

What Do You Mean? Exploring the Alleged Theory of Mind Abilities of Large Language Models

Agnese Lombardi

University School of Advanced Studies (IUSS)
Pavia, Italy

agnese.lombardi@iusspavia.it

Alessandro Lenci

University of Pisa
Pisa, Italy

alessandro.lenci@unipi.it

Abstract

This study explores the capacity of Large Language Models (LLMs) to perform tasks requiring Theory of Mind (ToM), a critical component of pragmatic language understanding. Although previous work suggests that LLMs may exhibit emergent ToM abilities, this research examines whether such capabilities genuinely involve reasoning about beliefs or merely reflect the reliance on shallow statistical cues. Through a series of controlled experiments featuring indirect speech acts and verbal irony, we assess how belief contexts influence LLM interpretations. The results reveal that, although LLMs occasionally succeed in decoding communicative intentions, their performance is not attributable to human-like ToM reasoning. This work underscores the limitations of LLMs in simulating human-like ToM and offers insight into their interpretive biases, contributing to a deeper understanding of their linguistic capabilities.¹

1 Introduction

An hallmark of human communication is the pragmatic ability to reconstruct the intended meaning of an utterance (that is, what the speaker intends to convey with it in a given situation), even when it goes beyond its literal interpretation (Grice, 1975; Levinson, 2000; Sperber and Wilson, 2002; Pinker et al., 2008). In this context, the **Theory of Mind** (ToM) – the ability to attribute mental states to others – is generally regarded to be essential for understanding intended meaning and crucial for successful social interactions in everyday life (de Villiers, 2007).

Recent advancements in Large Language Models (LLMs) and the unforeseen emergence of

new capabilities in tasks for which these models were neither explicitly trained nor instructed have prompted increased research into whether ToM could be among these newly emerging skills (Marchetti et al., 2023; Kosinski, 2024; Bubeck et al., 2023).

Determining whether these models can handle tasks that require the use of ToM would enable us to: 1) revise our understanding of ToM development from a linguistic or psycholinguistic perspective, 2) assess whether the language generated by models can be compared to human language and whether they use the same cues to perform tasks and produce texts, and 3) evaluate their effectiveness as communicators.

ToM is frequently implicated in the correct interpretation of **figurative language** (Melogno et al., 2021; Gavián and García-Albea, 2011). However, psycholinguistic research has shown that, when processing utterances with non-literal meanings, cues from the linguistic and situational context may allow hearers to access a non-literal interpretation directly, without first computing a literal one. This processing route is known as the *direct access view* (Wimmer et al., 2016; Gibbs, 2002). As a consequence, non-literal interpretations often function as the default, especially in the case of conventionalized expressions, thereby reducing, or even eliminating, the need to recruit ToM to derive an indirect meaning. This observation extends to phenomena traditionally assumed to require mentalizing, such as **indirect speech acts** (ISAs) (Gibbs, 1986b; Marocchini and Domaneschi, 2022) and **verbal irony** (Ackerman, 1983; Juez, 1998; Burnett, 2015).

In humans, an expression is *conventionalized* when a particular meaning (non-literal in our case) is strongly associated with a specific context due to prior knowledge and situational familiarity. For computational models, we assume that conventionalization is instead driven by the frequency

¹Dataset and Code on GitHub: https://github.com/agneselombardi/Alleged_tom.

with which expressions occur in the training data and are associated with particular meanings.

Building on this body of evidence, we hypothesize that LLMs, like humans, may rely on conventionalization when solving well-known ToM tasks, such as false-belief scenarios, especially when these tasks contain conventionalized expressions with figurative meanings. If this is the case, two consequences follow:

1. evaluations of LLMs on ToM benchmarks may be misleading if the degree of conventionalization is not controlled, as model outputs may reflect the exploitation of conventionalized patterns rather than genuine mentalizing abilities;
2. the ability of LLMs in processing figurative language might be simply due to their encoding frequent figurative meanings, rather than to a genuine capability to represent or reason about others' mental states.

To test these hypotheses, we adopt a false-belief paradigm in which we manipulate the degree of conventionalization associated with a target expression. In the first scenario, a figurative utterance such as *Can you open the window?* can be straightforwardly linked to its conventional non-literal interpretation. In the second scenario, we manipulate the beliefs of a character in the story so as to weaken this conventionalized association and promote a more literal interpretation of the same utterance. The experiment is designed to address the following research questions: 1) To what extent do LLMs accurately perform tasks that genuinely require human mentalizing? 2) Does conventionalization influence LLMs' resolution of false-belief tasks? In addition, we conduct an exploratory analysis to investigate: 3) Whether the size of the language model significantly affects performance accuracy.

2 Related Work

Mahowald et al. (2024) argue that models employing language in human-like ways, such as LLMs, must develop two distinct competences: **formal linguistic competence** – knowledge of linguistic rules and patterns – and **functional linguistic competence** – understanding and using language in real world contexts. These competences require the development of specialized mechanisms. This

highlights the need for separate testing and evaluation of formal and functional competences through benchmarking techniques and stimuli that isolate either of them.

Our research aligns with the perspective of Hu et al. (2023), who compare language models and humans across seven pragmatic phenomena using zero-shot prompting to evaluate the models' ability to interpret non-literal language.

A growing body of work investigates how large language models handle different types of figurative language, often comparing their performance to that of humans. This includes studies on the role of context in sarcasm recognition (Jang et al., 2024), idiomatic language processing (De Luca Fornaciari et al., 2024), and euphemism detection (Firsich and Rios, 2024), as well as work applying the Winograd Schema Challenge (Levesque et al., 2012) and its variants (Sun and Emami, 2024; Park et al., 2024; Gevers et al., 2025). While these studies provide valuable insights into how LLMs process figurative language, they do not explicitly examine these phenomena in relation to ToM. In contrast, our study differs from the works above for its focus on pragmatic phenomena that require explicit **mentalizing (i.e., ToM) abilities**, that is reasoning about the others' beliefs and intentions. Specifically, we seek to determine whether LLMs interpret non-literal language by inferring others' beliefs, or rely on other conventionalized cues to derive the intended meaning. Existing benchmarks for assessing ToM in models have predominantly addressed social commonsense (Sap et al., 2019), situational descriptions (Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2022; Zhou et al., 2023), or social interactions (Kim et al., 2023; Li et al., 2023).

Most studies in ToM primarily investigate 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 the observations of that agent (Jones et al., 2023; Shapira et al., 2023). These works focus on first or second-order ToM, while Hu et al. (2023) investigates third-order ToM and beyond.

Some studies argue that LLMs achieve human-level performance on ToM evaluation tasks (Kosinski, 2024; Strachan et al., 2024; Street et al., 2026). In contrast, other work suggests that LLMs do not exhibit ToM abilities in the robust, flexible, and generalizable manner characteristic of hu-

INDIRECT SPEECH ACTS		VERBAL IRONIES	
Indirect Requests	"Can you turn on the heater?"	Sarcasm	"Nice shot"
Indirect Suggestions	"Why don't you eat your meal?"	Hyperboles	"Haven't seen you in ages"
Indirect Declinations	"Do you want some more dessert?" "I'm full"	Rethorical Questions	"Is the Pope Catholic?"
Indirect Threats	"You will pay for this"		

Figure 1: Linguistic phenomena tested in the present study: Indirect Speech Acts and Verbal Ironies

mans. For example, LLMs have been shown to fail on minor modifications of the “unexpected contents” task that would presumably be trivial for human participants (Ullman, 2023), as well as on adversarially constructed examples (Shapira et al., 2023). As argued in Ma et al. (2023); Chen et al. (2025); Hu et al. (2025), these divergent findings can be attributed both to differences in evaluation methodologies and to variation in prompt design across studies, as also observed in Strachan et al. (2024). Moreover, many reported results may be explained by the exploitation of statistical shortcuts, stemming from the limited scope of tasks commonly used to assess ToM in deep learning systems (Aru et al., 2023). To the best of our knowledge, the present study is the first to explicitly control for the role of conventionalization across different belief contexts in tasks requiring ToM reasoning.

3 Experimental Materials

3.1 Linguistic phenomena

It is a common assumption that Indirect Speech Acts and Verbal Ironies require mentalizing for their highly context-dependent nature.

Indirect Speech Acts (ISAs) (Austin, 1975; Searle, 1975) occur when the speaker’s communicative intent (e.g., a request) differs from the literal meaning of their words (e.g., a question). They rely on the listener’s ability to infer the actual intent from the context and linguistic cues, often used to achieve politeness or convey subtle meanings indirectly.

We assessed four distinct types of ISAs, each differing in their level of conventionalization: **Indirect Requests** (Trott and Bergen, 2019), **Indirect Suggestions**, **Indirect Declinations**, and **Indirect Threats** (Figure 1). Accurately interpreting these ISAs may not solely depend on the ability to encode and represent others’ knowledge states and intentions, since it could also arise from conventional associations. In our experiments, we controlled for both literal and non-literal interpretations, so that differentiating between them really necessitates considering ToM, as **the relevant contexts differ based on the mutual beliefs held by the speaker and listener**. Conventionalization degree varies across phenomena. Indirect requests and indirect suggestions are treated as highly conventionalized, as all stimuli are introduced by fixed conventional formulae: “Can you” for indirect requests and “Why don’t you” for indirect suggestions (Marocchini and Domaneschi, 2022). Indirect threats occupy an intermediate position, as the stimuli were drawn from the most commonly recognized threatening expressions, validated through a Reddit survey in which participants judged whether a given expression would be perceived as a threat. Indirect declinations are considered the least conventionalized, since their interpretation relies primarily on situational context: an expression such as “I am full” functions as an indirect declination only within an appropriate context (e.g., following an offer of more food) and carries no such meaning in isolation.

