I don't Believe they Reason about Beliefs. Propositional Attitudes in Large Language Models

Alessandro Lenci

COmputational LINGuistics Laboratory
Università di Pisa
Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi)

Analyzing Language Understanding in Large Language Models (LLMs)

- LLMs give us the impression (illusion?) of having a super-human amount of knowledge they use to "understand" language and carry out different types of human-like reasoning

Some key questions:

- How do LLMs acquire the meaning of linguistic expressions?
- How do LLMs represent meaning and knowledge?
- Is language understanding in LLMs like human one?

Analyzing Language Understanding in Large Language Models (LLMs)

- LLMs give us the impression (illusion?) of having a super-human amount of knowledge they use to "understand" language and carry out different types of human-like reasoning

Some key questions:

- How do LLMs acquire the meaning of linguistic expressions?
- How do LLMs represent meaning and knowledge?
- Is language understanding in LLMs like human one?

Theory of Mind and Pragmatics

Theory of Mind (ToM)

Theory of Mind is the ability to track and reason about other people's mental states (beliefs, intentions, etc) and to use them to explain and predict their behavior.

- ToM is central to human social interaction and communication
- Pragmatic reasoning (e.g., Speech Act identification, Irony, etc.) is grounded on ToM, for instance in Grice's paradigm

Grice intended (utterer's) meaning (Levinson 2000)

S means *p* by uttering U to A iff S intends:

a. A to think *p*

b. A to recognize that S intends (a)

c. A's recognition of S's intending (a) to be the prime reason for A thinking *p*

Theory of Mind and Pragmatics

Theory of Mind (ToM)

Theory of Mind is the ability to track and reason about other people's mental states (beliefs, intentions, etc) and to use them to explain and predict their behavior.

- ToM is central to human social interaction and communication
- Pragmatic reasoning (e.g., Speech Act identification, Irony, etc.) is grounded on ToM, for instance in Grice's paradigm

Grice intended (utterer's) meaning (Levinson 2000)

S means *p* by uttering U to A iff S intends:

a. A to think *p*

b. A to recognize that S intends (a)

c. A's recognition of S's intending (a) to be the prime reason for A thinking *p*

Theory of Mind and Pragmatics

Theory of Mind (ToM)

Theory of Mind is the ability to track and reason about other people's mental states (beliefs, intentions, etc) and to use them to explain and predict their behavior.

- ToM is central to human social interaction and communication
- Pragmatic reasoning (e.g., Speech Act identification, Irony, etc.) is grounded on ToM, for instance in Grice's paradigm

Grice intended (utterer's) meaning (Levinson 2000)

S means *p* by uttering U to A iff S intends:

a. A to think *p*

b. A to recognize that S intends (a)

c. A's recognition of S's intending (a) to be the prime reason for A thinking *p*

Theory of Mind and Pragmatics

Theory of Mind (ToM)

Theory of Mind is the ability to track and reason about other people's mental states (beliefs, intentions, etc) and to use them to explain and predict their behavior.

- ToM is central to human social interaction and communication
- Pragmatic reasoning (e.g., Speech Act identification, Irony, etc.) is grounded on ToM, for instance in Grice's paradigm

Grice intended (utterer's) meaning (Levinson 2000)

S means *p* by uttering U to A iff S intends:

a. A to think *p*

b. A to recognize that S intends (a)

c. A's recognition of S's intending (a) to be the prime reason for A thinking *p*

Pragmatic Reasoning and ToM in GOFAI

- In Good Old-Fashioned AI (GOFAI), ToM was modeled with propositional attitudes explicitly represented with recursive symbolic statements
  - A BELIEVES P
- Pragmatic reasoning was modeled as an inferential process over sets of statements of propositional attitudes (e.g., Perrault et al. 1978)

```
REQUEST(SP,H,ACT)
preconditions:
SP BELIEVE H CANDO ACT
SP BELIEVE H BELIEVE H CANDO ACT
SP BELIEVE SP WANT TO REQUEST
effects:
H BELIEVE SP BELIEVE SP WANT H TO ACT

INFORM(SP,H,PROP)
preconditions:
SP BELIEVE PROP
SP BELIEVE SP WANT TO INFORM
effects:
H BELIEVE SP BELIEVE PROP
```

The Pragmatic Competence of LLMs

the case of indirect speech acts

Yesterday evening, John asked Mary to go to the cinema with him. Mary replied: "Tomorrow, I have an exam". What did Mary intend to say?

Mary intended to say that she couldn't go to the cinema with John because she needed to prepare for her exam, which is scheduled for the next day. Her response implies that she has to focus on studying and is not available to go out.

How Do LLMs Work?

- LLMs are deep (i.e., multi-layered) artificial neural networks pretrained on huge amounts of unlabeled data
  - the network acquires a large amount of knowledge about natural language structure and meaning from text corpora as a side effect of being trained with a self-supervised string prediction task (language modeling)
  - the model's knowledge is encoded in the vectors corresponding to the internal layers of the network
  - the model's knowledge consists of all and only the information that can be recovered from the distributional statistics in the training corpus (Lenci and Sahlgren 2023)
- Prompting
  - a task description is provided to the LLM as a natural language string (prompt)
  - the answer of the LLM is the most likely text string given the prompt

How Do LLMs Work?

- LLMs are deep (i.e., multi-layered) artificial neural networks pretrained on huge amounts of unlabeled data
  - the network acquires a large amount of knowledge about natural language structure and meaning from text corpora as a side effect of being trained with a self-supervised string prediction task (language modeling)
  - the model's knowledge is encoded in the vectors corresponding to the internal layers of the network
  - the model's knowledge consists of all and only the information that can be recovered from the distributional statistics in the training corpus (Lenci and Sahlgren 2023)
- Prompting
  - a task description is provided to the LLM as a natural language string (prompt)
  - the answer of the LLM is the most likely text string given the prompt

How Do LLMs Work?

- LLMs are deep (i.e., multi-layered) artificial neural networks pretrained on huge amounts of unlabeled data
  - the network acquires a large amount of knowledge about natural language structure and meaning from text corpora as a side effect of being trained with a self-supervised string prediction task (language modeling)
  - the model's knowledge is encoded in the vectors corresponding to the internal layers of the network
  - the model's knowledge consists of all and only the information that can be recovered from the distributional statistics in the training corpus (Lenci and Sahlgren 2023)
- Prompting
  - a task description is provided to the LLM as a natural language string (prompt)
  - the answer of the LLM is the most likely text string given the prompt

How Can LLMs Learn Pragmatic Competence?

- Beliefs and intentions are encoded in and recoverable from distributional statistics
  - "In the course of performing next-word prediction in context, current LMs sometimes infer approximate, partial representations of the beliefs, desires and intentions possessed by the agent that produced the context, and other agents mentioned within it." (Andreas 2022)
  - cf. Symbol Interdependency Hypothesis (Louwerse 2011): sensorimotor information is also encoded in language
- Some pragmatic meanings are strongly conventionalized in language (e.g., *Could you pass me the salt?*)
  - ToM is not always required to decode communicative intentions (the non-literal reading can be the default interpretation)
- Pragmatic abilities are shaped by fine-tuning LLMs with human data
  - cf. instruction tuning and Reinforcement Learning from Human Feedback

How Can LLMs Learn Pragmatic Competence?

- Beliefs and intentions are encoded in and recoverable from distributional statistics
  - "In the course of performing next-word prediction in context, current LMs sometimes infer approximate, partial representations of the beliefs, desires and intentions possessed by the agent that produced the context, and other agents mentioned within it." (Andreas 2022)
  - cf. Symbol Interdependency Hypothesis (Louwerse 2011): sensorimotor information is also encoded in language
- Some pragmatic meanings are strongly conventionalized in language (e.g., *Could you pass me the salt?*)
  - ToM is not always required to decode communicative intentions (the non-literal reading can be the default interpretation)
- Pragmatic abilities are shaped by fine-tuning LLMs with human data
  - cf. instruction tuning and Reinforcement Learning from Human Feedback

How Can LLMs Learn Pragmatic Competence?

- Beliefs and intentions are encoded in and recoverable from distributional statistics
  - "In the course of performing next-word prediction in context, current LMs sometimes infer approximate, partial representations of the beliefs, desires and intentions possessed by the agent that produced the context, and other agents mentioned within it." (Andreas 2022)
  - cf. Symbol Interdependency Hypothesis (Louwerse 2011): sensorimotor information is also encoded in language
- Some pragmatic meanings are strongly conventionalized in language (e.g., *Could you pass me the salt?*)
  - ToM is not always required to decode communicative intentions (the non-literal reading can be the default interpretation)
- Pragmatic abilities are shaped by fine-tuning LLMs with human data
  - cf. instruction tuning and Reinforcement Learning from Human Feedback

What Do LLMs Know about ToM and Pragmatics?

