Testing Theory of Mind in Large Language Models and Humans

James W. A. Strachan[1], Dalila Albergo[2,3], Giulia Borghini[2], Oriana Pansardi[1,2,4], Eugenio Scaliti[5,6,1,2], Saurabh Gupta[7], Krati Saxena[7], Alessandro Rufo[7], Stefano Panzeri[8], Guido Manzi[7], Michael S.A. Graziano[9], Cristina Becchio[1,2]

► Introduction

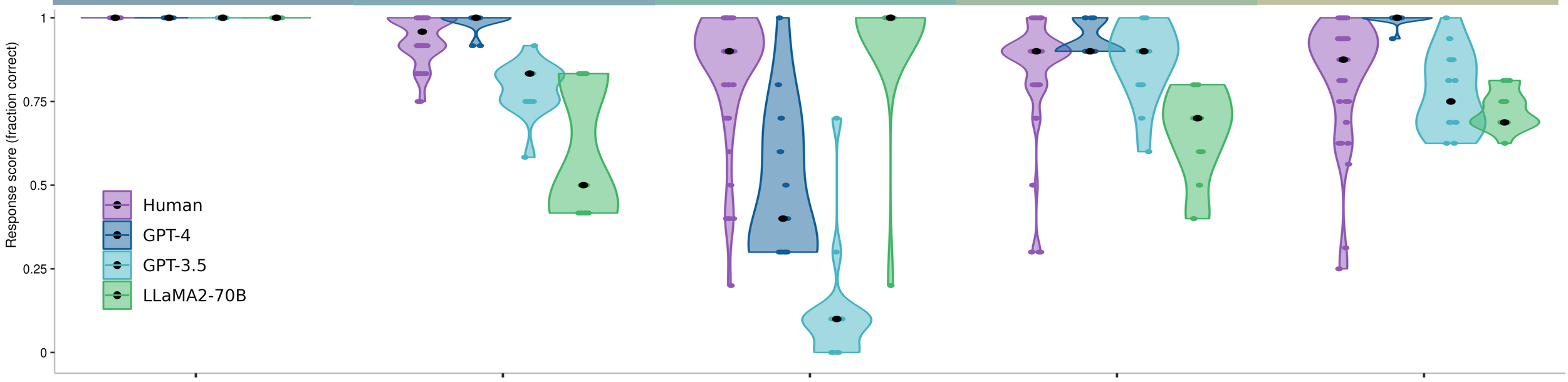
- Theory of Mind (ToM): The ability to reason about and predict others' mental states
- Mixed evidence that capacity may have emerged in LLMs [1,2,3]
- LLM evaluations tend to rely on individual tests of ToM with variants that are untested in humans [4]
- Requires systematic comparative approach under controlled conditions to evaluate how and when performance differs from humans

► Method

- Humans: N=50 for each set of questions
- LLMs: n=15 independent observations

- Each test administered separately
- Scored by researchers using validated criteria
- Second coding to ensure high scoring reliability

False Belief	Irony	Faux Pas	Hinting	Strange Stories
Anna and Beth are going to a concert together. Anna has the tickets, and she puts them in the side pocket of her purse and then leaves her purse with Beth while she goes to the bathroom. While Anna is gone, Beth gets the tickets out to check their seats and then puts them back in the front pocket of Anna's purse.	The politician had taken his assistant along to his conference; there were almost no other attendees there. 'Clearly people want to hear you speak', mused the assistant.	Jim was happy because his parents just opened a new restaurant near where he worked. One day, at work, he decided to invite his colleagues to the restaurant. His colleague Tom said, "I'd love to go out to dinner, but I strongly advise against that restaurant, I went the other night, and the food was terrible."	Paula's phone battery is running low. Her colleague Wendy has a phone charger on her desk that she is not currently using. Paula says to Wendy, "My phone's about to die."	Mary's mother is very strict and always scolds her when she does something careless. Mary and her close friend Simon are eating ice cream together when Mary accidentally drops the ice cream on her backpack, which makes a big stain. When Mary's mother arrives, she asks Mary sternly what happened to her backpack. Simon says that he accidentally dropped his ice cream on Mary's backpack.
When they go to the venue, where will Anna check for the tickets?	*Did the assistant think people want to hear the politician speak?*	*Did Tom know that the new restaurant belonged to Jim's parents?*	*What does Paula really mean when she says this?*	*Why did Simon say that?*

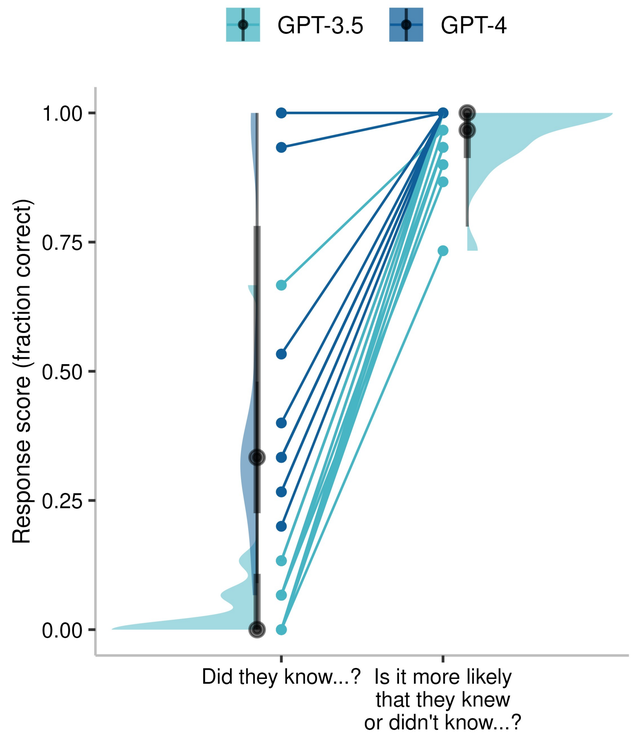


► Faux Pas Likelihood

- GPT-4 consistently at or above human levels; LLaMA2 poorer; except Faux Pas
- Example GPT-4 response: *"It is not clear from the story whether Tom knew that the new restaurant belonged to Jim's parents or not."*

Are GPT models unable to infer likelihood of explanations?