Verbal irony (Skalicky, 2024) encompasses utterances where the speaker feigns ignorance of an unfulfilled expectation, whether based on personal anticipations or cultural norms. This pretense creates a discrepancy between the spoken words (reflecting the expectation) and the actual situation. According to the graded salience hypothesis (Giora and Fein, 1999a), both literal and figurative meanings can be encoded in the lexicon based on their conventionality (e.g., *Thanks a lot* is used both sincerely and ironically).² Consequently, figurative meanings of conventional ironic expressions are more accessible and easier to comprehend than those of situation-specific ironic expressions, which may require additional contextual cues or processing. We investigated three kinds of verbal ironies: **Sarcasm** (Gibbs, 1986a), **Hyperboles** (Aguert et al., 2018; Carston and Wearing, 2015), and **Rethorical Questions** (Figure 1).

For verbal irony, conventionalization is

²Ackerman (1983) suggested that some ironic remarks might have become idiomatic, as children interpreted them ironically with minimal contextual cues.

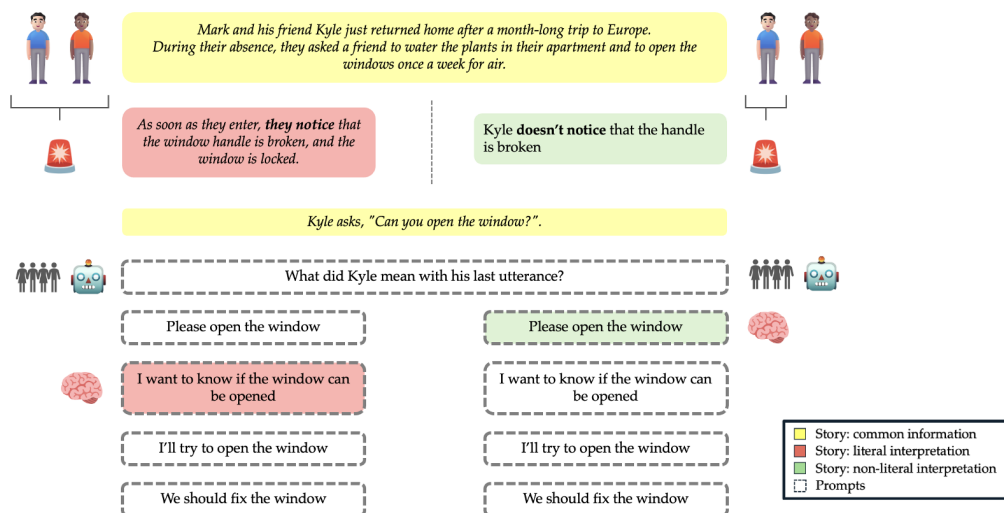


Figure 2: Stimulus from Indirect Requests. It illustrates two interpretations of the same story: non-literal (in green) and literal (in red). Each story has two versions differing for the knowledge states of the two characters. The yellow boxes highlight the common information retained across both versions. Both models and humans are presented with the same question and must select one of four possible answers.

achieved through conventional remarks following story contexts with varying degrees of salience (Giora and Fein, 1999b): higher salience (e.g., a negative context followed by a positive outcome) or lower salience (e.g., a positive context aligned with a positive outcome). This applies to sarcasm and indirect hyperbole, where interpretation remains context-dependent. Rhetorical questions, by contrast, can be considered highly conventionalized even in the absence of context: expressions such as “Is the Pope Catholic?” are virtually always interpreted as rhetorical and would rarely, if ever, be used otherwise.

3.2 Stimuli

The stimuli used in our study were inspired by the *false-belief task* design, and adapted from Trott and Bergen (2019), who created pragmatically ambiguous utterances to conclude stories. In these stories, an **obstacle** to fulfil a potential request was introduced, with the speaker either present or absent when the obstacle information was disclosed. We incorporated Trott and Bergen (2019)’s stimuli on Indirect Requests into our experiment, categorizing them into two groups: IR-Os (derived from Trott and Bergen (2019) IR-Ns (created by us). Stimuli for all other phenomena were newly developed to ensure they were not included in the

pretraining material.

Each story in the dataset presents two interpretations (Figure 2): in one version the belief scenario is congruent with the literal interpretation preventing any other non-literal meaning, while in other version the belief context triggers an indirect interpretation. The concluding target utterances in each story are typically conventionalized expressions of the non-literal interpretation.

This design enabled us to assess whether models adjusted their interpretations according to the different belief contexts of the two participants in the story. Successful task completion requires both humans and models to reason about the different epistemic states (i.e., beliefs, intentions and emotions) of the two characters. Indeed, the narrative remains consistent across both literal and non-literal versions, the only distinction being the characters’ knowledge of the obstacle.

In the non-literal version of the story, conventionalization is expected to play a critical role, as the belief context triggers the conventionalized (non-literal) meaning of the figurative expression. In contrast, **in the literal version, knowledge of the obstacle licenses a non-conventionalized interpretation** of the same utterance. Thus, in the former case, conventionalization should facilitate selection of the non-literal

interpretation, whereas in the latter it should act as a bias that must be overcome in order to select the literal meaning. Consequently, identifying the correct interpretation given a specific context requires accurate ToM reasoning.

We distinguish four possible outcome patterns:

- *Correct differentiation across contexts*: the model selects the non-literal interpretation in the non-literal context and the literal interpretation in the literal context. In the latter case, it successfully overcomes the conventionalization bias by relying on belief information, suggesting ToM-like reasoning.
- *Non-literal bias*: the model selects the non-literal interpretation in both contexts, indicating a conventionalization bias without proper inference of the characters’ beliefs (no ToM).
- *Literal bias*: the model selects the literal interpretation in both contexts, suggesting reliance on alternative shortcuts and a failure to incorporate belief information, excluding ToM-like reasoning.
- *Systematic reversal*: the model swaps literal and non-literal interpretations across contexts, indicating neither sensitivity to conventionalization nor appropriate use of belief information (no ToM).

Consistent with Hu et al. (2023), we also developed reduced context versions of each story to assess whether inference relies on shallow distributional cues within the linguistic context.

The final dataset comprises 160 stories for Indirect Speech Acts, including 80 with reduced context, and 96 stories for Verbal Ironies, 46 of which have a reduced context.

4 Experimental setup

4.1 Task

We assessed both humans and models using a multiple-choice question format to enable a direct comparison between the two groups. Each story was followed by a question regarding the intended meaning of the character’s utterance, phrased as: “What did ⟨name⟩ mean with his/her last utterance?”. Respondents selected from four options: the non-literal interpretation, the literal interpretation, and two semantically coherent but pragmatically inappropriate distractors.

Model	Parameters
GPT-2	775M (Radford et al., 2019)
T5 + Flan-L	780M (Chung et al., 2022)
T5-3B + Flan-XL	3B (Chung et al., 2022)
T5-11B + Flan-XXL	11B (Chung et al., 2022)
Tk-Instruct-3B	3B (Wang et al., 2022)
Tk-Instruct-11B	11B (Wang et al., 2022)
Falcon + Instruct	7B (Almazrouei et al., 2023)
Llama2 + Instruct	7B (Touvron et al., 2023)
Llama2-13B + Instruct	13B (Touvron et al., 2023)
Llama3 + Instruct	8B (AI@Meta, 2024)
GPT-3.5	175B (est.)
GPT-4	1.76T (est.)

Table 1: Models tested in our experiments.

4.2 Human data

We tested 360 participants using the Prolific platform. ISAs were categorized into eight groups, while Verbal Ironies were divided into four groups, with each group comprising 20-21 stories. The distribution of phenomena was randomized to ensure their equal representation and a balanced number of both indirect and direct interpretations across groups. Each group was tested with 30 speakers.

The participants were native English speakers, equally balanced in gender, from the USA and the UK. The test included three check questions, and participants were compensated only if they answered these questions correctly (Appendix A).

4.3 Models

This selection enabled us to systematically analyze how performance varies with model size, instruction tuning, and architectural family.

Each model was evaluated using the same prompt format (Appendix A), which included task instructions, the story with a question, four answer options, and the closing phrase “*The correct answer is:*”. To enhance robustness, each model received every item four times with randomized answer order. While varying the number of options does not necessarily improve accuracy, randomizing and repeating responses helps strengthen evaluation reliability (Li et al., 2024).

Moreover, we included a control task that did not require any ToM reasoning but followed the same format as the experimental tasks, in order to evaluate the models’ ability to adhere to the instructions provided in the prompt.³

³As a control prompt, we included the check questions originally used for human participants and extended them to a total of 20 items. See Appendix A for examples of these

4.4 Evaluation

We adopt a **zero-shot prompting** framework to evaluate ToM as a potentially emergent capability in language models. Recent studies identified a discrepancy between the probability models assign to word sequences, which is referred to as their “competence” since it directly relates to their language modeling training objective, and models’ “performance”, which is given by the accuracy of their responses elicited via prompting. This line of research suggests that prompting tends to underestimate the underlying linguistic knowledge of the models (Hu and Levy, 2023). Therefore, two complementary evaluation methods were used:

1. **Generation-based evaluation:** Models were prompted to produce a number (1 – 4) corresponding to the answer choice. Only valid numeric outputs were accepted, while all other outputs (e.g., text, missing numbers) were marked as invalid and considered task failures.