Hu et al. (2023), "A fine-grained comparison of pragmatic language understanding in humans and language models", *Proceedings of ACL*

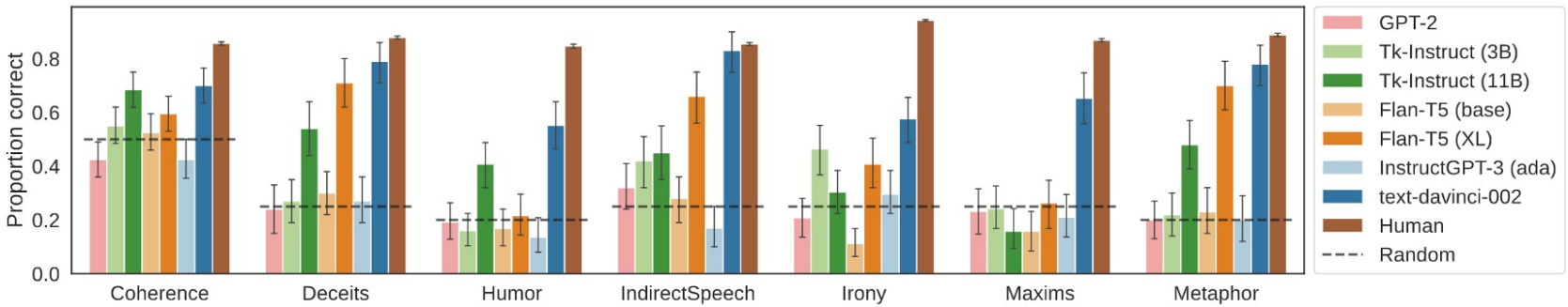


Figure 1: Accuracy for each task. Error bars denote 95% CI. Dashed line indicates task-specific random baseline.

- cf. Barattieri di San Pietro et al. (2023) for similar results in Italian

What Do LLMs Know about ToM and Pragmatics?

- Ongoing debate about the true ToM abilities of LLMs (Kosinski 2023, Marchetti et al. 2023, Strachan 2024)
- Benchmarks are ToM battery tests designed for psychological experiments with humans
  - false belief, strange stories, faux pas, indirect requests, irony, etc.
- The ToM and pragmatic abilities of LLMs are still controversial
  - some experiments report performances equal or even above humans in some tasks, but not in others (Strachan et al. 2024)
  - models do not have robust ToM abilities and can fail on small alterations of the original task (Ullman 2023, Shapira et a. 2024)
  - LLMs are likely to rely on shallow statistical correlations in the data (clever Hans effect)
  - methodological problems in the same benchmarks used to test ToM in LLMs

What Do LLMs Know about ToM and Pragmatics?

- Ongoing debate about the true ToM abilities of LLMs (Kosinski 2023, Marchetti et al. 2023, Strachan 2024)
- Benchmarks are ToM battery tests designed for psychological experiments with humans
  - false belief, strange stories, faux pas, indirect requests, irony, etc.
- The ToM and pragmatic abilities of LLMs are still controversial
  - some experiments report performances equal or even above humans in some tasks, but not in others (Strachan et al. 2024)
  - models do not have robust ToM abilities and can fail on small alterations of the original task (Ullman 2023, Shapira et a. 2024)
  - LLMs are likely to rely on shallow statistical correlations in the data (clever Hans effect)
  - methodological problems in the same benchmarks used to test ToM in LLMs

What Do LLMs Know about ToM and Pragmatics?

- Ongoing debate about the true ToM abilities of LLMs (Kosinski 2023, Marchetti et al. 2023, Strachan 2024)
- Benchmarks are ToM battery tests designed for psychological experiments with humans
  - false belief, strange stories, faux pas, indirect requests, irony, etc.
- The ToM and pragmatic abilities of LLMs are still controversial
  - some experiments report performances equal or even above humans in some tasks, but not in others (Strachan et al. 2024)
  - models do not have robust ToM abilities and can fail on small alterations of the original task (Ullman 2023, Shapira et a. 2024)
  - LLMs are likely to rely on shallow statistical correlations in the data (clever Hans effect)
  - methodological problems in the same benchmarks used to test ToM in LLMs

Pragmatic Explorations of LLMs

Joint work with Agnese Lombardi, Univ. Pisa

Question

Can LLMs infer the correct pragmatic interpretation of an utterance, when it requires reasoning on the beliefs of the participants in a story?

Pragmatic Explorations of LLMs

Joint work with Agnese Lombardi, Univ. Pisa

Question

Can LLMs infer the correct pragmatic interpretation of an utterance, when it requires reasoning on the beliefs of the participants in a story?

What Does it Mean Having a ToM?

ToM and Communication

A communicative agent has a ToM iff

- it represents information in terms of the content of different propositional attitudes (e.g., beliefs and intentions)
  - e.g., A *believes* p, A *intends* p, etc.
- it represents the fact that agents have recursive propositional attitudes and different propositional attitudes about the same information content
  - e.g., A *believes* that B *believes* that p, but B *does not believe* that p
- it reasons and draws inferences based on the representation of its own and other agents' mental states
- it uses its representation of mental states and inferences about other agents' mental states to generate and interpret utterances

What Does it Mean Having a ToM?

ToM and Communication

A communicative agent has a ToM iff

- it represents information in terms of the content of different propositional attitudes (e.g., beliefs and intentions)
  - e.g., A *believes p*, A *intends p*, etc.
- it represents the fact that agents have recursive propositional attitudes and different propositional attitudes about the same information content
  - e.g., A *believes* that B *believes* that *p*, but B *does not believe* that *p*
- it reasons and draws inferences based on the representation of its own and other agents' mental states
- it uses its representation of mental states and inferences about other agents' mental states to generate and interpret utterances

What Does it Mean Having a ToM?

ToM and Communication

A communicative agent has a ToM iff

- it represents information in terms of the content of different propositional attitudes (e.g., beliefs and intentions)
  - e.g., A *believes p*, A *intends p*, etc.
- it represents the fact that agents have recursive propositional attitudes and different propositional attitudes about the same information content
  - e.g., A *believes* that B *believes* that *p*, but B *does not believe* that *p*
- it reasons and draws inferences based on the representation of its own and other agents' mental states
- it uses its representation of mental states and inferences about other agents' mental states to generate and interpret utterances

What Does it Mean Having a ToM?

ToM and Communication

A communicative agent has a ToM iff

- it represents information in terms of the content of different propositional attitudes (e.g., beliefs and intentions)
  - e.g., A *believes p*, A *intends p*, etc.
- it represents the fact that agents have recursive propositional attitudes and different propositional attitudes about the same information content
  - e.g., A *believes* that B *believes* that *p*, but B *does not believe* that *p*
- it reasons and draws inferences based on the representation of its own and other agents' mental states
- it uses its representation of mental states and inferences about other agents' mental states to generate and interpret utterances

What Does it Mean Having a ToM?

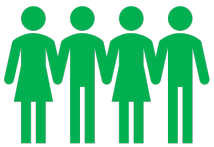
ToM and Communication

A communicative agent has a ToM iff

- it represents information in terms of the content of different propositional attitudes (e.g., beliefs and intentions)
  - e.g., A *believes p*, A *intends p*, etc.
- it represents the fact that agents have recursive propositional attitudes and different propositional attitudes about the same information content
  - e.g., A *believes* that B *believes* that *p*, but B *does not believe* that *p*
- it reasons and draws inferences based on the representation of its own and other agents' mental states
- it uses its representation of mental states and inferences about other agents' mental states to generate and interpret utterances

