevier.

W.C. Booth. 1974. *A Rhetoric of Irony*. Literature/Criticism - The University of Chicago Press. University of Chicago Press.

Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Two Minds, One Dialog: Coordinating Speaking and Understanding. In Brian H. Ross, editor, *Psychology of Learning and Motivation*, volume 53, pages 301–344. Academic Press.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Penelope Brown, Mark A. Sicoli, and Olivier Le Guen. 2021. Cross-speaker repetition and epistemic stance in Tzeltal, Yucatec, and Zapotec conversations. *Journal of Pragmatics*, 183:256–272.

Stephen A. Butterfill and Ian A. Apperly. 2013. How to Construct a Minimal Theory of Mind. *Mind & Language*, 28(5):606–637. Publisher: John Wiley & Sons, Ltd.

Carl Camden, Michael T. Motley, and Ann Wilson. 1984. White lies in interpersonal communication: A taxonomy and preliminary investigation of social motivations. *Western Journal of Speech Communication*, 48(4):309–325. Publisher: Routledge.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv preprint.

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.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. Evaluating Compositionality in Sentence Embeddings. In *Proceedings of the Cognitive Science Society*.

Lambert Deckers and Philip Kizer. 1975. Humor and the Incongruity Hypothesis. *The Journal of Psychology*, 90(2):215–218.

Judith Degen. 2023. The Rational Speech Act Framework. *Annual Review of Linguistics*, 9(1):519–540.

Marie-Julie Demedardi, Claire Brechet, Edouard Gentaz, and Catherine Monnier. 2021. Prosocial lying

in children between 4 and 11 years of age: The role of emotional understanding and empathy. *Journal of Experimental Child Psychology*, 203:105045.

Christina Dideriksen, Morten H Christiansen, Mark Dingemanse, Malte Højmark-Bertelsen, Christer Johansson, Kristian Tylén, and Riccardo Fusaroli. 2022. Language specific constraints on conversation: Evidence from Danish and Norwegian. PsyArXiv preprint.

Judit Dombi, Tetyana Sydorenko, and Veronika Timpe-Laughlin. 2022. Common ground, cooperation, and recipient design in human-computer interactions. *Journal of Pragmatics*, 193:4–20.

Ivan Enrici, Bruno G. Bara, and Mauro Adenzato. 2019. Theory of Mind, pragmatics and the brain: Converging evidence for the role of intention processing as a core feature of human communication. *Pragmatics & Cognition*, 26(1):5–38.

Sammy Floyd, Olessia Jouravlev, Zachary Mineroff, Leon Bergen, Evelina Fedorenko, and Edward Gibson. In prep. Deciphering the structure of pragmatics: A large-scale individual differences investigation.

Simeon Floyd. 2021. Conversation and Culture. *Annual Review of Anthropology*, 50(1):219–240. Publisher: Annual Reviews.

Michael C Frank, Andrés Goméz Emilsson, Benjamin Peloquin, Noah D. Goodman, and Christopher Potts. 2018. Rational speech act models of pragmatic reasoning in reference games. PsyArXiv preprint.

Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998.

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1):3–44.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Raymond W. Gibbs. 1979. Contextual effects in understanding indirect requests. *Discourse Processes*, 2(1):1–10. Publisher: Routledge.

Herbert P. Grice. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Francesca G.E. Happé. 1993. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119.

Cecilia Heyes. 2014. Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, 9(2):131–143.

Laurence R. Horn. 1972. *On the semantic properties of logical operators in English*. PhD Thesis, University of California Los Angeles.

Yik Kwan Hsu and Him Cheung. 2013. Two mentalizing capacities and the understanding of two types of lie telling in children. *Developmental Psychology*, 49:1650–1659.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Nir Jacoby and Evelina Fedorenko. 2020. Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. *Language, Cognition and Neuroscience*, 35(6):780–796.

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.

Justine T. Kao, Leon Bergen, and Noah D. Goodman. 2014. Formalizing the Pragmatics of Metaphor Understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Justine T. Kao and Noah D. Goodman. 2014. Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Melissa Kline Struhl, Jeanne Gallée, Zuzanna Balewski, and Evelina Fedorenko. 2018. Understanding jokes draws most heavily on the Theory of Mind brain network. PsyArXiv preprint.

Michal Kosinski. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. arXiv preprint.

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. 2022. Concadia: Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4684, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

G. Lakoff and M. Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the Evaluation of Theory of Mind through Question Answering. 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 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Alan M. Leslie, Ori Friedman, and Tim P. German. 2004. Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, 8(12):528–533.

Stephen Levinson. 2000. *Presumptive meaning: The theory of generalized conversational implicature*. MIT Press.