2. **Probability-based evaluation:**⁴ Using the `minicons` library (Misra, 2022), we extracted: a) *Number Probability* (likelihood of the correct digit), b) *Choice Probability* (likelihood of the full answer sentence), and c) *Continuation Probability* (the probability of the answer given the prompt without the final question; see Appendix A).

5 Experimental Results

5.1 Probabilities vs. Generation

Figures 3 and 4 compare model performance across three probability-based evaluation methods. Among them, Choice Probability yields the highest performance overall, followed by Continuation Probability, while Number Probability shows minimal variation across models, suggesting it behaves as a random effect.

When comparing **Generation accuracy** with the best probability measure, that is **Choice Probability** (henceforth referred to simply as probability), we find that for all models probability evaluation is generally better than Generation (except for Indirect Threats). This contrast may stem from smaller or non-instruction-tuned models (e.g., GPT-2, T5, T5 3B, Falcon, Llama 2 7B, Llama 2 13B, T5 11B) struggling to follow instructions during text generation. The control task performances in Appendix B support this result:

check questions and Appendix B for control experiments.

⁴Probability estimates were not available for OpenAI models due to API limitations.

Smaller models often fail to follow instructions and show low generation accuracy despite assigning high probability to the correct answer, indicating that the answer is encoded but not generated.

Conversely, certain models that achieve the highest scores in generation, such as Flan-T5 XXL, Llama 3 8B Instruct, Flan-T5 XL, and Llama 2 7B Instruct, exhibit lower accuracy when evaluated using probability. This pattern is further illustrated in Figure 5, which shows that probability scores tend to plateau or even decline beyond a certain model size. This suggests the presence of a threshold after which generation-based evaluation outperforms probability-based measures. This discrepancy underscores the need for caution when interpreting performance metrics and highlights the importance of comparing different evaluation methods to understand their influence on results.

Notably, these trends are consistent across all ISA categories and types of irony, indicating that the choice of evaluation method systematically affects performance assessment.

5.2 Accuracy Across Phenomena

Figure 6 reveals substantial variability in accuracy across phenomena within each model family. For ISAs, the Llama family tends to perform better on Indirect Threats than on other categories, while Tk-Instruct models show the opposite pattern, performing relatively worse on Indirect Threats but better on Indirect Suggestions. Across most models, Indirect Suggestions yield the highest accuracy among ISA phenomena, whereas Indirect Declinations and Indirect Requests tend to elicit weaker performance. For irony, Indirect Sarcasm consistently attracts the highest accuracy among instruction-tuned models (most notably Llama 2 7B Instruct, which reaches approximately 0.80) while Indirect Hyperbole yields the lowest scores across nearly all models. Indirect Rhetorical Questions show a more stable pattern, with most models clustering around 0.40–0.65. Overall, ISA accuracy tends to be higher and more variable across phenomena, while irony scores are generally lower and more uniform, suggesting that irony poses a more uniformly difficult challenge for current LLMs.

6 Discussion and Analysis

In this section, we carry out a more comprehensive analysis of the experimental results to address the

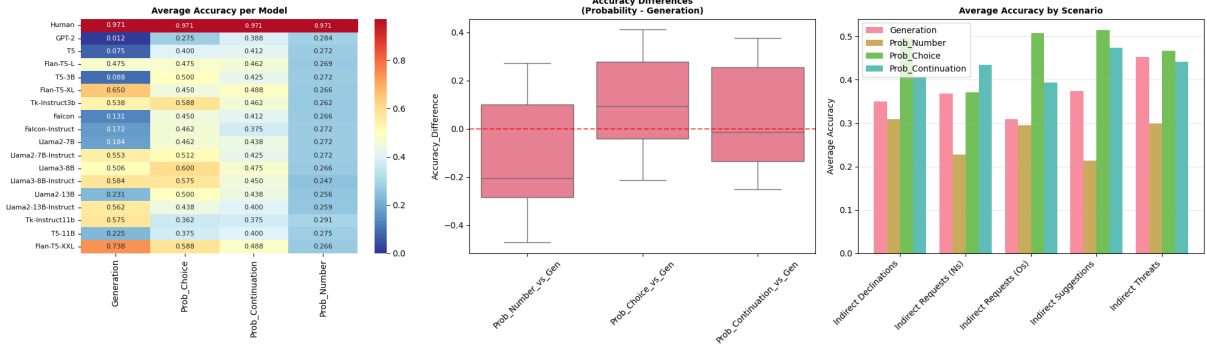


Figure 3: Comparison of evaluation methods for ISAs. (Left) Average accuracy per model across four methods; humans are near-ceiling (0.971), models perform lower with inconsistent patterns. (Center) Accuracy differences (probability-based minus generation) per model; Prob Choice shows the largest positive shift, tending to overestimate accuracy. (Right) Average accuracy by scenario; Prob Choice consistently scores highest, especially for indirect suggestions and affirmative requests, showing that method choice substantially affects conclusions about models' pragmatic competence.

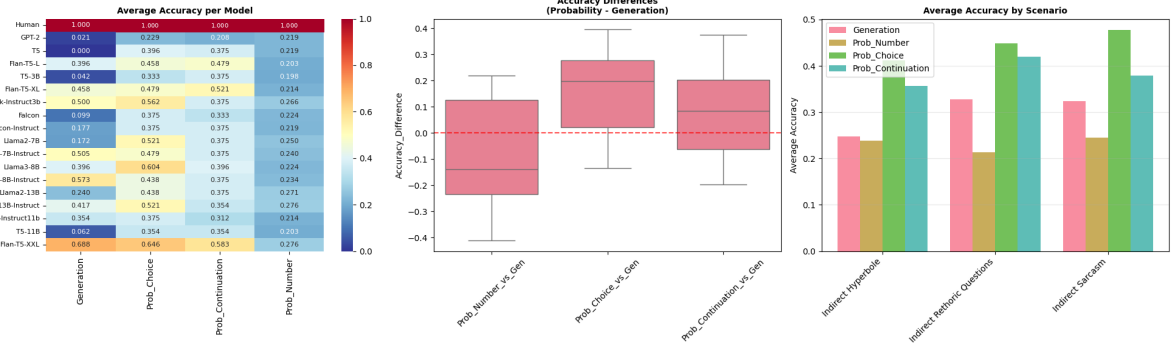


Figure 4: Comparison of evaluation methods for Irony. (Left) Average accuracy per model across four methods; humans reach ceiling (1.000), models perform lower with inconsistent patterns. (Center) Accuracy differences (probability-based minus generation) per model; Prob Choice shows the largest positive shift, tending to overestimate accuracy. (Right) Average accuracy by scenario; Prob Choice consistently scores highest, especially for sarcasm, showing that method choice substantially affects conclusions about models' pragmatic competence.

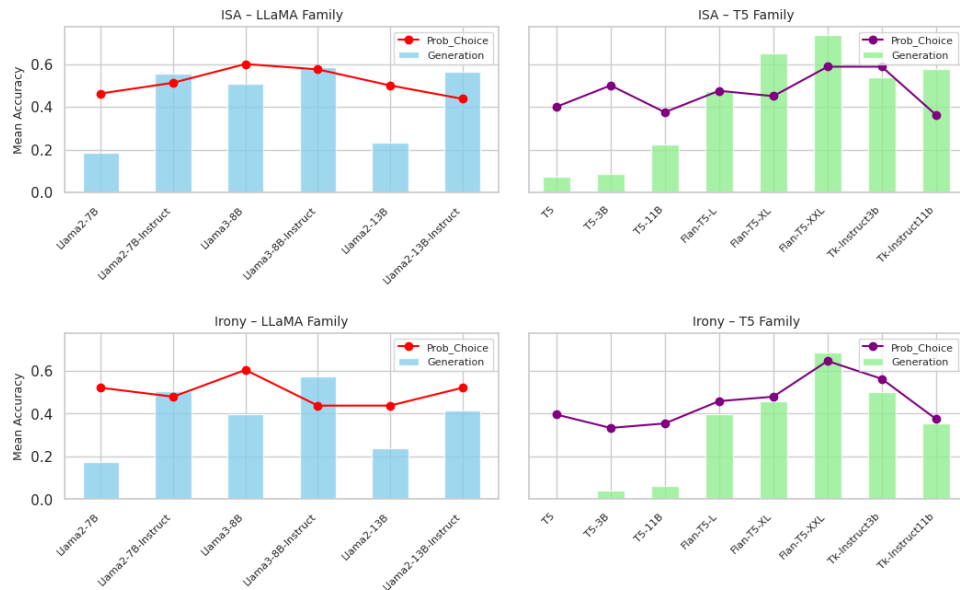


Figure 5: Comparison of Generation and Probability accuracy between the two main model families evaluated: T5-based models and LLaMA-based models.