Experimental Approach

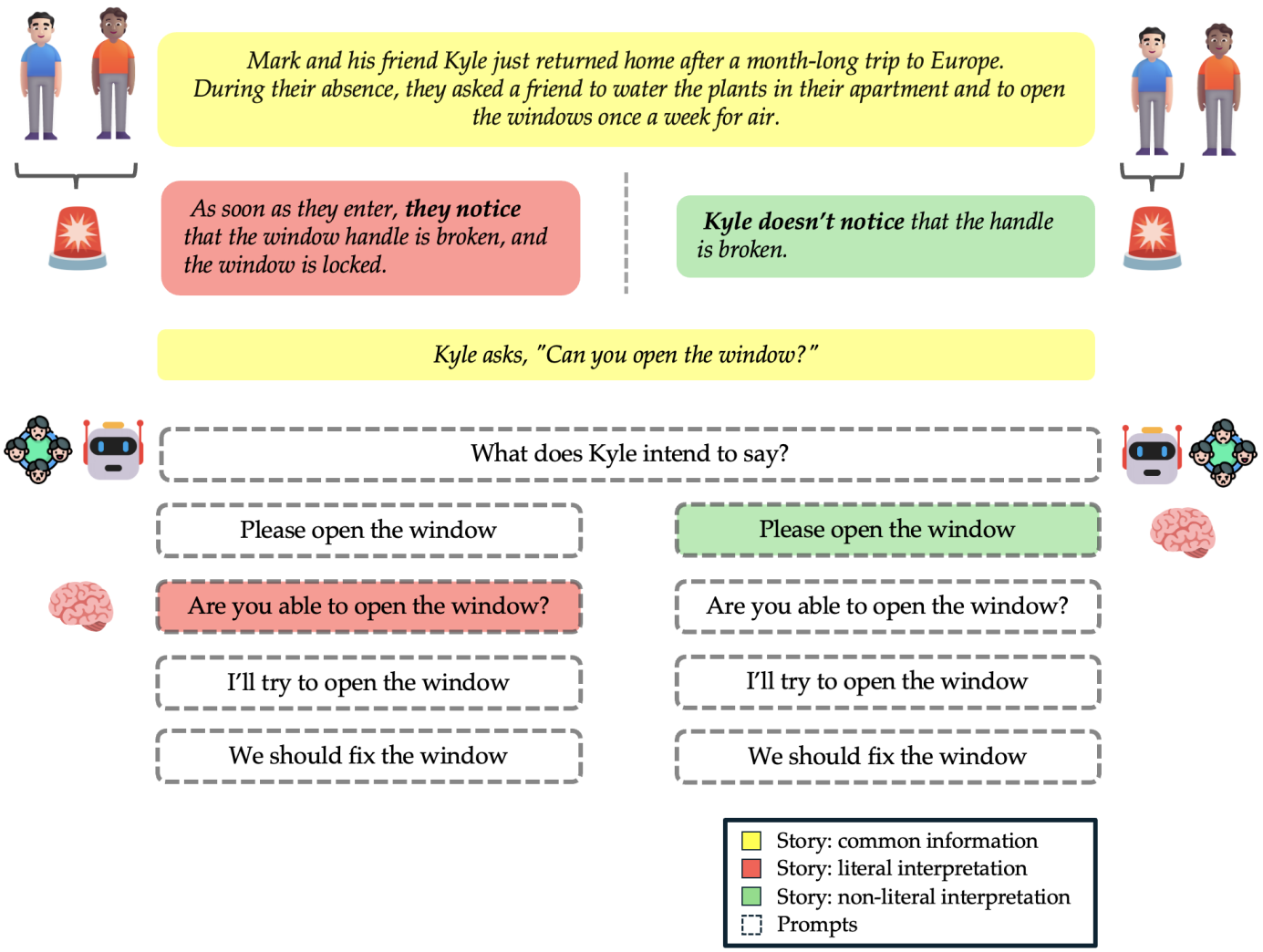
False-Belief inspired design



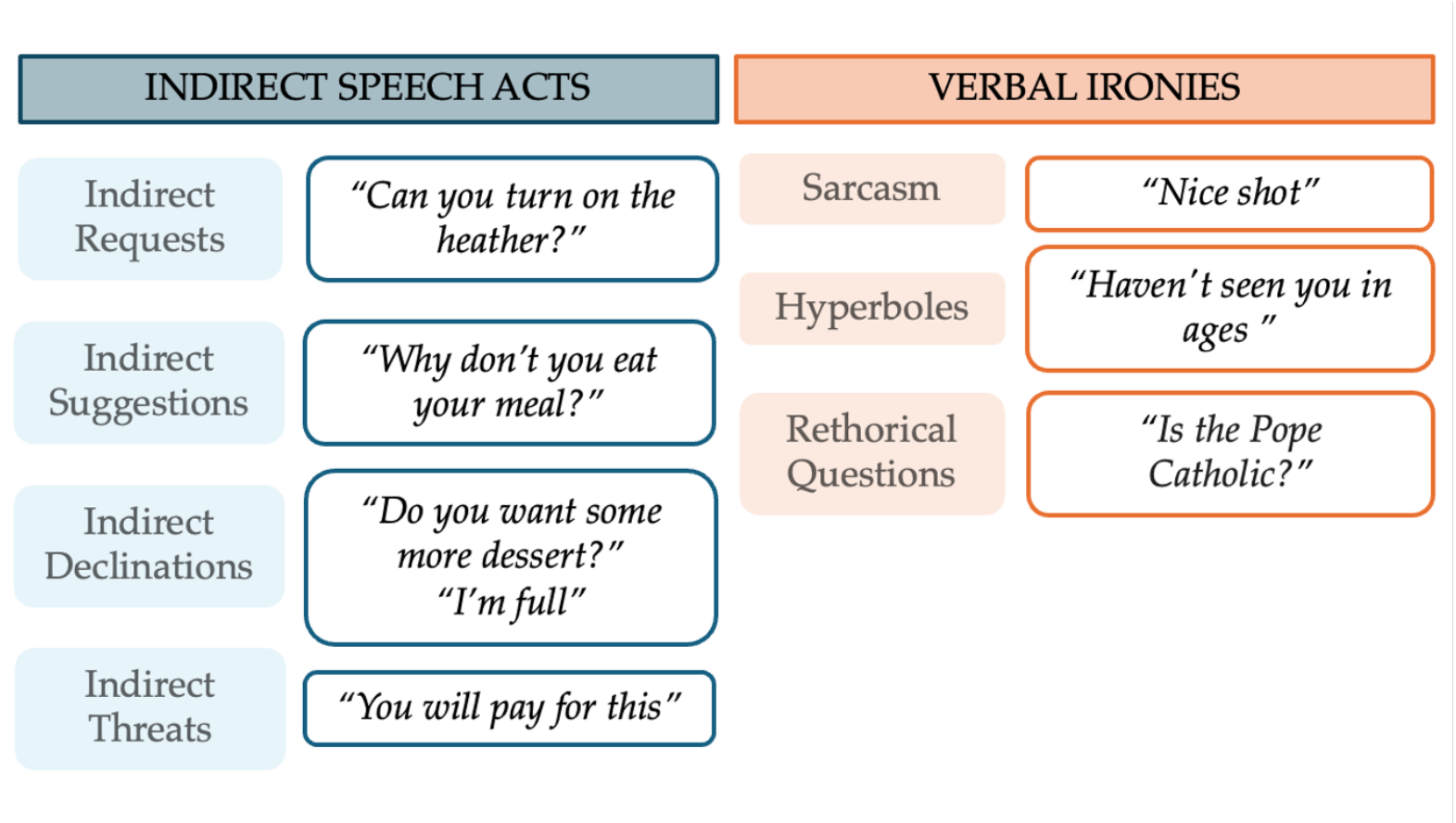
Obstacle to derive a non-literal interpretation of an utterance

Interpretation depends on **reasoning about beliefs and intentions** (mentalizing)

Experimental Setting



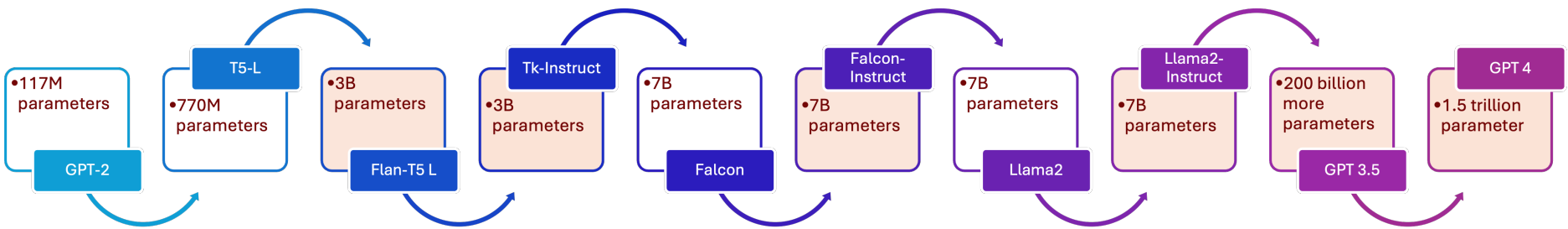
Experimental Material



Experimental Material

- Human judgments
  - 8 balanced groups for ISAs and 4 balanced groups for Irony, each tested on 30 subjects recruited with Prolific
  - 360 subjects in total

Models


•117M parameters
GPT-2
T5-L
•770M parameters
•3B parameters
Flan-T5 L
Tk-Instruct
•3B parameters
•7B parameters
Falcon
Falcon-Instruct
•7B parameters
•7B parameters
Llama2
Llama2-Instruct
•7B parameters
•200 billion more parameters
GPT 3.5
GPT 4
•1.5 trillion parameter

