Poster presented at the ToM4AI (Theory of Mind for Artificial Intelligence) Workshop, AAAI '2025, Association for the Advancement of Artificial Intelligence, March 2, 2025, Philadelphia, PA

Projecting Characters' Knowledge from Utterances in Narratives: A Psycholinguistic Baseline for LLMs

ADIL SOUBKI[1], AMIE PAIGE[3], JOHN MURZAKU[1], OWEN RAMBOW[12], & SUSAN E. BRENNAN[3]

[1]Department of Computer Science; [2]Department of Linguistics; [3]Department of Psychology

(For more information: asoubki@cs.stonybrook.edu, susan.brennan@stonybrook.edu)

Background

When reading narratives, human readers rely on their Theory of Mind (ToM) to infer not only *what the characters know* from their utterances, but also whether characters are likely to share common ground. As in human conversation, such decisions are not infallible but probabilistic, based on the evidence available in the narrative.

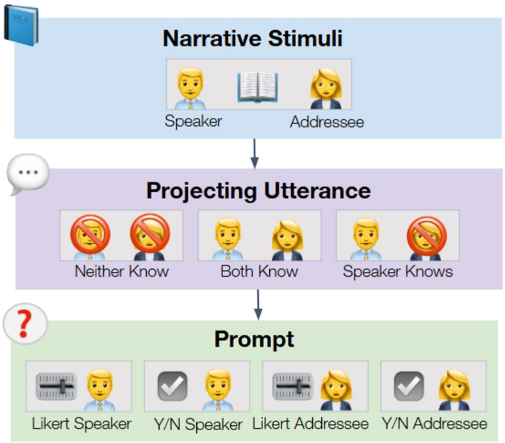
By responding on a scale (rather than Yes/No), humans can indicate commitment to their inferences about what characters know (ToM). We use two prompting approaches to explore (i) how well LLM judgments align with human judgments, and (ii) how well LLMs infer the author's intent from utterances intended to project knowledge in narratives.

Motivation

Can LLMs use utterances in narratives to project characters' knowledge, as the authors intended? These models do surprisingly well at making some but not all kinds of ToM-related inferences (see, e.g., Jones et al., 2024; Kim et al., 2023). LLM benchmarks tend to treat ToM inferences as a binary phenomenon; but humans make ToM inferences in a more nuanced way.

Narrative Stimuli & Experimental Design

Five psycholinguistics experiments by Gerrig et al. (2001) systematically probed readers' ToM-related pragmatic inferences from utterances that projected knowledge for neither character, both (common ground), or just the speaker.



EXAMPLE NARRATIVE (from Gerrig et al. (2001)

Cathy and Bob were flying from New York to Chicago for a good friend's wedding. This was the first time they'd ever flown first class.

The flight was also special for the flight attendant, Maureen. It was her last flight before she retired, after 40 years with the airline.

Bob said to Cathy ***[one of these three]***

- "Did you remember to order our special meals?" *(projects knowledge to neither)*
- "Can you imagine what flying was like 40 years ago?" *(projects knowledge to both)*
- "While you were in the restroom, I had an interesting conversation about what flying was like 40 years ago." *(projects knowledge to speaker, Bob)*

[Bob/Cathy] knows that the flight attendant is retiring. Do you agree? **(A) NO (B) YES**

How much do you agree? ***(Likert scale)***
(1) Completely ------------------------ (9) Completely
disagree **agree**

Gerrig et al.'s (2021) studies demonstrated that utterances in narratives trigger readers' inferences about characters' knowledge (see *Human* in Tables 1 and 2, below). **But how well do LLMs do with such ToM inferences?**

LLM Prompting Method

We used Gerrig et al.'s (N=20) narratives to test several classes of open and closed LLMs, including larger models (e.g., 70B parameters), smaller ones (e.g., <=8B), reasoning-distilled models (e.g., DeepSeek's R1-distilled and OpenAI's o1 models), and base models (without R1). For each of the 20 narratives, each model was presented in separate contexts/windows with each of 6 versions of the entire narrative with projecting utterance (Projects-Neither, Projects-Both, Projects-Speaker) X knowledge test (Speaker, Addressee). This prompting was done for both Yes/No and Likert scale tests.

Results

Likert-Scale Prompt (1-9)

Note: Shading indicates agreement that a particular character likely knows (low/high).