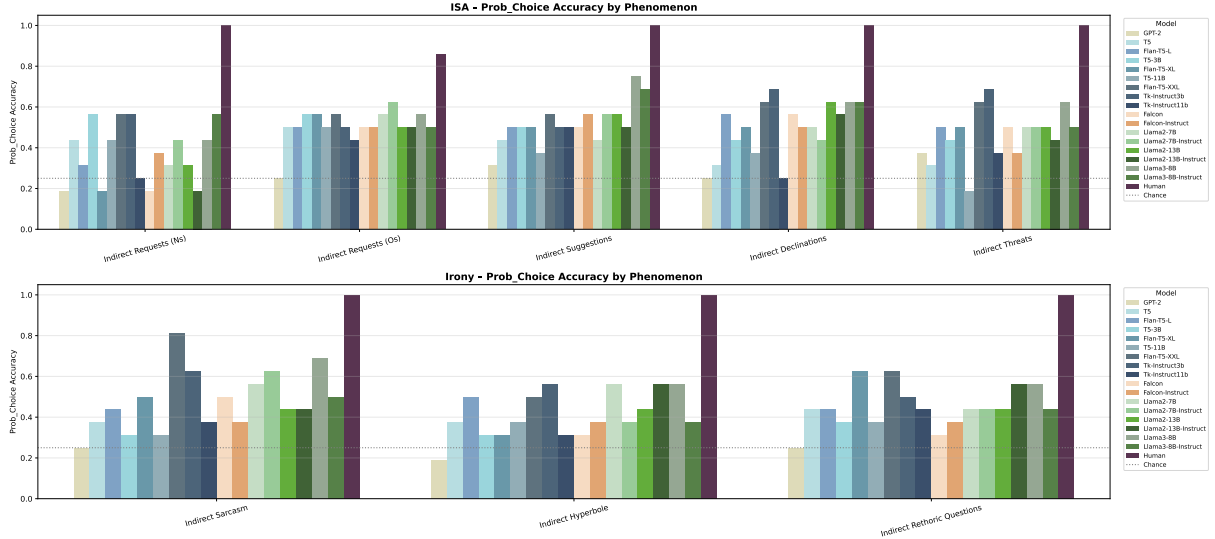


Figure 6: Probability-based accuracy across phenomena and models, compared against the human baseline

research questions outlined in Section 1.

6.1 LLMs and Humans in Mentalizing task

In our experiments, we controlled for both literal and non-literal interpretations. Crucially, distinguishing between them necessitates considering ToM, as the relevant contexts differ based on the mutual beliefs held by the speaker and listener in the stories.

Can LLMs perform tasks that involve understanding mutual beliefs between characters? While LLMs do not achieve human-level accuracy in these tasks, certain models exhibit relatively high accuracy, particularly those that have undergone post-training on specialized tasks. Still, **the gap between them and humans is still very large across all phenomena** (cf. Figures 3, 4, and 6). Therefore, if one were to frame the question in a yes/no form, the response would be negative. The assessment of LLMs performance in mentalizing tasks should focus on the errors made rather than the number of correct answers. If LLMs genuinely possessed the ability to mentalize and apply such cognitive skills in task completion, their accuracy would likely approach that of humans. In humans, responses are consistently guided by their capacity for mentalizing, as this is needed to understand the knowledge contexts in the stories.

It is worth noting that conventionalization is not, in itself, evidence against ToM. Humans also default to direct, conventionalized readings when available, and doing so is not typically considered a failure of mentalizing. The relevant question is

therefore not whether a system relies on conventionalized cues per se, but whether its behavior remains sensitive to belief-context manipulation independently of, and when in conflict with, conventionalization strength.

Since mental states are not directly observable, there is no absolute gold standard about *genuine* mentalizing in humans. Moreover, talking about mentalizing tasks in LLMs does not entail assuming that such models might have anything like mental states. What is at stake here is the humans’s and models’ abilities to perform *belief-tracking*, and to use it for language interpretation. Following the logic of the false-belief paradigm itself, we treat belief-tracking as demonstrated when behavior differentiates correctly across matched contexts that hold conventionalized cues constant. Given this operational criterion, humans in our data show such differentiation (i.e., near-ceiling accuracy in both the literal and non-literal conditions), while models largely do not, reverting to literal readings once any context is present regardless of its belief content. Model evaluation here is thus necessarily comparative and human-referenced.

In this regard, our findings are consistent with numerous studies in the scientific literature (Ullman, 2023), which underscore the limitations of LLMs in executing tasks that necessitate mentalizing. However, if true ToM-like capabilities are still lacking in LLMs, what mechanisms or factors account for the particularly accurate task completion observed in several cases?

Step	ΔR^2	F	p -value
ISA			
1 $\rightarrow$ 2	0.298	76.6	2.2e-15
2 $\rightarrow$ 3	0.000031	0.0079	9.2e-01
3 $\rightarrow$ 4	0.007	1.76	1.8e-01
Irony			
1 $\rightarrow$ 2	0.247	34	6.9e-08
2 $\rightarrow$ 3	0.001	0.19	6.6e-01
3 $\rightarrow$ 4	0.010	1.42	2.3e-01

Table 2: Model comparisons: R^2 gain and F-test statistics

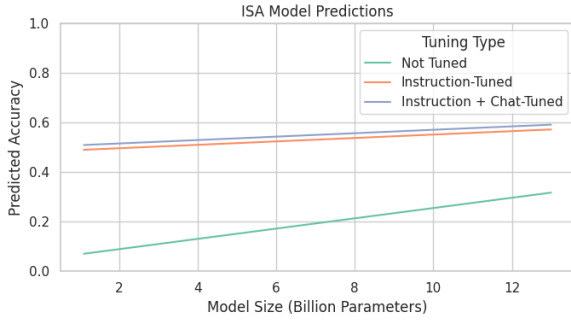


Figure 7: Results of a four-step regression analysis examining the impact of model size, instruction tuning, and chat tuning on accuracy for ISA.

6.1.1 Statistical Analyses of the Models' Accuracy

To answer this question, we conducted a series of additional statistical analyses to explore the respective contributions of size, instruction tuning, and chat tuning to model accuracy. As shown in Figure 5, instruction-tuned models consistently outperform their base versions across both the T5 and LLaMA model families, suggesting that the ability to follow instructions plays a critical role in task performance.

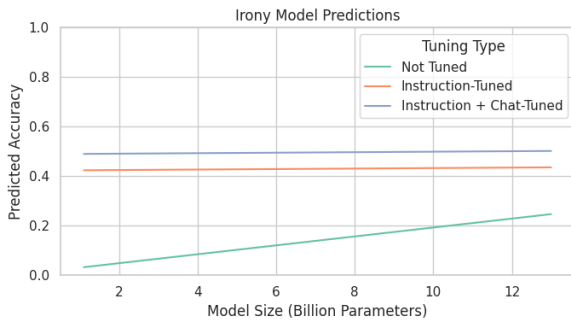


Figure 8: Results of a four-step regression analysis examining the impact of model size, instruction tuning, and chat tuning on accuracy for Irony.

To better quantify these effects, we conducted a four-step regression analysis with increasing factor complexity. In the first step, we examined whether model size alone could account for variation in probability-based accuracy. The second step introduced a binary variable for instruction tuning, allowing us to measure its additional impact beyond size. In the third step, we added a chat tuning variable to evaluate whether conversational fine-tuning provided further improvements. Finally, in the fourth step, we included an interaction term between size and instruction tuning to explore whether the effect of model size depends on the presence of instruction tuning. This step-wise approach allowed us to isolate the relative influence of each factor and examine how they interact in shaping model accuracy.

Figures 7 and 8 and Table 2 provide further support for the central role of instruction tuning in determining model performance. Compared to model size, instruction tuning shows a significantly stronger effect on accuracy across both ISA and irony tasks. Additionally, chat tuning and interaction effects do not provide statistically meaningful gains in either task (see F-tests), suggesting that the main performance gains come from the initial instruction tuning step rather than further refinement through chat-specific tuning.

Such results reinforce the conclusion that instruction tuning is a key factor influencing model effectiveness, surpassing the impact of scale alone. This also calls for caution in treating the alleged pragmatic abilities of LMs as emergent properties of their predictive training objective (Kosinski, 2024). Rather, a key factor in shaping such abilities lies in the highly sophisticated supervised post-training processes that instruction-tuned models undergo.

The divergence between LLaMA and T5/Tk families on specific phenomena (e.g., LLaMA outperforming on Indirect Threats, Tk-Instruct on Indirect Suggestions) is not random but likely reflects systematic differences in fine-tuning data composition. Crucially, this family-level patterning is itself evidence against ToM reasoning: if models were truly mentalizing, performance should be more uniformly sensitive to belief context rather than tracking the statistical profile of specific conventionalized expressions in training data.

While the results thus far partially address our

Model	ISAs		Irony	
	Literal	Non-Literal	Literal	Non-Literal
Falcon	0.208	0.542	0.425	0.475
Falcon-Instruct	0.333	0.417	0.526	0.400
Flan-T5-L	0.500	0.417	0.450	0.500
Flan-T5-XL	0.542	0.417	0.475	0.425
Flan-T5-XXL	0.708	0.583	0.550	0.625
GPT-2	0.125	0.333	0.375	0.175
Human	1.000	0.947	1.000	1.000
Llama2-13B	0.542	0.333	0.675	0.325
Llama2-13B-Instruct	0.667	0.375	0.525	0.325
Llama2-7B	0.583	0.458	0.500	0.425
Llama2-7B-Instruct	0.583	0.375	0.500	0.525
Llama3-8B	0.708	0.500	0.675	0.525
Llama3-8B-Instruct	0.458	0.417	0.700	0.450
T5	0.333	0.458	0.600	0.200
T5-11B	0.208	0.500	0.550	0.200
T5-3B	0.208	0.458	0.675	0.325
Tk-Instruct11b	0.333	0.417	0.475	0.250
Tk-Instruct3b	0.792	0.333	0.875	0.300

Table 3: Accuracy by interpretation type (Literal vs. Non-Literal) across models for ISAs and Irony in ToM context. Bold values indicate the dominant interpretation (literal bias when Literal > Non-Literal, or non-literal bias otherwise).

first research question (the extent to which LLMs perform tasks requiring mentalizing) and indicate that models do not match human-level performance overall, some models indeed achieve high accuracy. This improvement has been partly attributed to the role of instruction tuning. However, accuracy alone does not address one of the central aims of our analysis: Determining whether the degree of conventionalization influences models’ successful completion of ToM tasks.