Elissa Li, Sebastian Schuster, and Judith Degen. 2021. Predicting Scalar Inferences From "Or" to "Not Both" Using Neural Sentence Encoders. In *Proceedings of the Society for Computation in Linguistics*, volume 4.

Jin Li. 2012. *Cultural Foundations of Learning: East and West*. Cambridge University Press, Cambridge.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. Place: Cambridge, MA Publisher: MIT Press.

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*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Eleonore Lumer and Hendrik Buschmeier. 2022. Modeling Social Influences on Indirectness in a Rational Speech Act Approach to Politeness. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: A cognitive perspective. arXiv preprint.

R.A. Martin and T. Ford. 2018. *The Psychology of Humor: An Integrative Approach*. Academic Press.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Julian Michael. 2020. To Dissect an Octopus: Making Sense of the Form/Meaning Debate.

Jarrod Moss and Christian D. Schunn. 2015. Comprehension through explanation as the interaction of the brain's coherence and cognitive control networks. *Frontiers in Human Neuroscience*, 9.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing Compositionality-Sensitivity of NLI Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Christopher Potts. 2020. Is it possible for language models to achieve language understanding?

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank. 2016. Embedded Implicatures as Pragmatic Inferences under Compositional Lexical Uncertainty. *Journal of Semantics*, 33(4):755–802.

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.

Paula Rubio-Fernandez. 2021. Pragmatic markers: the missing link between language and Theory of Mind. *Synthese*, 199(1):1125–1158.

Paula Rubio-Fernandez and Julian Jara-Ettinger. 2020. Incrementality and efficiency shape pragmatics across languages. *Proceedings of the National Academy of Sciences*, 117(24):13399–13404.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. arXiv preprint.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Ayse Pinar Saygin and Ilyas Cicekli. 2002. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

John R. Searle. 1975. Indirect Speech Acts. In *Speech Acts*, pages 59–82. Brill, Leiden, The Netherlands.

Chirag Shah and Emily M. Bender. 2022. Situating Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232, New York, NY, USA. Association for Computing Machinery. Event-place: Regensburg, Germany.

Nicola Spotorno, Eric Koun, Jérôme Prado, Jean-Baptiste Van Der Henst, and Ira A. Noveck. 2012. Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, 63(1):25–39.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI Models' Performance on Figurative Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. arXiv preprint.

Michael Henry Tessler and Michael Franke. 2018. Not unreasonable: Carving vague dimensions with contraries and contradictions. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Anna Trosborg, editor. 2010. *Pragmatics across Languages and Cultures*. De Gruyter Mouton.

Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv preprint.

Thomas C. Veatch. 1998. A theory of humor. *Humor*, 11(2):161–216.

Corrie Vendetti, Deepthi Kamawar, and Katherine E. Andrews. 2019. Theory of mind and preschoolers' understanding of misdeed and politeness lies. *Developmental Psychology*, 55(4):823–834.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

D. Wilson and D. Sperber. 2012. *Meaning and Relevance*. Cambridge University Press.

Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua*, 87(1):53–76.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin

Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Erica J. Yoon, Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank. 2016. Talking with tact: Polite language as a balance between informativity and kindness. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Erica J. Yoon, Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank. 2020. Polite Speech Emerges From Competing Social Goals. *Open Mind*, 4:71–87.

George Yule. 1996. *Pragmatics*, 1 edition. Oxford Introduction to Language Study. Oxford University Press.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809.

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 Example prompts

This section contains example prompts for each task in our experiments. See Section 3 and Table 1 for details on the materials, and Section 4.1 for discussion of how prompts were constructed.

A.1 Deceits

Task: You will read short stories that describe two characters interacting with each other. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. Your task is to decide why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: Henry is sitting at his desk and watching TV, and reluctantly switches off the TV with the remote control and picks up a textbook. Shortly after, his mother comes in the room and asks, "What have you been doing up here?" Henry responds: "Reading." Why has Henry responded in such a way?
Options:
1) He has been reading for some time.
2) He does not want to offend his mom by not reading the books that she gave him.
3) He does not want to get into trouble for not studying.
4) He wants his mom to believe that he has been watching TV.
Answer:

A.2 IndirectSpeech

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, 2, 3, or 4.

Scenario: Nate is about to leave the house. His wife points at a full bag of garbage and asks: "Are you going out?" What might she be trying to convey?
Options:
1) She wants Nate to spend more time with the family.
2) She wants to know Nate's plans.
3) She wants Nate to take the garbage out.
4) She wants Nate to bring his friends over.
Answer:

A.3 Irony

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, 2, 3, or 4.