Prompt

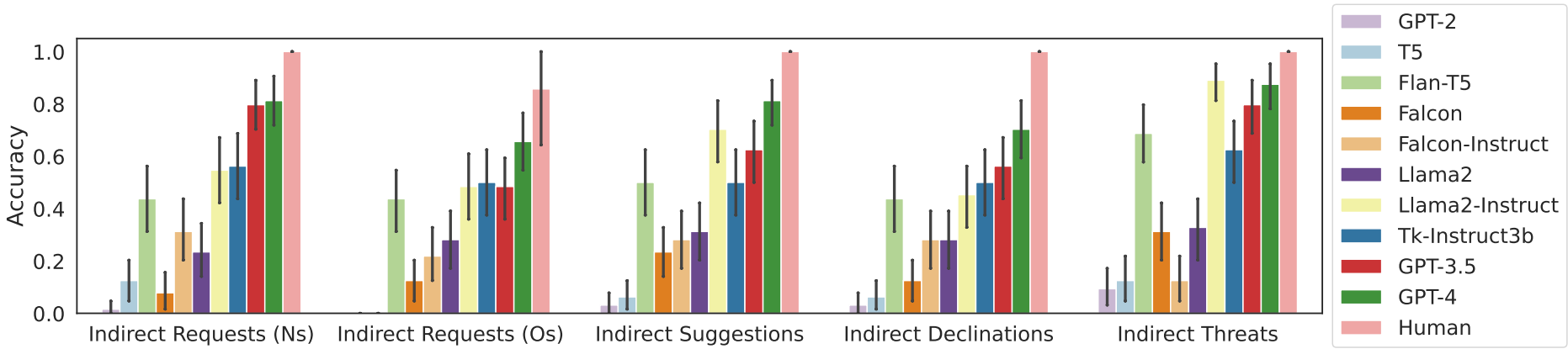
"Task: **I'll give you a story and I'll ask you to answer a question about one character of the story. I'll give you four possible answers to the question and you must choose the right one. The possible answers are 1, 2, 3 and 4.**

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”.

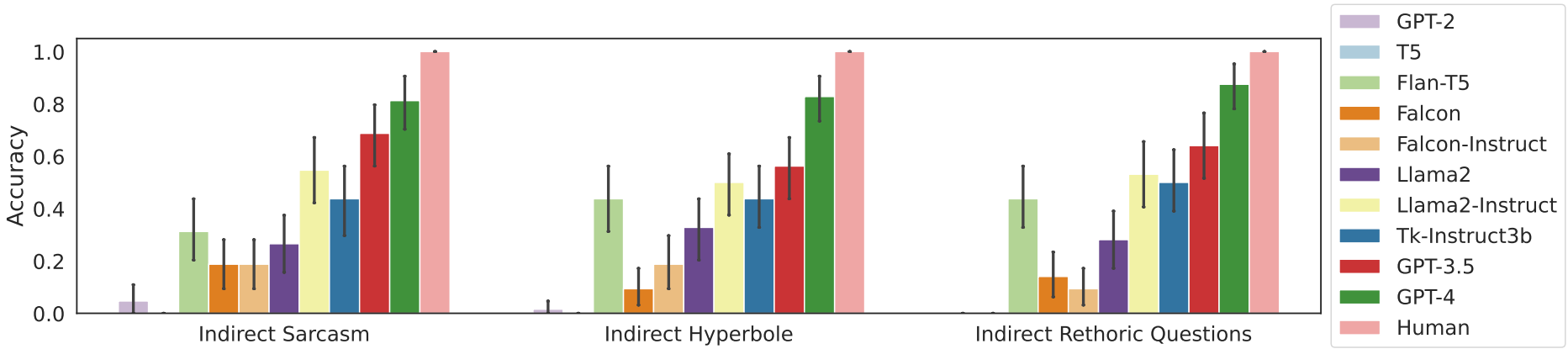
What does Kevin intend to say?
Options:
1 Please, do not pour me more wine.
2 I drive tonight
3 I don’t like this wine.
4 We should drink French wine.

The correct answer is:"

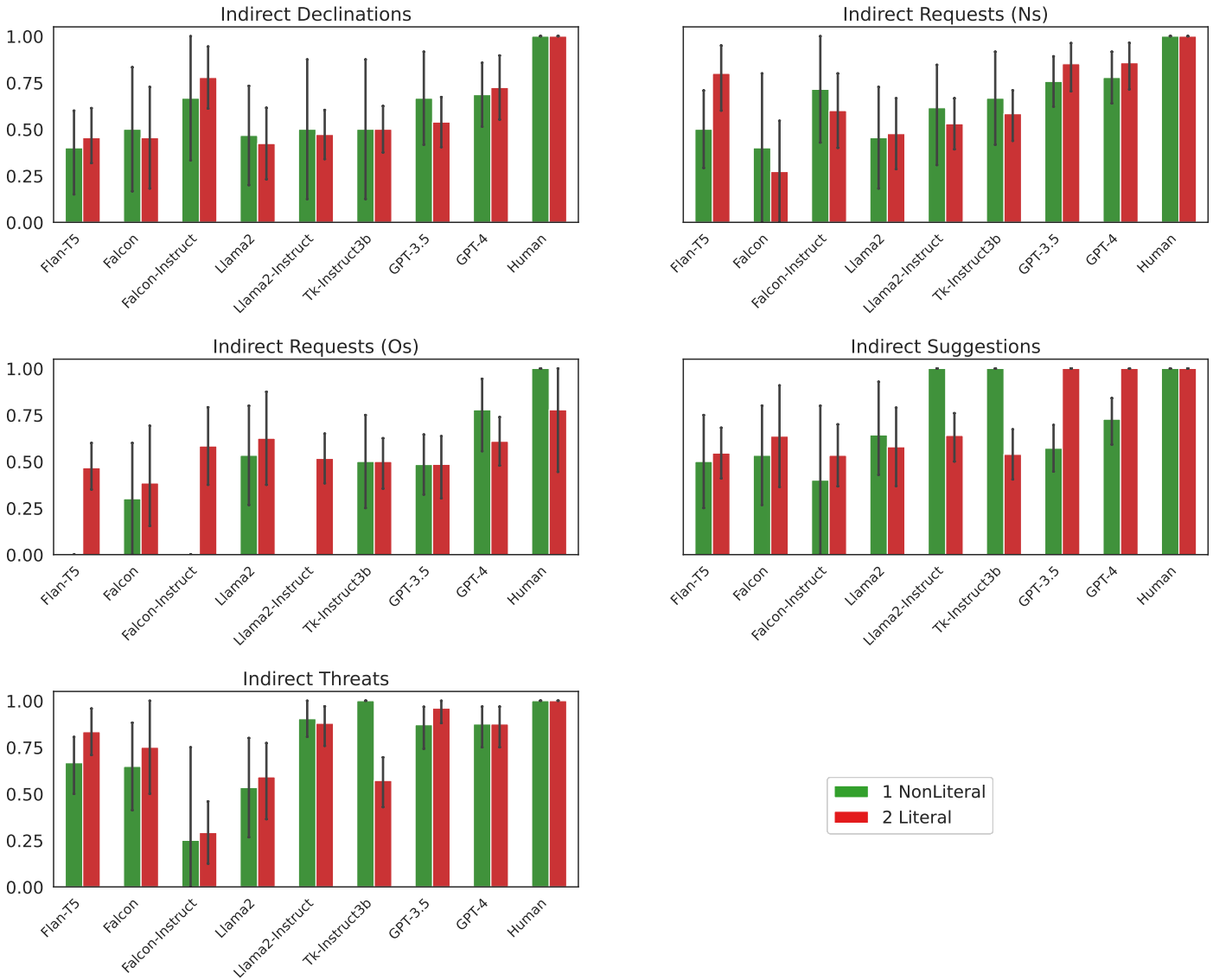
Results - ISA


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Indirect Requests (Ns)
Indirect Requests (Os)
Indirect Suggestions
Indirect Declinations
Indirect Threats
GPT-2
T5
Flan-T5
Falcon
Falcon-Instruct
Llama2
Llama2-Instruct
Tk-Instruct3b
GPT-3.5
GPT-4
Human