6.2 Do Belief Contexts Really Drive LLMs’ Interpretation?

The key design feature of our test items is the difference in the belief contexts of the two characters triggering either the literal or the non-literal interpretation of the target sentences. We refer to these as **belief-biased contexts**. A key question, then, is: **Does a genuine causal relationship exist between such biased contextual belief states and the answers produced by LLMs?**

This experiment was motivated by the hypothesis that conventionalization, rather than belief attribution, drives model performance on ToM tasks. To test this, we examine the contrast between literal interpretations, which are non-conventionalized in our test items and dependent on belief reasoning, and non-literal interpretations, which are instead strongly conventionalized.

Table 3 reports probability-based accuracy for literal and non-literal interpretations. A clear pattern emerges for ISAs: models within the same family tend to behave consistently, with the Llama

series and instruction-tuned T5 models achieving higher accuracy on literal stories, while non-instruction-tuned T5 and Falcon models show a preference for non-literal interpretations. For irony, the pattern is more uniform: almost all models exhibit a literal bias, or show negligible differences between literal and non-literal accuracy.

One interpretation for the models dissociation displayed in ISAs might be that instruction tuning, by sharpening the model’s sensitivity to contextual cues, pushes models toward literal readings, because the belief-biased context provides explicit, surface-level information that instruction-tuned models are particularly good at exploiting. This would suggest that instruction tuning improves task compliance but not pragmatic reasoning, and may actually suppress the non-literal associations acquired during pretraining.

However, our findings suggest that reducing contextual information does not result in different interpretations by the model (full context vs. short context, Appendix D). To further investigate whether the models’ behavior is indeed driven by sensitivity to differences in the belief contexts and whether the default interpretation of the conventionalized utterance in isolation would be non-literal, we conducted additional experiments designed to investigate whether the presence or absence of belief-related context in stories influences a model’s ability to select the correct interpretation.

We designed two new test sets: i.) **ToM-N**, in which the target utterances are preceded by a *neutral context without belief states triggering either a literal or non-literal reading* (see Appendix D), and ii.) **ToM-0**, in which the target utterances appear *without contextual information*. The original stimuli with belief-biased context requiring ToM reasoning are henceforth designated as **ToM**.

To examine variation in interpretations across context types, we employed a soft clustering method based on Expectation Maximization (EM) with a Gaussian Mixture Model. The analysis was constrained to two clusters, labeled **Literal** and **Non-Literal**, which were determined from the distribution of the preference score (Δ) for the non-literal answer option in the ToM literal and ToM non-literal conditions (corresponding to the two possible interpretations). Δ is defined as $\alpha - \beta$ where α denotes the model’s probability of a non-literal interpretation and β the probabil-

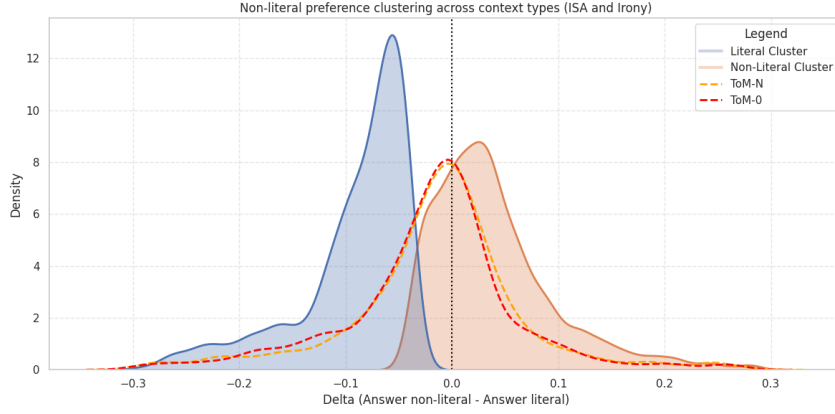


Figure 9: Non literal probability clustering across context types for ISA and Irony

ity of a literal interpretation. The EM algorithm estimates these distributions and assigns posterior probabilities to new data points, allowing us to assess how interpretations from ToM-N and ToM-0 align with either the Literal or Non-Literal cluster. This approach provides a probabilistic estimate of whether a given target sentence in the belief-neutral or zero-context conditions is interpreted more like the non-literal or literal cases in the belief-biased ToM condition.

Figure 9 shows that both ToM-0 and ToM-N contexts align more closely with the non-literal cluster, indicating a model preference for non-literal interpretations in the absence of explicit belief-biased contextual cues, a pattern further reflected in the Mean Δ values reported in Table 4.

Condition	Mean Δ
ToM	-0.0122
ToM-N	-0.0075
ToM-0	0.0047

Table 4: Mean Δ accuracy across conditions

Positive values indicate a bias toward non-literal interpretations, while negative values indicate a bias toward literal interpretations.

As shown in Table 4, ToM and ToM-N mean Δ values are slightly negative, indicating a preference for literal interpretations when context is present. The ToM-0 condition, by contrast, yields a positive value, suggesting that in the absence of context, models tend toward non-literal interpretations, consistently with its strong conventionalization in our stimuli.

Moreover, a one-way ANOVA conducted on the models’ probability-based accuracy reveals

ANOVA	SS	df	F	p
Label	0.52	3	24.94	4.30×10^{-16}
Residual	120.55	17404	—	—
Tukey HSD (FWER = 0.05)				
G1	G2	ΔM	p-adj	Sig.
ToM-L	ToM-NL	+0.005	.054	No
ToM-N	ToM-L	-0.005	.017	Yes
ToM-N	ToM-NL	+0.010	< .001	Yes
ToM-0	ToM-L	+0.010	< .001	Yes
ToM-0	ToM-N	+0.005	.041	Yes
ToM-0	ToM-NL	+0.015	< .001	Yes

Table 5: ANOVA and Tukey HSD. L = Literal, NL = Non Literal.

a significant effect across the three context (belief-biased, belief-neutral, and zero) conditions ($F(3, 17404) = 24.94$, $p < .001$). Post-hoc Tukey HSD comparisons (Table 5) indicated that ToM-0 differed significantly from all other conditions. ToM-L(iteral) differed significantly from ToM-0 and ToM-N, but not from ToM-N(on)L(iteral) ($p = .054$), suggesting that the literal and non-literal belief-biased conditions elicit similar model behavior. ToM-N and ToM-NL also differed significantly from each other. These results suggest that models do encode conventionalized non-literal meanings, as evidenced by the non-literal bias observed in ToM-0, but that this default is overridden as soon as any context is introduced. Notably, this shift occurs even under belief-neutral conditions (ToM-N, Mean $\Delta = -0.0075$), indicating that it is not belief-reasoning per se, but **the mere presence of contextual information, that drives models toward literal interpretations.**

This might be interpreted as a dissociation between **stored pragmatic knowledge** and **applied pragmatic reasoning**: Models have learned that

“Can you open the window?” is typically a request, but when a story context is introduced, they leverage its surface content rather than using it to enrich interpretation.

This behavior contrasts with human processing as described by the **conventionalization hypothesis** (Gibbs, 1983, 2002), according to which highly conventionalized non-literal meanings are accessed directly and preferentially, without requiring deliberate pragmatic inference. In skilled human comprehenders, this direct access route makes ToM largely unnecessary for interpreting established ISAs. The present results suggest an inverted pattern in LLMs: whereas humans default to non-literal readings for conventionalized expressions regardless of context, models exhibit this default only in the complete absence of context, and revert to literal readings once any contextual information is available. This inversion suggests that LLMs not only lack robust ToM-based pragmatic reasoning, but also fail to replicate the processing route that renders such reasoning unnecessary in humans.

7 Conclusions

Previous studies have reported impressive performances of LLMs on psycholinguistic tasks designed to require ToM. This paper critically assesses these claims by examining whether LLMs can infer the intended meaning in two pragmatic tasks: ISAs and irony. We used short narratives in which the correct interpretation, literal or non-literal, depended on reasoning about the differing mental states of two characters, allowing us to test whether models really adjust their interpretations based on belief contexts. We evaluated nineteen models, including instruction-tuned variants, across four contextual conditions.

Our results reveal several key findings:

1. LLMs consistently underperform compared to humans in the pragmatic tasks tested in the experiments.
2. LLMs struggle to correctly interpret non-literal meanings, even when such interpretations are more conventionalized than their literal counterparts, showing a divergent processing pattern from humans.
3. Model interpretation of target utterances is not driven by the belief context, suggesting

that even correct responses may not reflect human-like ToM abilities. Models are indeed affected by the presence of context, but crucially not by differences in the belief states expressed in such contexts.

4. Generic contextual information (e.g., descriptive or neutral situational cues) influences interpretation, suggesting that the model behavior is driven by surface statistical clues in the context, rather than by the actual belief states of the characters in the story.

Our analyses further reveal that while instruction tuning and increased model size improve performance on tasks requiring reasoning about characters and beliefs, these improvements alone are not sufficient to simulate ToM-like capabilities.

Overall, our findings underscore a critical distinction between human and model behavior in interpreting intended meanings. Unlike humans, who often default to non-literal interpretations in stereotypical contexts, models tend to favor literal interpretations ignoring different belief-contexts. This suggests that LLMs may rely on shallow pattern-matching rather than true inferential reasoning. As such, attributing ToM abilities to model outputs without considering alternative explanations is problematic. Extreme caution is therefore warranted when interpreting the behavior of LLMs as evidence of ToM-like reasoning. The observed performance may instead reflect systematic biases in the models’ interpretive strategies, shaped by pretraining and fine-tuning data distributions rather than belief-based reasoning.