Scenario: It is a holiday. Stefan and Kim are sitting in the backseat of the car. They are fighting all the time. Their father says: "Oh, it is so pleasant here." What did the father want to convey?
Options:
1) He enjoys listening to his kids fighting.
2) He remembers about his wife's birthday.
3) He does not want to listen to his kids' arguments.
4) AC gives them some needed cool.
Answer:

A.4 Maxims

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 why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: Leslie and Jane are chatting at a coffee shop. Leslie asks, "Who was that man that I saw you with last night?" Jane responds, "The latte is unbelievable here." Why has Jane responded like this?
Options:
1) She does not want to discuss the topic that Leslie has raised.
2) The man who Leslie saw makes unbelievable lattes.
3) She thinks that it is the best latte in the town.
4) A coffee break is not a good time to discuss men.
Answer:

A.5 Metaphor

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 to each question. The answer options are 1, 2, 3, 4, or 5.

Scenario: Andrew and Bob were discussing the investment company where Andrew works. Bob said: "The investors are squirrels collecting nuts." What does Bob mean?
Options:
1) The investors dress and eat well.
2) Squirrels were hired to work in the company.
3) Bob is allergic to nuts.
4) They buy stocks hoping for future profit.
5) The investors enjoy picking nuts as much as squirrels do.

Answer:

A.6 Humor

Task: You will read jokes that are missing their punch lines. A punch line is a funny line that finishes the joke. Each joke will be followed by five possible endings. Please choose the ending that makes the joke funny. The answer options are 1, 2, 3, 4, or 5.

Joke: Martha walked into a pastry shop. After surveying all the pastries, she decided on a chocolate pie. "I'll take that one," Martha said to the attendant, "the whole thing." "Shall I cut it into four or eight pieces?" the attendant asked.
Punchlines:
1) Martha said, "My leg is hurting so much."
2) Martha said, "Four pieces, please; I'm on a diet."
3) Martha said: "Well, there are five people for dessert tonight, so eight pieces will be about right."
4) Then the attendant squirted whipped cream in Martha's face.
5) Martha said, "You make the most delicious sweet rolls in town."
Answer:

A.7 Coherence

Task: You will read pairs of sentences. Reach each pair and decide whether they form a coherent story. The answer options are 1 or 2.

Scenario: Cleo brushed against a table with a vase on it. She decided to study harder to catch up.
Options:
1) Incoherent
2) Coherent
Answer:

B Timestamps of OpenAI model queries

Table 3 shows timestamps of requests sent to the OpenAI API.

Model	Phenomenon	Timestamp
text-ada-001	Coherence	2022-10-11 12:28 -0400
text-ada-001	Deceits	2022-10-11 12:28 -0400
text-ada-001	IndirectSpeech	2022-10-11 12:28 -0400
text-ada-001	Irony	2022-10-11 12:28 -0400
text-ada-001	Humor	2022-10-11 12:28 -0400
text-ada-001	Maxims	2022-10-11 12:29 -0400
text-ada-001	Metaphor	2022-10-11 12:29 -0400
text-davinci-002	Coherence	2022-10-11 11:56 -0400
text-davinci-002	Deceits	2022-10-11 11:55 -0400
text-davinci-002	IndirectSpeech	2022-10-11 11:55 -0400
text-davinci-002	Irony	2022-10-11 11:54 -0400
text-davinci-002	Humor	2022-10-11 11:53 -0400
text-davinci-002	Maxims	2022-10-11 11:56 -0400
text-davinci-002	Metaphor	2022-10-11 11:57 -0400

Table 3: Timestamps of OpenAI API model queries.

C No-context analysis

C.1 Details of human experiments

Below, we discuss details of the no-context human experiments described in Section 5.3.1. This study was approved by the Institutional Review Board at the home institution of the authors (protocol 2010000243).

Participants. We collected data from 30 participants using Amazon.com's Mechanical Turk. All participants were recruited from IP addresses in the US, Canada, and other English-speaking countries and passed a brief English proficiency task to participate. We pre-screened participants using a qualification task in which they were asked to perform 10 simple sentence completions, which were judged for basic levels of coherence and grammaticality. Participants were paid 7 USD for completing the study, which took around 20 minutes to complete. The resulting hourly rate was around 21 USD, which is well above federal minimum wage in the United States.

Procedure. Participants completed these tests during one individual testing session. After giving informed consent, which included assurance of anonymity, participants were shown instructions and a training trial, in which they were told they would be answering questions about a character in a short interaction. They then saw 105 trials (similar to those described in Appendix A), without the scenario context. For example:

Bob said: "The investors are squirrels collecting nuts." What does Bob mean?
1) The investors dress and eat well.
2) Squirrels were hired to work in the company.
3) Bob is allergic to nuts.
4) They buy stocks hoping for future profit.
5) The investors enjoy picking nuts as much as squirrels do.