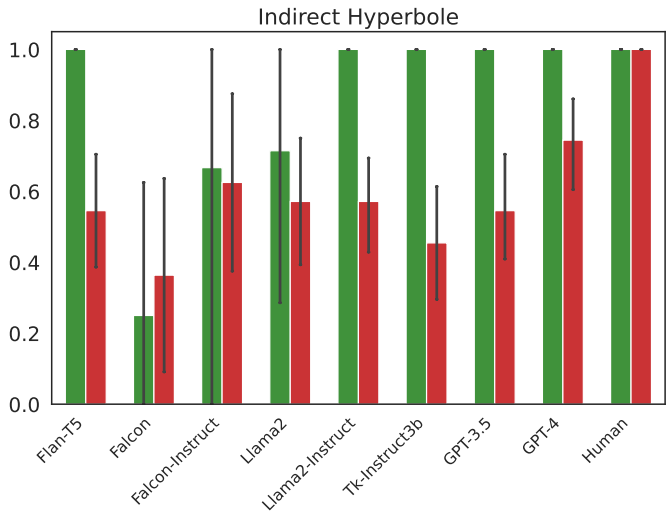
Results - Irony

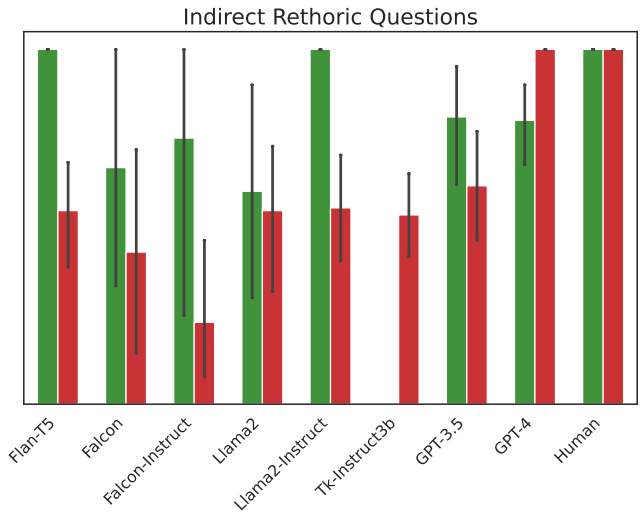


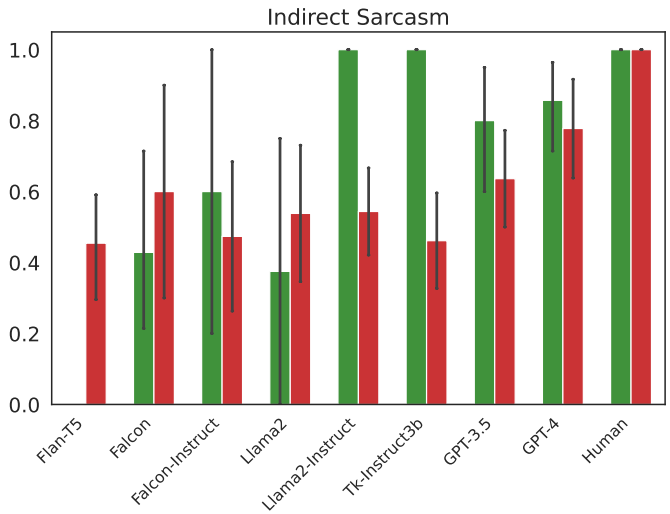
Literal vs. Non-Literal Interpretation - ISA

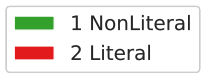


Literal vs. Non-Literal Interpretation - Irony









Do LLMs Really Consider Belief States?

- The model is prompted to answer the same question, both without story context (ToM-0) and with a neutral story context (ToM-N)

Questions

- Are LLMs biased towards an interpretation independently of the belief context?
- Do LLMS really take believe states into account when solving a ToM task?

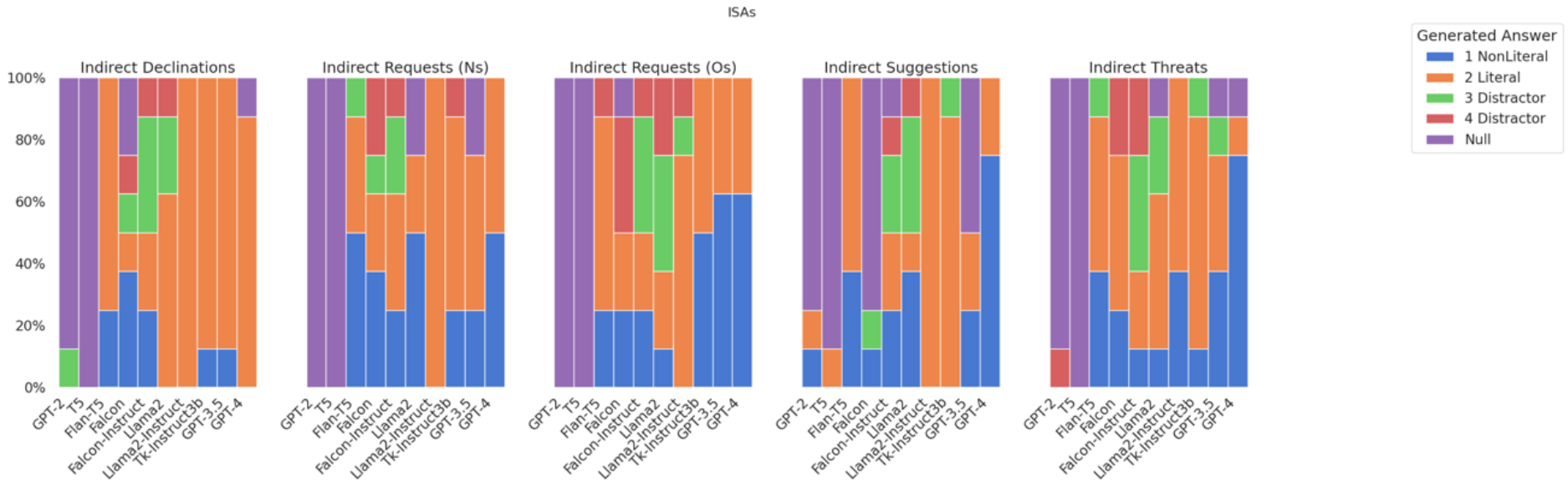
ToM - prompt

```
"Task: I will give you a sentence
said from one character. Then I'll
ask you the meaning of the
sentence. 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. Sentence: Kevin says, “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:"
```

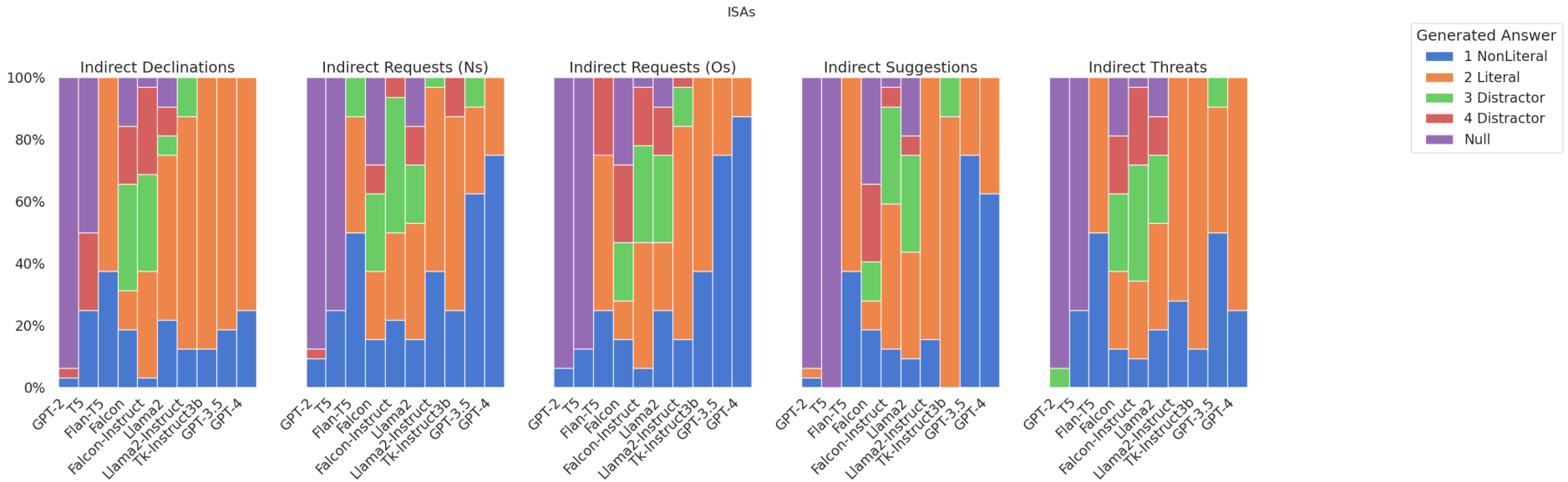
ToM + prompt

```
"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 favourite. After a
couple of drinks, Kevin says “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:"
```