References

- Brian P. Ackerman. 1983. [Form and function in children’s understanding of ironic utterances](#). *Journal of Experimental Child Psychology*, 35(3):487–508.
- Marc Aguert, Coralie Le Vallois, Karine Martel, and Virginie Laval. 2018. [“that’s really clever!” ironic hyperbole understanding in children](#). *Journal of Child Language*, 45(1):260–272.
- AI@Meta. 2024. [Llama 3 model card](#).
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra

- Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malarctic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. [The falcon series of open language models](#).
- Jaan Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. 2023. [Mind the gap: Challenges of deep learning approaches to theory of mind](#). *Artificial Intelligence Review*, 56(9):9141–9156.
- John L. Austin. 1975. *How to Do Things With Words*. Harvard University Press.
- Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. [Does the autistic child have a “theory of mind”?](#) *Cognition*, 21(1):37–46.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrike, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#).
- Debra L. Burnett. 2015. [Exploring the role of conventionality in children’s interpretation of ironic remarks](#). *Journal of Child Language*, 42(6):1267–1288.
- Robyn Carston and Catherine Wearing. 2015. [Hyperbolic language and its relation to metaphor and irony](#). *Journal of Pragmatics*, 79:79–92.
- Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. [Theory of mind in large language models: Assessment and enhancement](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558, Vienna, Austria. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, 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](#).
- Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. [A hard nut to crack: Idiom detection with conversational large language models](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Todd Firsich and Anthony Rios. 2024. [Can GPT4 detect euphemisms across multiple languages?](#) In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 65–72, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- José M. Gavilán and José E. García-Albea. 2011. [Theory of mind and language comprehension in schizophrenia: Poor mindreading affects figurative language comprehension beyond intelligence deficits](#). *Journal of Neurolinguistics*, 24(1):54–69.
- Ine Gevers, Victor De Marez, Luna De Bruyne, and Walter Daelemans. 2025. [WinoWhat: A parallel corpus of paraphrased WinoGrande sentences with common sense categorization](#). In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 68–80, Vienna, Austria. Association for Computational Linguistics.
- Raymond W. Gibbs. 1983. [Do people always process the literal meanings of indirect requests?](#) *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(3):524–533.
- Raymond W. Gibbs. 1986a. [On the psycholinguistics of sarcasm](#). *Journal of Experimental Psychology: General*, 115(1):3–15.
- Raymond W. Gibbs. 1986b. [What makes some indirect speech acts conventional?](#) *Journal of Memory and Language*, 25(2):181–196.
- Raymond W. Gibbs. 2002. [A new look at literal meaning in understanding what is said and implicated](#). *Journal of Pragmatics*, 34(4):457–486.

- Rachel Giora and Ofer Fein. 1999a. [Irony comprehension: The graded salience hypothesis](#). *Humor*, 12(4):425–436.
- Rachel Giora and Ofer Fein. 1999b. [Irony: Context and salience](#). *Metaphor and Symbol*, 14(4):241–257.
- Herbert P. Grice. 1975. Logic and conversation. *Syntax and Semantics: Speech Acts*, 3(1):41–58.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. [A fine-grained comparison of pragmatic language understanding in humans and language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.
- Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of EMNLP*.
- Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions of the Royal Society B*.
- Hyewon Jang, Moritz Jakob, and Diego Frassinelli. 2024. [Context vs. human disagreement in sarcasm detection](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 1–7, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Cameron Robert Jones, Sean Trott, and Ben Bergen. 2023. [Epitome: Experimental protocol inventory for theory of mind evaluation](#). In *First Workshop on Theory of Mind in Communicating Agents*.
- Laura Alba Juez. 1998. [On the conventionalization of verbal irony in english](#). *Perspectivas Pragmáticas en Lingüística Aplicada*, pages 11–18.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. [Fantom: A benchmark for stress-testing machine theory of mind in interactions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.
- Michał Kosinski. 2024. [Evaluating large language models in theory of mind tasks](#). *Proceedings of the National Academy of Sciences*, 121(45).
- 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.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR’12*, page 552–561. AAAI Press.
- Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.
- Huaoli Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. [Theory of mind for multi-agent collaboration via large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.
- Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. [Can multiple-choice questions really be useful in detecting the abilities of llms?](#)
- Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. [Towards a holistic landscape of situated theory of mind in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. [Dissociating language and thought in large language models](#). *Trends in Cognitive Sciences*, 28(6):517–540.
- Antonella Marchetti, Cinzia Di Dio, Angelo Cangelosi, Federico Manzi, and Davide Massaro. 2023. [Developing chatgpt’s theory of mind](#). *Frontiers in Robotics and AI*, 10.
- Eleonora Marocchini and Filippo Domaneschi. 2022. [“can you read my mind?” conventionalized indirect requests and theory of mind abilities](#). *Journal of Pragmatics*, 193:201–221.
- Sergio Melogno et al. 2021. [Case report: Theory of mind and figurative language in a child with agenesis of the corpus callosum](#). *Frontiers in Psychology*, 11:596804.
- Kanishka Misra. 2022. [minicons: Enabling flexible behavioral and representational analyses of transformer language models](#).
- 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.
- Brendan Park, Madeline Janecek, Naser Ezzati-Jivan, Yifeng Li, and Ali Emami. 2024. [Picturing ambiguity: A visual twist on the Winograd schema challenge](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–374, Bangkok, Thailand. Association for Computational Linguistics.
- Steven Pinker, Martin A. Nowak, and James J. Lee. 2008. [The logic of indirect speech](#). *Proceedings of the National Academy of Sciences*, 105(3):833–838.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- 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.
- John R. Searle. 1975. Indirect speech acts. *P. Cole and J. Morgan (Eds.), Syntax and Semantics 3: Speech Acts*, pages 59–82.
- Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. [Clever hans or neural theory of mind? stress testing social reasoning in large language models](#).
- Stephen Skalicky. 2024. *Verbal Irony Processing*. Cambridge University Press.
- Dan Sperber and Deirdre Wilson. 2002. [Pragmatics, modularity and mind-reading](#). *Mind and Language*, 17(1-2):3–23.
- James W. A. Strachan, Davide Albergo, Gabriele Borghini, et al. 2024. [Testing theory of mind in large language models and humans](#). *Nature Human Behaviour*, 8:1285–1295.
- Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Agüera y Arcas, and Robin I. M. Dunbar. 2026. [Llms achieve adult human performance on higher-order theory of mind tasks](#). *Frontiers in Human Neuroscience*, 19:1633272.
- Jing Han Sun and Ali Emami. 2024. [EvoGrad: A dynamic take on the Winograd schema challenge with human adversaries](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6701–6716, Torino, Italia. ELRA and ICCL.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- Sean Trott and Benjamin Bergen. 2019. [Individual differences in mentalizing capacity predict indirect request comprehension](#). *Discourse Processes*, 56(8):675–707.
- Tomer Ullman. 2023. [Large language models fail on trivial alterations to theory-of-mind tasks](#).
- Jill de Villiers. 2007. [The interface of language and theory of mind](#). *Lingua. International Review of General Linguistics*, 117(11):1858–1878.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. [Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks](#).
- Lena Wimmer, Ursula Christmann, and Elisabeth Ihmels. 2016. [Non-conventional figurative language as aesthetics of everyday communication](#). *Metaphor and the Social World*, 6(2):243–275.
- Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023. [How far are large language models from agents with theory-of-mind?](#)

A Stimuli and Prompts

In Table 6 we present stimuli for each phenomena in our dataset. Stimuli were presented both to model and humans with slight changes. We tested humans through the Prolific platform with the following premise:

In this test, you will read short stories and you will have to answer a question about one character of the story. READ EACH STORY CAREFULLY! You must chose one among 4 possible answers. Choose the answer that you think to be correct given the story. The test contains thee check questions. IMPORTANT. The participant will not be paid unless it answers to these questions correctly.

The check questions included in stimuli submitted to humans were the following:

Today is Sunday and after school Tommaso goes out to play with Red in the garden. When he hears Lisa call "Lunch is ready, time to stop playing now and come in". He runs in and sits at the table.

What does Lisa inted to say?

1 The lunch is ready; 2 The lunch isn't ready; 3 I want to go out fot lunch; 4 It's time for playing.

"Amy!" Mrs. Grove called from the door that opened towards the garden. But no answer came.

What does Mrs. Grove intend to say?

1 She is asking a question; 2 She is calling Amy; 3 She is answering a question; 4 She wants to change her name.

My boyfriend David, is a dentist. David says that he is happy to work in the centre of a small city. He can go shopping in his lunch hour, or go for a walk and look at the ducks on the river. He doesn't need a car, unless he wants to drive out of York into the countryside.

What does David intens to say?

1 He is happy to work in the centre of a small city; 2 He doesn't like working in a small city; 3 He wants to move abroa; 4 He hates the small cities.

The sequence of stories within the form, as well as the options presented in each story, were randomized. This randomization minimizes subjective error in human psycholinguistic text construc-

tion, as the entire process of form creation and stimuli randomization was automated. To generate forms from the table, we employed App Scripts. The Excel files containing human responses, along with the order of stimuli and samples, are stored on GitHub.