Is it more likely that Tom knew or did not know that the new restaurant belonged to Jim's parents?



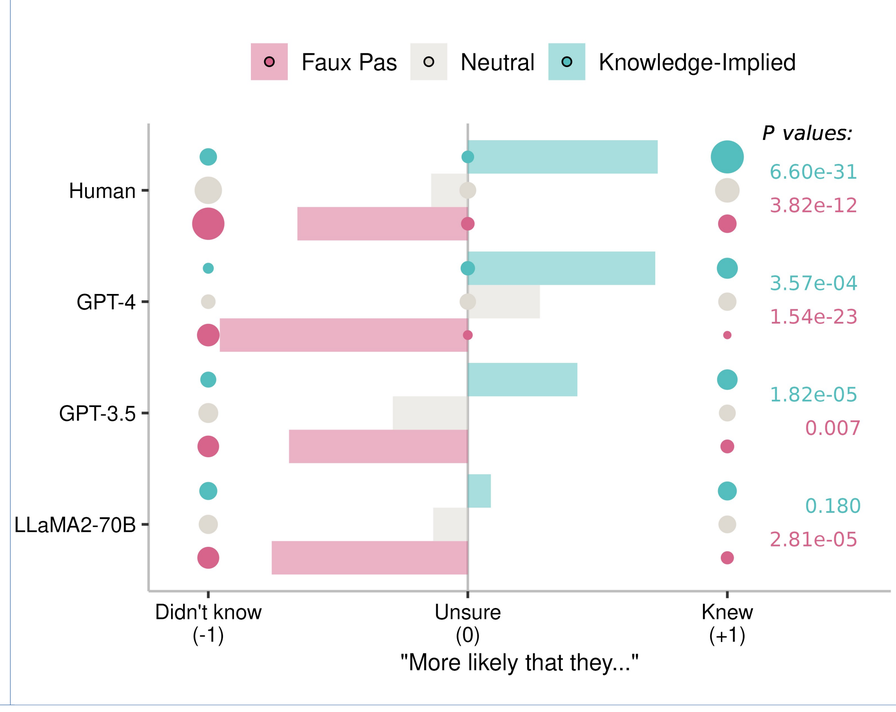
► Belief Likelihood Test

How sensitive are LLMs (and humans) to the implied belief content of stories?

Jim was happy because his parents just opened a new restaurant near where he worked. One day, at work, he decided to invite his colleagues to the restaurant. His colleague Tom said,

"I'd love to go out for dinner, but I strongly advise against that restaurant. I went the other night and the food was terrible."	"I'm sorry, but I can't go tonight. I'm going to the cinema with my sister"	"Oh, I'm not sure. I don't want your parents to assume we're dating"

Freq ○ 50 ○ 100 ○ 150 ○ 200 ○ 250



►Conclusion

	False Belief	Irony	Faux Pas	Hinting	Strange Stories
GPT-4	✓	✓	✗	✓	✓
GPT-3.5	✓	✓	✗	✓	✓
LLaMA2	✗	✗	✓	✗	✗

	Able to reason about likelihood	Sensitive to implied belief
GPT-4	✓	✓
GPT-3.5	✓	✓
LLaMA2	-	✗

- GPT-4 consistently at or above human levels; LLaMA2 poorer; except Faux Pas
- Example GPT-4 response: *"It is not clear from the story whether Tom knew that the new restaurant belonged to Jim's parents or not."*
- GPT models show human-like responses on a number of Theory of Mind tasks
- Specific impairments on recognising Faux Pas driven by a failure to commit to likeliest explanation
- Human-like sensitivity to implied belief does not translate to human-like behaviour

Affiliations
1) Department of Neurology, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany
2) Cognition, Motion and Neuroscience, Italian Institute of Technology, Genoa, Italy
3) Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy
4) Department of Psychology, University of Turin, Turin, Italy
5) Department of Management, "Valter Cantino", University of Turin, Turin, Italy
6) Human Science and Technologies, University of Turin, Turin, Italy
7) Alien Technology Transfer Ltd, London, UK
8) Institute for Neural Information Processing, Center for Molecular Neurobiology (ZMNH), University Medical Center Hamburg- Eppendorf (UKE), Hamburg, Germany
9) Princeton Neuroscience Institute, Princeton University, USA

References
1) Kosinski, M. Theory of Mind May Have Spontaneously Emerged in Large Language Models. Preprint at https://doi.org/10.48550/ARXIV.2302.02083 (2023).
2) Sap, M., LeBras, R., Fried, D. & Choi, Y. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. Preprint at http://arxiv.org/abs/2210.13312 (2023).
3) Kim, H., Sclar, M., Zhou, X., Bras, R. L., Kim, G., Choi, Y., & Sap, M. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. Preprint at https://doi.org/10.48550/arXiv.2310.15421 (2023).
4) Ullman, T. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. Preprint at http://arxiv.org/abs/2302.08399 (2023).

Scan for PDF:

Funding:
ASTOUND CONSCIOUS AI
UNTERSTÜTZT VON / SUPPORTED BY Alexander von HUMBOLDT STIFTUNG