			Neither		Both		Speaker	
Model	MAD	Mean	Addressee	Speaker	Addressee	Speaker	Addressee	Speaker
Human*	0.0	4.8	3.1	3.0	5.8	6.3	4.6	6.2
o1-2024-12-17	0.4	4.4	2.9	2.6	5.2	6.1	3.6	6.0
Deepseek-R1-Llama-70B	0.6	5.0	3.5	4.1	5.0	6.1	4.2	6.8
GPT-4o	0.6	4.9	4.0	4.0	4.8	6.0	4.4	6.1
Llama-3.3-70B	0.8	4.5	3.5	3.9	4.0	5.0	4.5	6.0
Deepseek-R1-Llama-8B	1.2	4.7	4.5	4.7	4.0	5.2	4.0	5.7
Llama-3.2-1B	1.2	5.3	4.8	5.8	5.7	5.0	5.2	5.4
GPT-3.5-turbo	1.7	6.6	5.6	5.8	6.5	7.5	6.5	7.7
Llama-3.1-8B	1.8	6.6	6.7	6.7	5.9	6.8	6.8	7.0
Llama-3.2-3B	1.9	6.7	6.3	7.0	6.8	7.2	6.5	6.7

Table 1. Model Performance on the Likert-Scale Prompt. *Mean Average Difference* (MAD) denotes the extent to which each LLM diverged from human performance (abs value) for each condition.

Yes-No Prompt

Note: Shading indicates proportion of YES answers that a character

			Neither		Both		Speaker	
Model	MAD	Mean	Addressee	Speaker	Addressee	Speaker	Addressee	Speaker
Human*	0.0	65.3	22.4	21.4	59.2	70.6	39.0	70.4
Deepseek-R1-Llama-70B	10.8	40.0	20.0	25.0	35.0	60.0	35.0	65.0
Llama-3.3-70B	11.1	43.3	30.0	30.0	35.0	65.0	35.0	65.0
Llama-3.1-8B	19.6	33.3	25.0	30.0	20.0	55.0	20.0	50.0
GPT-4o	21.4	28.3	10.0	10.0	35.0	50.0	15.0	50.0
Deepseek-R1-Llama-8B	25.0	25.0	15.0	10.0	25.0	20.0	40.0	40.0
o1-2024-12-17	26.4	23.3	0.0	5.0	35.0	45.0	5.0	50.0
Llama-3.2-1B	35.6	14.2	5.0	10.0	20.0	15.0	25.0	10.0
Llama-3.2-3B	46.9	96.7	85.0	100.0	95.0	100.0	100.0	100.0
GPT-3.5-turbo	48.6	98.3	95.0	95.0	100.0	100.0	100.0	100.0
Author Intent	-	-	*Low*	*Low*	*High*	*High*	*Low*	*High*

Table 2. Model Performance on the Yes-No Prompt. In addition to Mean Average Difference (MAD), *Author Intent* denotes the expectations of the narratives' author (the experimenter) for each condition.

Discussion

- Results vary with prompting method!
- *No* models perform particularly well at correctly attributing CG to characters (Y/N). Some match author intent better than others (o1 and GPT-4o).
- Smaller models do not attribute knowledge as the author intended, nor as humans do.
- R1-distilled Llama-70B performs well, and strikingly similarly to humans.
- In future work, we will compare LLM performance to a new human baseline with identical prompting and significance testing.
- Future question: Does reinforcement learning lead to ToM that is more human-like?

***** *Human baseline is from Gerrig et al. (2001). Table 1 (Likert) is from GBO's Table 5, Expt. 3 and Table 4, Expt 2. Table 2 (Y/N) is from GBO's Tables 6 and 7.*

References

Gerrig, R. J., Brennan, S. E., & Ohaeri, J. O. (2001). What characters know: Projected knowledge and projected co-presence. *Journal of Memory & Language, 44*(1), 81-95.

Kim, H., Sclar, M., Zhou, X., Le Bras, R., Kim, G., Choi, Y. Choi, & Sap, M. (2023). FANTOM: A Benchmark for stress-testing machine theory of mind in Interactions. *Proc. 2023 Conf. on Empirical Methods in NLP*, 14397–14413.

Jones, C. R., Trott, S., & Bergen, B. (2024). Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (EPITOME). *Transactions of the Association of Computational Linguistics.*

ACKNOWLEDGMENT: This material is based upon work supported by NSF under Grant # NRT-HDR 2125295. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.