The query used to prompt the models, instead, was the following (example extraxted from indirect declinations):

"Task: I'll give you a story and I'll ask you to answer a question about one character of the story. I'll give you four possible answers to the question and you must choose the right one. The possible answers are 1, 2, 3 and 4. Story: I invited my brother Kevin to dinner to celebrate my birthday. Kevin and I rarely get to spend time together. For the occasion I prepared a fish dinner, and I bought some excellent Italian wine, our favorite. After a couple of drinks, I ask Kevin to pass me the glass so I can pour him more wine, but he tells me: "I need to drive". What does Kevin intend to say? Options: 1 Please, do not pour me more wine. 2 I drive tonight 3 I don't like this wine. 4 We should drink French wine. The correct answer is:"

The query used to collect continuation probabilities is:

"Task: I'll give you a story and I'll ask you to answer a question about one character of the story. I'll give you four possible answers to the question and you must choose the right one. The possible answers are 1, 2, 3 and 4. Story: I invited my brother Kevin to dinner to celebrate my birthday. Kevin and I rarely get to spend time together. For the occasion I prepared a fish dinner, and I bought some excellent Italian wine, our favorite. After a couple of drinks, I ask Kevin why he stopped drinking. Kevin replies "I need to drive". What I understand is... Options: 1 that Kevin is asking me to do not pour him more wine. 2 that Kevin has to drive tonight 3 that Kevin doesn't like this wine. 4 that we should drink French wine . The correct answer is:"

B Control Task

Figure 10 compares the accuracy of three response measures: free generation, full sentence probability (choice prob), and the probability assigned to the answer token (number prob). The results illustrate how generation-based accuracy

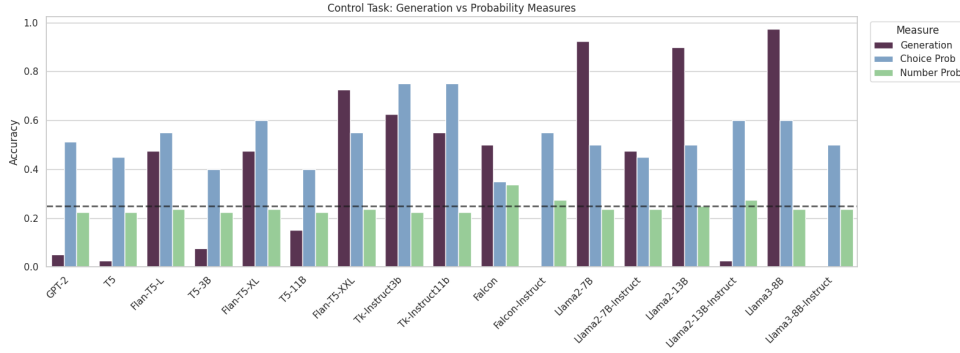


Figure 10: Accuracy of generation and probability-based measures on the control task

can be unreliable, particularly for smaller and non instruction-tuned models that frequently fail to produce well-formed option selections. In contrast, probability-based measures remain above chance across all models, suggesting that the underlying task is within the capability of all models evaluated. Given this, main analyses rely on probability-based measures.

C Generation Results

We report generation accuracy as it is the only metric consistently available across all models, due to the lack of probability outputs for GPT-3.5 and GPT-4.

C.1 Indirect Speech Acts

Figure 11 presents the accuracy scores for each type of Indirect Speech Act (ISA) across all models evaluated with the Generation method.

Out of the 19 models assessed, 11 outperform the baseline accuracy of 25%, indicating that their predictions are not the result of random selection.

The best-performing models across all ISA categories include GPT-3.5 and GPT-4, Flan-T5-L, Flan-T5-XL, Flan-T5-XXL, Tk-Instruct 3B and 11B (for the T5 architecture), as well as LLaMA 2 7B Instruct, LLaMA 2 13B Instruct, LLaMA 3 8B, and LLaMA 3 8B Instruct. Notably, nearly all of these models are instruction-tuned, with the exception of LLaMA 3 8B and for each pair, the instruction-tuned version consistently outperforms its non-instruction tuned counterpart. These results suggest that instruction tuning has a greater effect on ISA performance than model size alone.

Performance varies considerably across different ISA types. For example, LLaMA 2 7B Instruct achieves the highest accuracy on Indirect Threats (0.8), but its performance drops significantly on

Indirect Declinations and Indirect Requests (0.5). This pattern of variability across ISA types is observed across all models tested.

Importantly, even the best performing models fall short of human accuracy in all tasks. The ISA category with the lowest model accuracy is Indirect Requests (Os), in which human performance is also comparatively lower, though still consistently higher than that of the models.

Finally, tasks presented with reduced context show increased variability in human performance compared to models (see Appendix D, Figure 13).

C.2 Irony

Figure 12 indicates that LLMs exhibit similar behavior when processing irony as they do with ISAs. In particular, instruction tuning proves more impactful than model size. Nevertheless, only 9 out of the 19 models (Flan-T5-L, Flan-T5-XL, Flan-T5-XXL, Tk-Instruct 3B, LLaMA 2 7B Instruct, LLaMA 2 13B Instruct, LLaMA 3 8B Instruct, GPT-3.5, and GPT-4) outperform the baseline. Like with ISAs, model performance remains substantially below that of humans.

When context is reduced there is a decline in accuracy. However, the relative performance differences among models persist in these reduced-context settings (see Appendix D, Figure 14).

D Reduced Contexts

Figures 13 and 14 show generation accuracy for tasks with reduced contexts.

An example of ToM-N is:

Mark and his friend Kyle just returned home after a month-long trip to Europe. During their absence, they asked a friend to water the plants in their apartment and to open the windows once a week for air. As soon as they enter, Kyle asks,

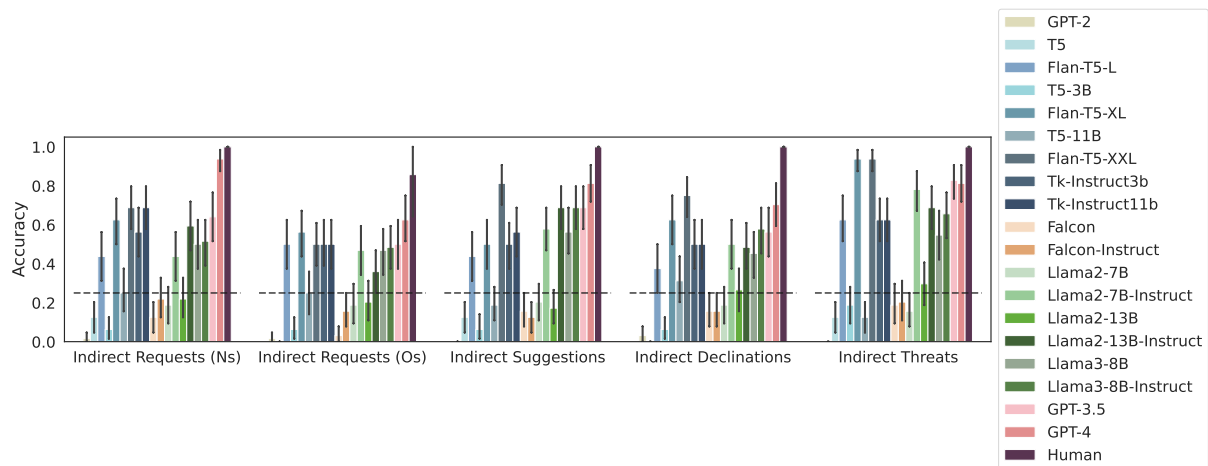


Figure 11: ISAs - Accuracy for tasks with non-reduced context.

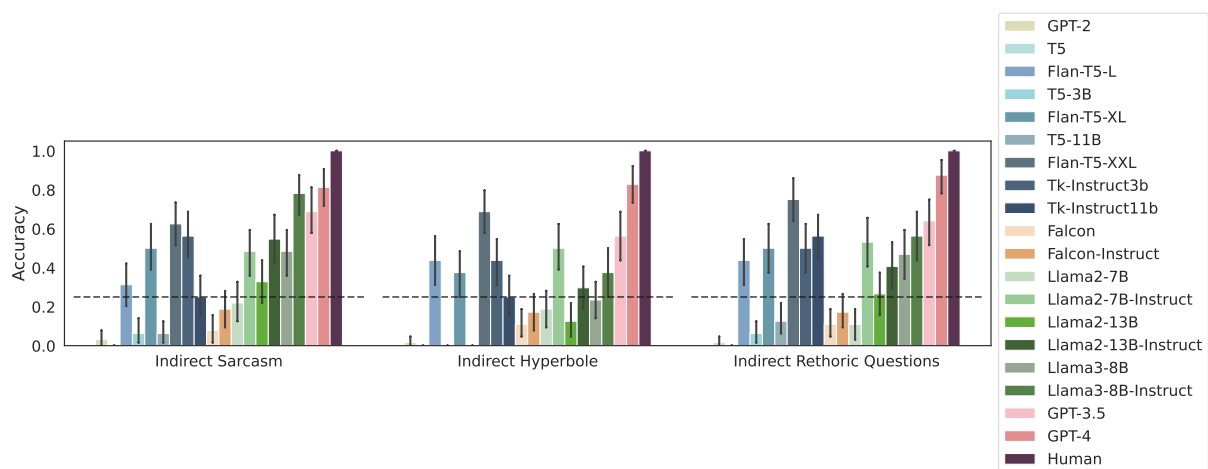


Figure 12: Irony - Accuracy for tasks with non-reduced context.

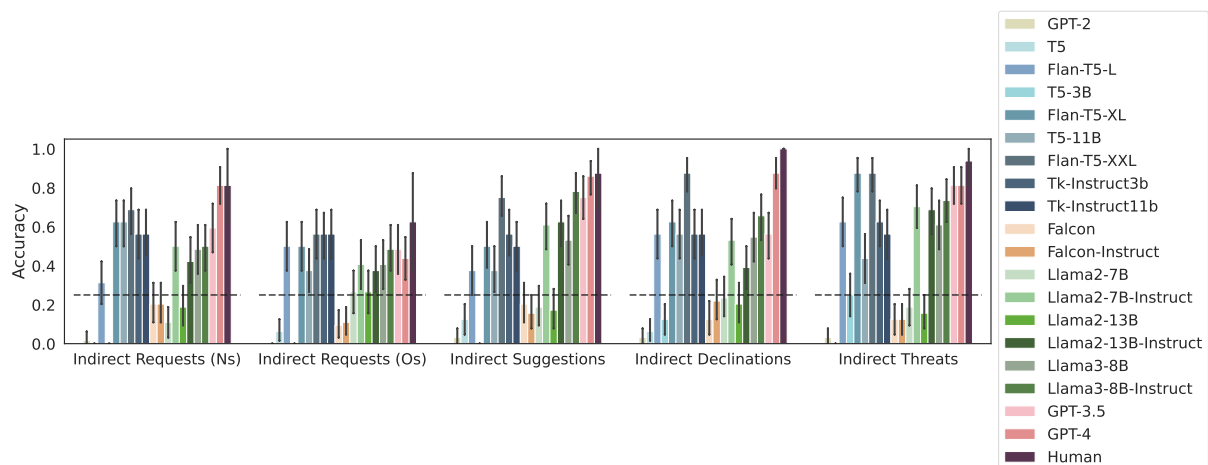


Figure 13: ISAs - Accuracy for tasks with reduced contexts.

"Can you open the window?".

An example of ToM-0, instead, is:

Kyle asks, "Can you open the window?".

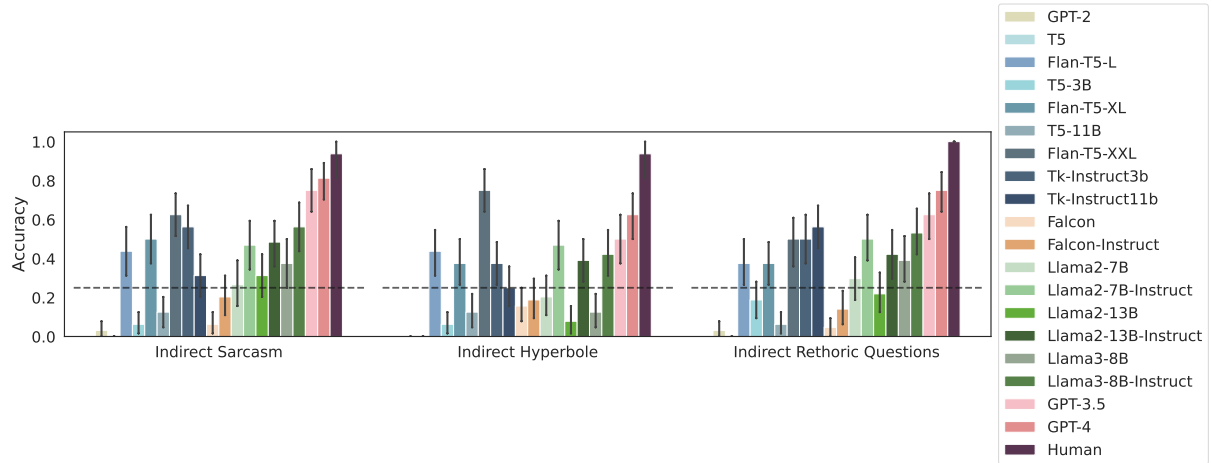


Figure 14: Irony - Accuracy for tasks with reduced contexts.

Phenomenon	Literal Story	Non-Literal Story	Question	Options
Indirect Request (Os)	You and your friend Jonathan are taking a road trip. You began in California, and are now passing through Michigan. It's almost winter, so it's very cold outside-especially for Southern California dwellers like you and Jonathan. You see that you're almost out of gas, so you stop at a gas station in a small town. You fill up the tank, and then the two of you go inside the gas station to buy some water and snacks. When you return to the car and start up the engine, you and Jonathan notice with some dismay a blinking light, which indicates that the car's heating system is broken. You both bundle up. As you leave the station, Jonathan shivers in his seat. He turns to you and says, "Man, it's really cold in here"	You and your friend Jonathan are taking a road trip. You began in California, and are now passing through Michigan. It's almost winter, so it's very cold outside-especially for Southern California dwellers like you and Jonathan. You see that you're almost out of gas, so you stop at a gas station in a small town. While you fill up the tank, Jonathan goes inside to buy some water and snacks. As you're checking the meter, you notice with some dismay a blinking light, which indicates that the car's heating system is broken. You finish filling up the gas and wait for Jonathan. Jonathan returns with some snacks, and you both set off. As you leave the station, he shivers in his seat. He turns to you and says, "Man, it's really cold in here."	What did Jonathan mean with his last utterance?	1. Could you turn on the heater? 2. I'm really cold; it's too bad the heater is broken. 3. We should fix the heater now. 4. We should get out of the car to warm up.
Indirect Suggestions	Noah lives in New Jersey, but his girlfriend Rachel lives in California. Noah often complains to his friend Jordan that he never gets to spend time with Rachel and that he misses her a lot. Today, Jordan and Noah are out to lunch and Noah is talking about how he decided to quit his job and move to Florida. Jordan is surprised by this decision and says: "Why don't you move to California?"	Noah lives in New Jersey, but his girlfriend Rachel lives in California. Noah often complains to his friend Jordan that he never gets to spend time with Rachel and that he misses her a lot. Today, Jordan and Noah are out to lunch and Noah is complaining about his long-distance relationship. Jordan knows that Noah would really like to move in with Rachel and that he isn't fully satisfied with his life in New Jersey. Jordan at one point says: "Why don't you move to California?"	What did Jordan mean with his last utterance?	1. I suggest you move to California. 2. Why are you moving to Florida and not California? 3. You shouldn't move to California. 4. California is uglier than Florida.
Indirect Declinations	I invited my brother Kevin to dinner to celebrate my birthday. Kevin and I rarely get to spend time together. For the occasion I prepared a fish dinner, and I bought some excellent Italian wine, our favorite. After a couple of drinks, I ask Kevin why he stopped drinking. Kevin replies "I need to drive".	I invited my brother Kevin to dinner to celebrate my birthday. Kevin and I rarely get to spend time together. For the occasion I prepared a fish dinner, and I bought some excellent Italian wine, our favorite. After a couple of drinks, I ask Kevin to pass me the glass so I can pour him more wine, but he tells me: "I need to drive".	What did Kevin mean with his last utterance?	1. Please, do not pour me more wine. 2. I drive tonight. 3. I don't like this wine. 4. We should drink French wine.
Indirect Threats	Janet and Monica are planning to go on a trip to the lake. They decide to invite their friend Brittany too. However, Brittany decides to join them after the picnic at Janet's house. Janet asks if she wants them to pick her up before they go home, but Brittany replies: "I know where you live".	Janet and Monica are planning to go on a trip to the lake. They decide to invite their friend Brittany too. However, Brittany decides to join them after the picnic at Janet's house. Once at Janet's house, Brittany reminds Janet that she has to pay her back the money she lent her and adds "I know where you live."	What did Brittany mean with her last utterance?	1. Give me my money back or I will come looking for you. 2. I know where your house is. 3. I really like your house. 4. I don't want to come to your house.
Sarcasm	Kyle and Mark play soccer on a team in their town. Today their team is playing the final. Kyle and Mark both play offense. During the match Mark makes a pass to Kyle and Kyle scores a goal. Mark tells him "Nice shot."	Kyle and Mark play soccer on a team in their town. Today their team is playing the final. Kyle and Mark both play offense. During the match Mark makes a pass to Kyle but Kyle misses the shot and fails to score a goal. Mark tells him "Nice shot."	What did Mark mean with his last utterance?	1. It was a really bad shot. 2. It was a really good shot. 3. You should try again. 4. You should take a shot.
Indirect Hyperbole	Monica and Rachel are two old friends who haven't met since university. After speaking by phone, they decide to meet for a coffee to see each other again and update on their lives. When they meet, Monica says "Haven't seen you in ages."	Monica and Rachel are two best friends. They went to the same high school and the same university, live in the same house, and practically spend all their time together. Tonight, Monica decides to go out with some of her friends and while she is at a restaurant she meets Rachel who is having dinner with her boyfriend. Monica tells Rachel "Haven't seen you in ages."	What did Monica mean with her last utterance?	1. We always see each other. 2. Haven't seen each other for a long time. 3. I don't want to see you again. 4. I would like to see you again soon.
Rhetorical Questions	Clara attends the art academy and often during the courses they must copy famous paintings. During today's lesson, Clara must reproduce a painting of her teacher. Clara is sitting in the last row and cannot distinguish the colors of the painting well, also because the colors are very dark. After drawing the draft, she begins to color, but is undecided about which are the right tempera colors to use, so she says to the teacher "Is the sky blue?"	Clara attends the art academy and often during the courses they must copy famous paintings. During today's lesson, Clara must reproduce a painting of her teacher. Clara is sitting in the last row and cannot distinguish the colors of the painting well, also because the colors are very dark. After drawing the draft, she begins to color. At a certain point one of her classmates asks Clara if she will use the brush to color and Clara says, "Is the sky blue?"	What did Clara mean with her last utterance?	1. The answer is obvious. 2. I want to know if the color of the sky is blue. 3. I think the sky is blue. 4. I don't think the sky is blue.

Table 6: Stimuli for each phenomenon, illustrating the two versions of the story. Literal correct answers are with number 2 and non-literal correct answers with number 1.