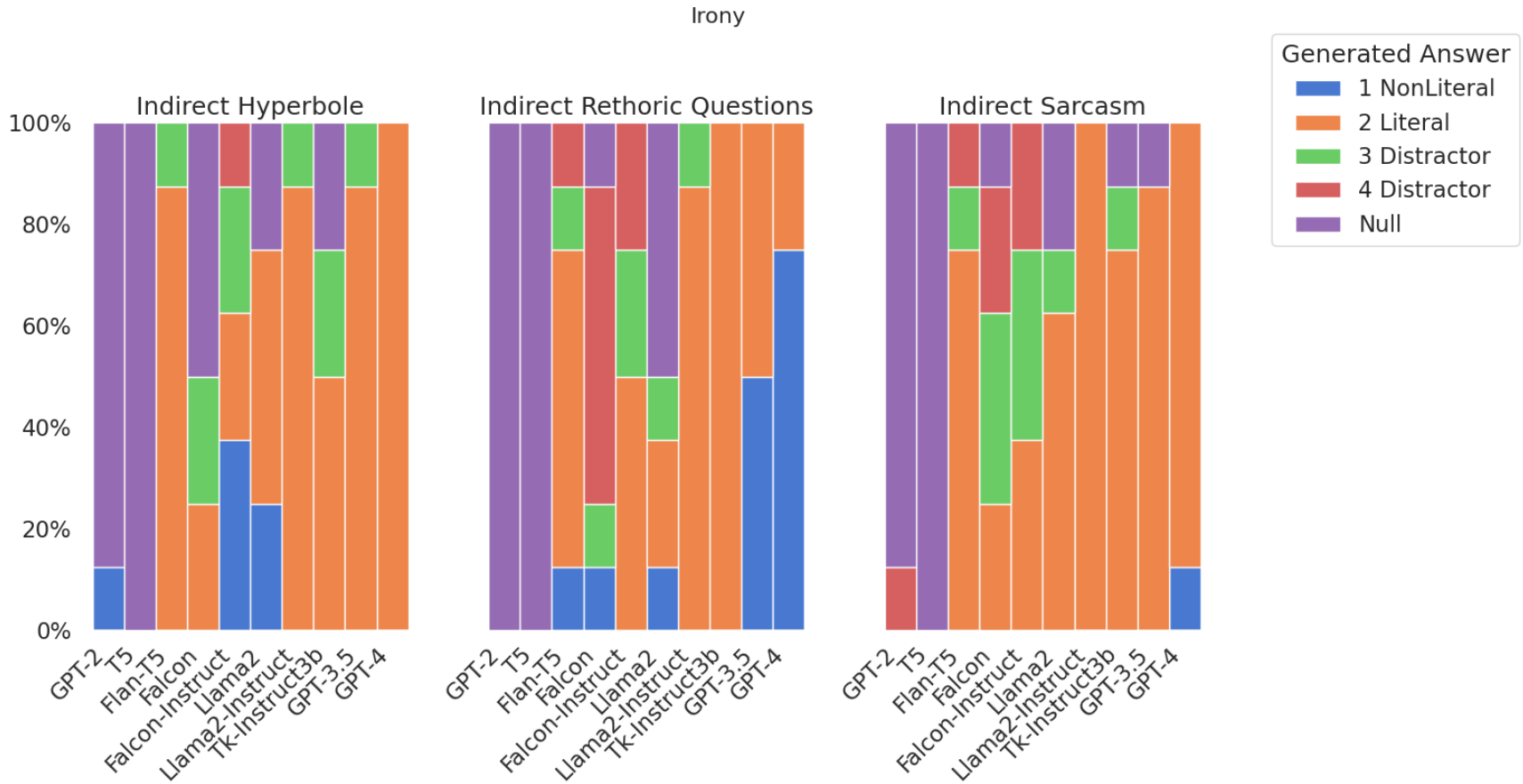
Indirect Speech Acts without Story Context (ToM-0)



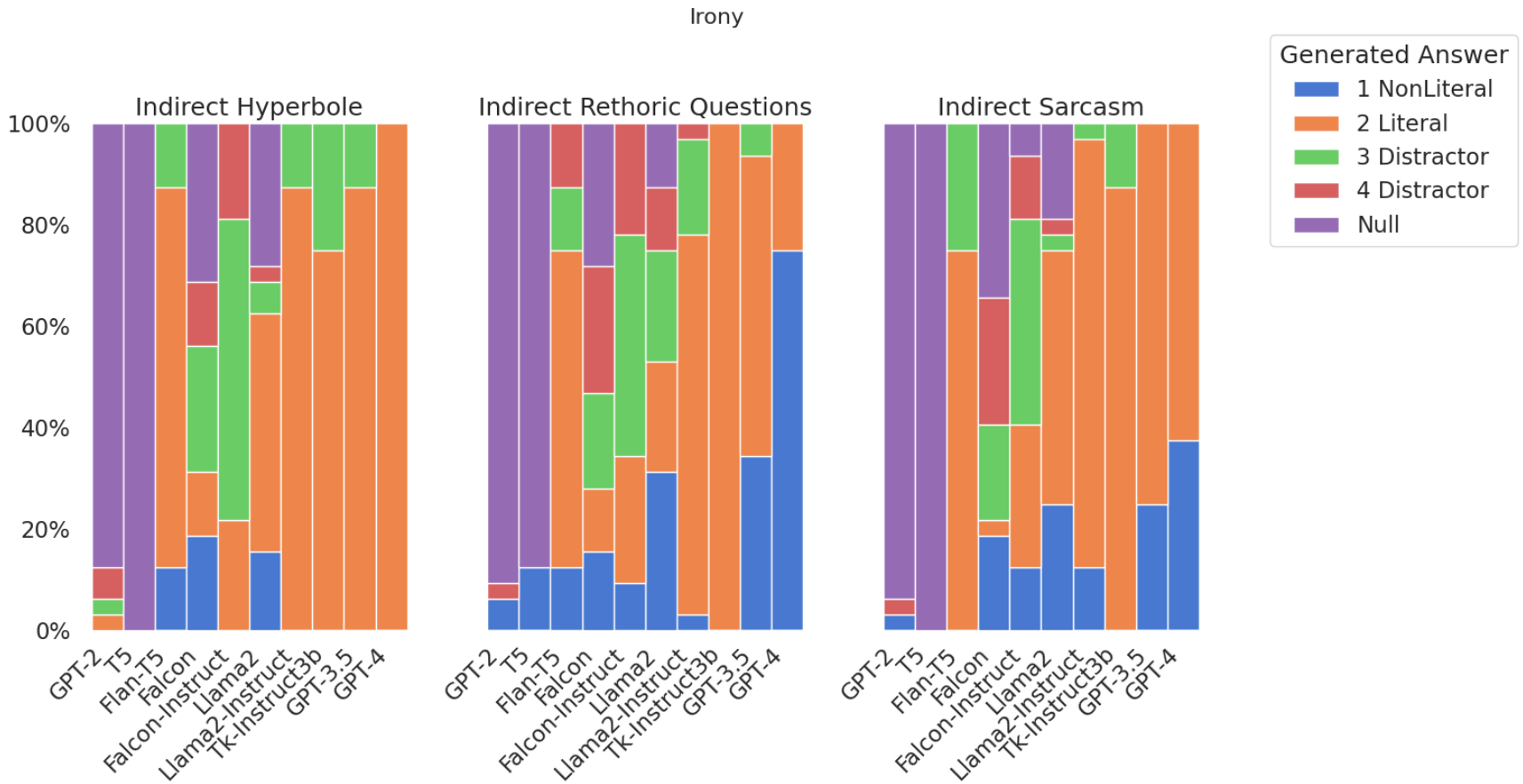
Indirect Speech Acts with ToM-N Context



Verbal Irony without story context (ToM-0)



Verbal Irony without story context with a neutral story context (ToM-N)



Pearson's Residual Significance Evaluation

- The significance values of the ToM- and ToM+ prompts are compared with those of the original prompts that contain belief alternation (ToM), to determine whether the observed difference is statistically significant
  - † = no significant difference

Model	Indirect Declinations	Indirect Requests (Ns)	Indirect Requests (Os)	Indirect Suggestions	Indirect Threats
Flan-T5	†	†	†	†	†
Falcon	†	†	†	†	†
Falcon-Instruct	†	†	†	†	†
Llama2	†	†	†	†	†
Llama2-Instruct	†	†	†	†	†
Tk-Instruct3b	†	†	†	†	†
GPT-3.5	†	†	†	†	†
GPT-4	*	†	†	†	†

Pearson's Residual Significance Evaluation

- The significance values of the ToM-0 and ToM-N prompts are compared with those of the original prompts that contain belief change (ToM), to determine whether the observed difference is statistically significant
  - † = no significant difference

Model	Indirect Hyperbole	Indirect Rhetoric Questions	Indirect Sarcasm
Flan-T5	†	†	-
Falcon	†	†	†
Falcon-Instruct	†	†	†
Llama2	†	†	†
GPT-3.5	†	*	*
GPT-4	*	*	†

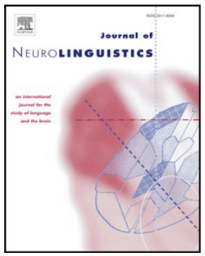
Plausible vs. Implausible Events in Propositional Attitudes

Journal of Neurolinguistics 53 (2020) 100877

Contents lists available at ScienceDirect

Journal of Neurolinguistics

journal homepage: www.elsevier.com/locate/jneuroling



Embedding (im)plausible clauses in propositional attitude contexts: Modulatory effects on the N400 and late components

Lia Călinescu[1], Anna Giskes[1], Mila Vulchanova, Giosuè Baggio[*]

Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway

Plausible vs. Implausible Events in Propositional Attitudes

- Propositional attitude verbs can change the plausibility of embedded events

(1) a. Cars have wheels. (plausible)
b. Cars have wings. (implausible)

(2) a. Magnus knows that cars have wheels. (plausible)
b. Magnus knows that cars have wings. (implausible)

(3) a. Magnus believes that cars have wheels. (plausible)
b. Magnus believes that cars have wings. (plausible)

Plausible vs. Implausible Events in Propositional Attitudes

- Propositional attitude verbs can change the plausibility of embedded events

(1) a. Cars have wheels. (plausible)
b. Cars have wings. (implausible)

(2) a. Magnus knows that cars have wheels. (plausible)
b. Magnus knows that cars have wings. (implausible)

(3) a. Magnus believes that cars have wheels. (plausible)
b. Magnus believes that cars have wings. (plausible)

Plausible vs. Implausible Events in Propositional Attitudes

- Factuality Scale: know > believe > dream > doubt > imagine

- Declerk (2011), "The definition of modality"
  - factive verbs (e.g., know) evoke a world which is "automatically interpreted as being the factual world" (p. 41)
  - attitude verbs (e.g., believe, doubt, dream, imagine, etc.) create an "intensional world which may or may not coincide with the factual world"

Plausible vs. Implausible Events in Propositional Attitudes

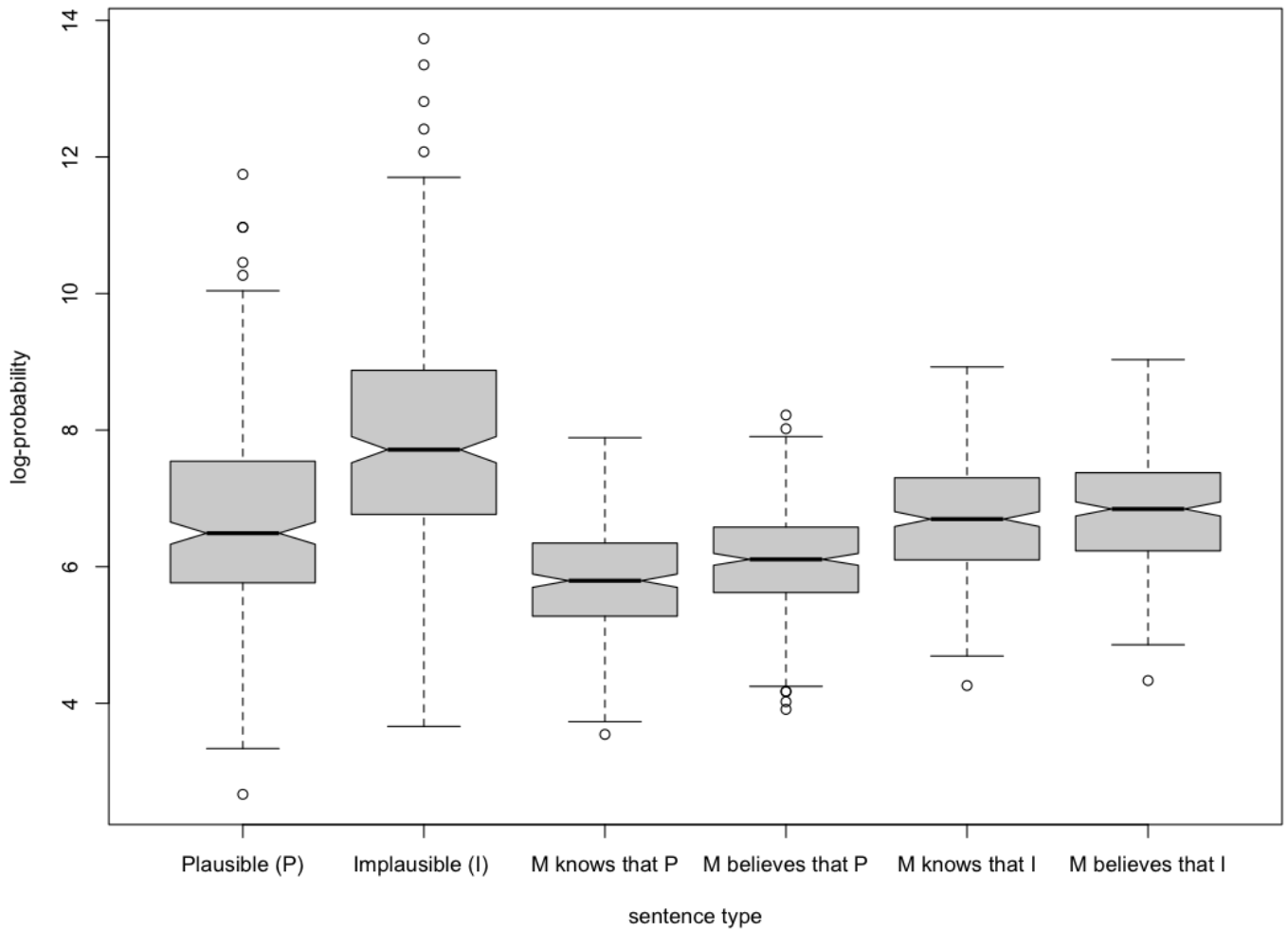
- Factuality Scale: know > believe > dream > doubt > imagine

- Declerk (2011), "The definition of modality"
  - factive verbs (e.g., know) evoke a world which is "automatically interpreted as being the factual world" (p. 41)
  - attitude verbs (e.g., believe, doubt, dream, imagine, etc.) create an "intensional world which may or may not coincide with the factual world"

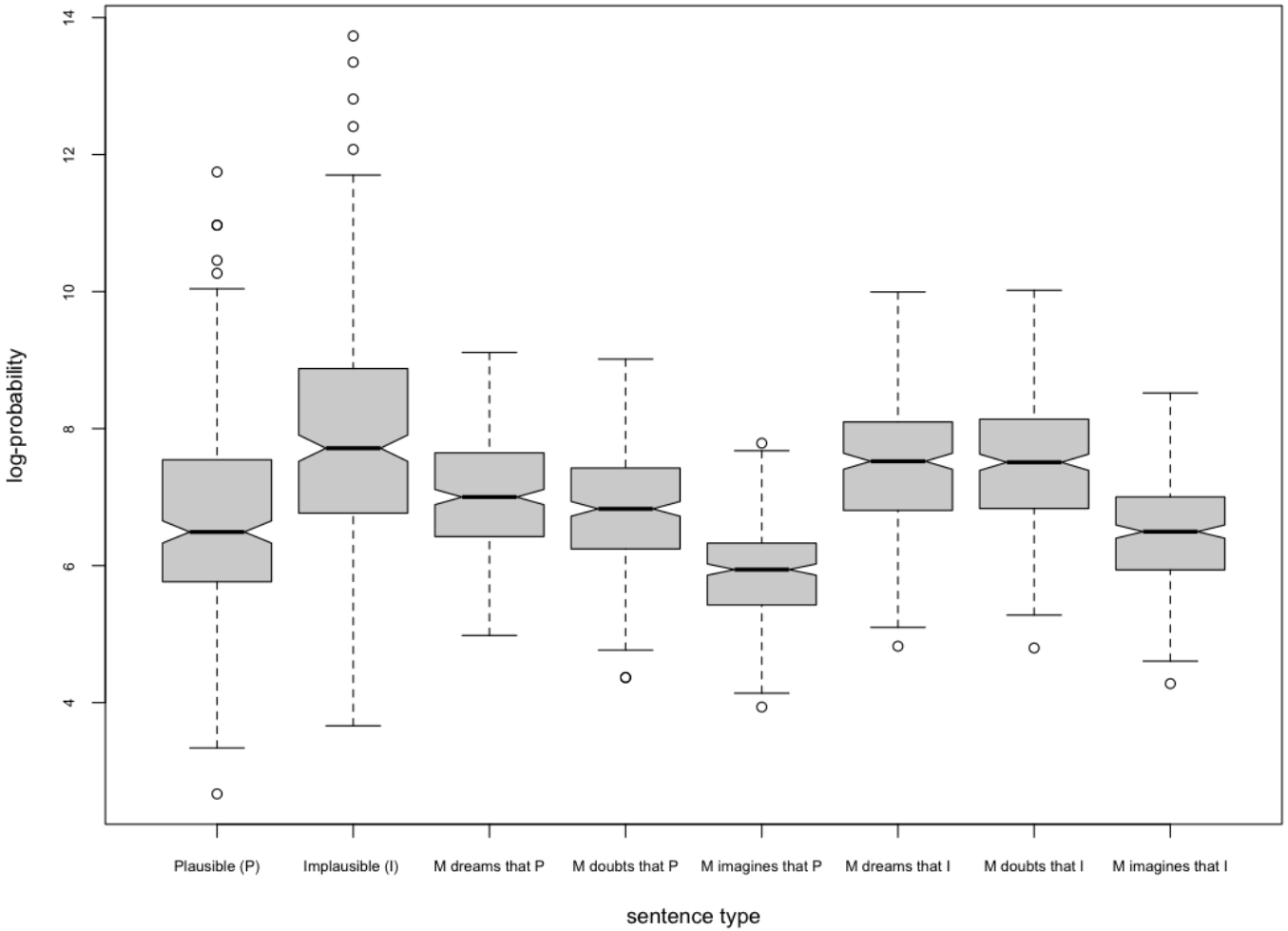
Experimental Setting

- Stimuli from Călinescu et al. (2020)
  - 300 plausible sentences (P) + 300 implausible sentences (I)
  - the P and I sentences were embedded in propositional attitude contexts with 5 different verbs differing for factuality (know, believe, dream, doubt, imagine), for a total of 3,600 data points
- Model: Llama-3 8B Instruct
- Measure: the LLM computed the log-probability scores (Kauf et al. 2023, 2024) of the P and I sentences both as main clauses and as embedded ones

Results


log-probability
4
6
8
10
12
14
Plausible (P)
Implausible (I)
M knows that P
M believes that P
M knows that I
M believes that I
sentence type

Results


log-probability
4
6
8
10
12
14
Plausible (P)
Implausible (I)
M dreams that P
M doubts that P
M imagines that P
M dreams that I
M doubts that I
M imagines that I
sentence type

Language Without Thought

Mahowald et al. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*

"good at language → good at thought" fallacy

If an entity (be it human or a machine) generates long coherent stretches of text, it must possess rich knowledge and reasoning capacities

- Mahowald et al. (2024) distinguish between:
  - formal linguistic competence, that is knowledge of linguistic rules and patterns
  - functional competence, which roughly corresponds to inferential competence
    formal reasoning (logical reasoning and novel problem solving), world knowledge (knowledge of objects and events and their properties, participants and relations), situation modeling (the ability of building a representation of the stories), social reasoning (Theory of Mind)
- LLMs have an almost human-like formal competence, but still fall short of functional competence
  - cf. LLMs as "cultural technology" that only imitates human language production (Yiu et al. 2023; also termed as bibliotechnism by Lederman & Mahowald 2024)

Language Without Thought

Mahowald et al. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*

"good at language → good at thought" fallacy

If an entity (be it human or a machine) generates long coherent stretches of text, it must possess rich knowledge and reasoning capacities

- Mahowald et al. (2024) distinguish between:
  - formal linguistic competence, that is knowledge of linguistic rules and patterns
  - functional competence, which roughly corresponds to inferential competence formal reasoning (logical reasoning and novel problem solving), world knowledge (knowledge of objects and events and their properties, participants and relations), situation modeling (the ability of building a representation of the stories), social reasoning (Theory of Mind)
- LLMs have an almost human-like formal competence, but still fall short of functional competence
  - cf. LLMs as "cultural technology" that only imitates human language production (Yiu et al. 2023; also termed as bibliotechnism by Lederman & Mahowald 2024)

Language Without Thought

Mahowald et al. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*

"good at language → good at thought" fallacy

If an entity (be it human or a machine) generates long coherent stretches of text, it must possess rich knowledge and reasoning capacities

- Mahowald et al. (2024) distinguish between:
  - formal linguistic competence, that is knowledge of linguistic rules and patterns
  - functional competence, which roughly corresponds to inferential competence
    formal reasoning (logical reasoning and novel problem solving), world knowledge (knowledge of objects and events and their properties, participants and relations), situation modeling (the ability of building a representation of the stories), social reasoning (Theory of Mind)
- LLMs have an almost human-like formal competence, but still fall short of functional competence
  - cf. LLMs as "cultural technology" that only imitates human language production (Yiu et al. 2023; also termed as bibliotechnism by Lederman & Mahowald 2024)

Language Without Thought

Mahowald et al. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*

"good at language → good at thought" fallacy

If an entity (be it human or a machine) generates long coherent stretches of text, it must possess rich knowledge and reasoning capacities

- Mahowald et al. (2024) distinguish between:
  - formal linguistic competence, that is knowledge of linguistic rules and patterns
  - functional competence, which roughly corresponds to inferential competence
    formal reasoning (logical reasoning and novel problem solving), world knowledge (knowledge of objects and events and their properties, participants and relations), situation modeling (the ability of building a representation of the stories), social reasoning (Theory of Mind)
- LLMs have an almost human-like formal competence, but still fall short of functional competence
  - cf. LLMs as "cultural technology" that only imitates human language production (Yiu et al. 2023; also termed as bibliotechnism by Lederman & Mahowald 2024)

Is There Any Sparkle of ToM in LLMs?

The alleged ToM of LLM

- Do LLMs represent information in terms of the content of different propositional attitudes?
  - UNLIKELY!
- Do LLMs represent the fact that agents have recursive propositional attitudes and may have different propositional attitudes about the same information content?
  - UNLIKELY!
- Do LLMs reason and draw inferences based on the representation of their own and other agents' mental states?
  - NO!
- Do LLMs use their representation of mental states and inferences about other agents' mental states to generate and interpret utterances?
  - NO!

Is There Any Sparkle of ToM in LLMs?

The alleged ToM of LLM

- Do LLMs represent information in terms of the content of different propositional attitudes?
  - UNLIKELY!
- Do LLMs represent the fact that agents have recursive propositional attitudes and may have different propositional attitudes about the same information content?
  - UNLIKELY!
- Do LLMs reason and draw inferences based on the representation of their own and other agents' mental states?
  - NO!
- Do LLMs use their representation of mental states and inferences about other agents' mental states to generate and interpret utterances?
  - NO!

Is There Any Sparkle of ToM in LLMs?

The alleged ToM of LLM

- Do LLMs represent information in terms of the content of different propositional attitudes?
  - UNLIKELY!
- Do LLMs represent the fact that agents have recursive propositional attitudes and may have different propositional attitudes about the same information content?
  - UNLIKELY!
- Do LLMs reason and draw inferences based on the representation of their own and other agents' mental states?
  - NO!
- Do LLMs use their representation of mental states and inferences about other agents' mental states to generate and interpret utterances?
  - NO!

Is There Any Sparkle of ToM in LLMs?

The alleged ToM of LLM

- Do LLMs represent information in terms of the content of different propositional attitudes?
  - UNLIKELY!
- Do LLMs represent the fact that agents have recursive propositional attitudes and may have different propositional attitudes about the same information content?
  - UNLIKELY!
- Do LLMs reason and draw inferences based on the representation of their own and other agents' mental states?
  - NO!
- Do LLMs use their representation of mental states and inferences about other agents' mental states to generate and interpret utterances?
  - NO!

The Role of LLMs in Cognitive and Linguistic Research

- The "magic" of LLMs is simply the "magic" of distributional learning
- The real scientific revelation brought by LLMs is that the range of semantic and pragmatic aspects that language encodes and can be recovered from distributional statistics is far greater than we could have ever imagined before (at least if we have enough amount of data)
- LLMs can be used to understand which aspects of language processing might be solved with shallow surface cues only, without explicit "mentalizing".
  - humans too often behave like "stochastic parrots"!
- They still lack adequate representational structures of propositional attitudes that are crucial for ToM, situation modeling and inference

The Role of LLMs in Cognitive and Linguistic Research

- The "magic" of LLMs is simply the "magic" of distributional learning
- The real scientific revelation brought by LLMs is that the range of semantic and pragmatic aspects that language encodes and can be recovered from distributional statistics is far greater than we could have ever imagined before (at least if we have enough amount of data)
- LLMs can be used to understand which aspects of language processing might be solved with shallow surface cues only, without explicit "mentalizing".
  - humans too often behave like "stochastic parrots"!
- They still lack adequate representational structures of propositional attitudes that are crucial for ToM, situation modeling and inference

The Role of LLMs in Cognitive and Linguistic Research

- The "magic" of LLMs is simply the "magic" of distributional learning
- The real scientific revelation brought by LLMs is that the range of semantic and pragmatic aspects that language encodes and can be recovered from distributional statistics is far greater than we could have ever imagined before (at least if we have enough amount of data)
- LLMs can be used to understand which aspects of language processing might be solved with shallow surface cues only, without explicit "mentalizing".
  - humans too often behave like "stochastic parrots"!
- They still lack adequate representational structures of propositional attitudes that are crucial for ToM, situation modeling and inference

The Role of LLMs in Cognitive and Linguistic Research

- The "magic" of LLMs is simply the "magic" of distributional learning
- The real scientific revelation brought by LLMs is that the range of semantic and pragmatic aspects that language encodes and can be recovered from distributional statistics is far greater than we could have ever imagined before (at least if we have enough amount of data)
- LLMs can be used to understand which aspects of language processing might be solved with shallow surface cues only, without explicit "mentalizing".
  - humans too often behave like "stochastic parrots"!
- They still lack adequate representational structures of propositional attitudes that are crucial for ToM, situation modeling and inference

Grazie!!!

Merci!!!

Thank you!!!