Items were presented within blocks according to their phenomenon, as in Floyd et al.'s (In prep) original experiments. Blocks and items were presented in a random order.

C.2 Raw accuracy scores

Figure 6 shows accuracy scores achieved by humans and the three best-performing models on the original (shaded bars) and no-context (empty bars) versions of the test items.

D Sentence- and word-level scrambling

Figure 7 shows accuracy scores achieved by the three best-performing models on each task, across three scrambling conditions: none (original, unmodified items), sentence-level, and word-level. Example prompts are provided below.

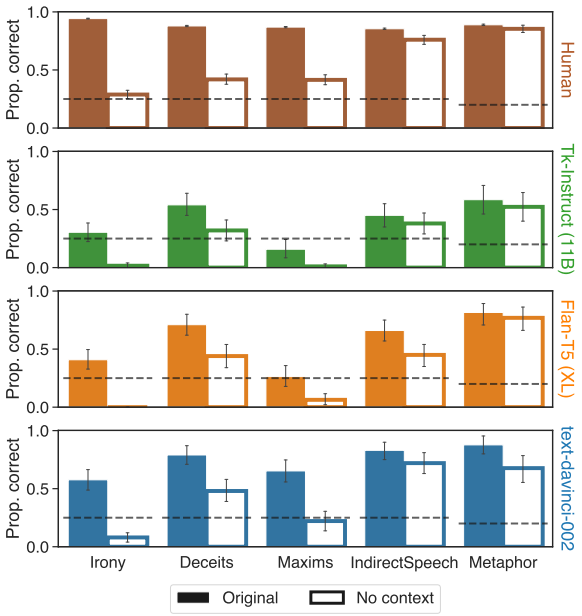

Figure 6: Proportion of items where humans and models select the correct pragmatic answer, on both original (shaded bars) and no-context (empty bars) versions.

D.1 Sentence-level scrambled prompt

Task: You will read short stories that describe two characters interacting with each other. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. Your task is to decide why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: Dan says,"The dog knocked it over." The vase falls down on the floor and breaks. He brushes against his mother's vase. When Dan's mother comes home, she asks Dan: "What happened to my vase?" Dan is playing in the living room. Why has Dan responded in such a way?
Options:
1) Dan does not want his mom to be angry with him for breaking the vase.
2) Dan finds this vase ugly and wants to get rid of it.
3) Dan wants his mom to know that he knocked it over.
4) Dan thinks that the dog has knocked over the vase.
Answer:

D.2 Word-level scrambled prompt

Task: You will read short stories that describe two characters interacting with each other. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. Your task is to decide why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: to happened Dan "The against in it she comes "What living Dan the vase floor on down The Dan: He dog my brushes vase?" mother When falls breaks. vase. and playing room. his asks knocked says, home, over." the mother's is Dan's Why has Dan responded in such a way?
Options:
1) Dan does not want his mom to be angry with him for breaking the vase.
2) Dan finds this vase ugly and wants to get rid of it.
3) Dan wants his mom to know that he knocked it over.
4) Dan thinks that the dog has knocked over the vase.
Answer:

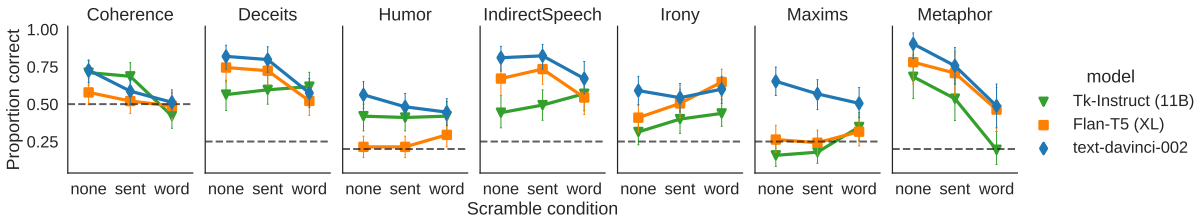

Figure 7: Model performance across scrambling conditions (none = original, unmodified items). Error bars denote 95% CI. Dashed line indicates random baseline.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
"Limitations" section after Section 6

☑ A2. Did you discuss any potential risks of your work?
"Limitations" and "Ethics statement" sections after Section 6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract; Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Section 3; Section 4.2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All artifacts are open to scientific research use.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
All data and models used in our study were intended for scientific research.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We only used publicly available pre-trained models.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix C

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix C

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix C

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix C

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix C