Are explicit belief representations necessary?

A comparison between Large Language Models and Bayesian probabilistic models

UC San Diego

Dingyi Pan[1], Benjamin Bergen[2]

[1]Department of Linguistics, [2]Department of Cognitive Science

{dipan, bkbergen}@ucsd.edu

Clearly defined belief states might be necessary for belief attribution at least in the case of projection inferences, and the recursive reasoning between the interlocutors is crucial in pragmatic inference in general.

Background

Large Language Models (LLMs) show certain indirect pragmatic capabilities [1,2], although they lack explicit belief representations. It's unclear if they could also succeed in phenomena that directly require belief attributions.

Projection inferences: Inferences about speaker's commitment to the embedded content [3].

Scott says:
"John *knows* that Julian dances salsa."
"Does John *know* that Julian dances salsa?"
⇝*Scott* is certain that *Julian dances salsa*.

There are several factors that modulate projection inferences in humans, including the predicates [3,4], (not) at-issueness of the embedded clause [5-7] and speakers' prior knowledge [8,9].

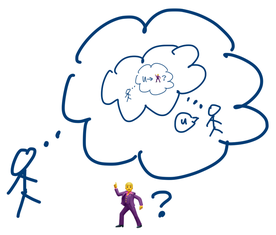
Research questions

Are explicit representations of mental states needed to model human pragmatic inferences?

- Are LLMs sensitive to factors that modulate human projection inferences?
- Do LLMs or Bayesian probabilistic models better capture the inference process in humans?

Bayesian model: Mixture RSA

- Propose in [10,11], the mixture RSA model is partially couched in the Rational Speech Act (RSA) framework [12,13].
- utterance set: $u \in U$ = {"know p", "know not p", "think p", "think not p", "BARE"}
- **Literal listener** reasons about the meaning of the utterance: the utterance is felicitous if the belief exceeds the threshold [14].
- **Pragmatic speaker** soft-maximizes the utility of the utterance, balancing the informativeness, as modeled in the *literal listener* model, and the costs.
- **Pragmatic listener** samples from either the prior belief distribution or the *speaker production distribution*, weighing by how likely the embedded content is at-issue, given the predicate.



Task 1: Prior knowledge

Prompt
Fact: Julian is Cuban./Julian is German.
Question: How likely is it that Julian dances salsa?

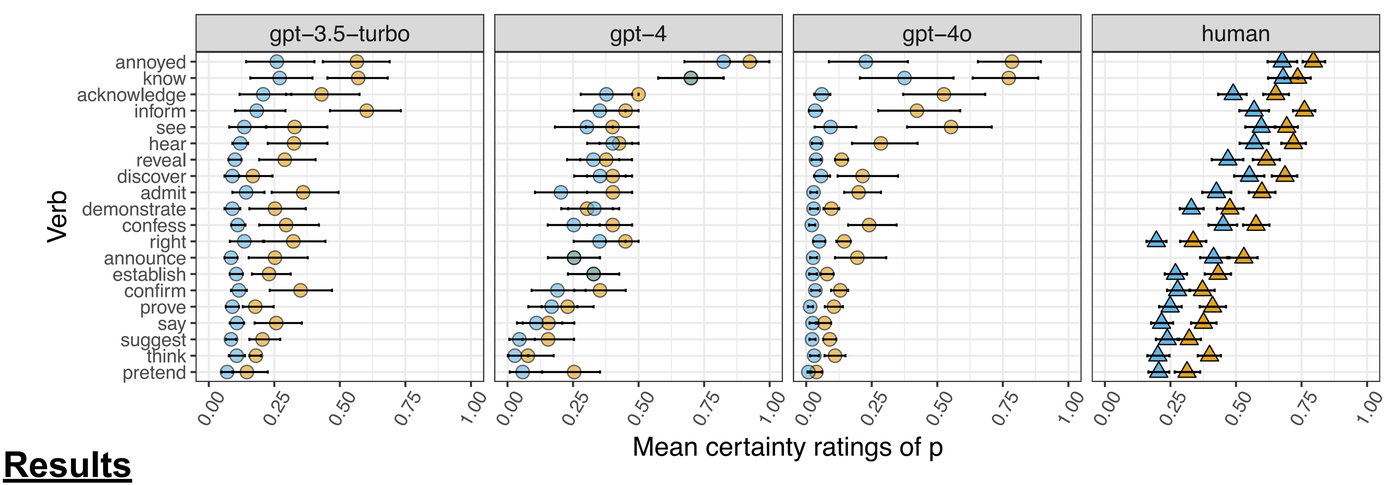
Task
Provide a rating between 0 and 1.

Results
Models capture world knowledge, such that each fact in the two prior conditions makes the content more or less likely a priori for LLMs, similar to humans.

Task 2: Projection inferences

Prompt
Fact: Julian is Cuban./Julian is German.
Sentence: Paul asks: Does John know that Julian dances salsa?
Question: Is Paul certain that Julian dances salsa?

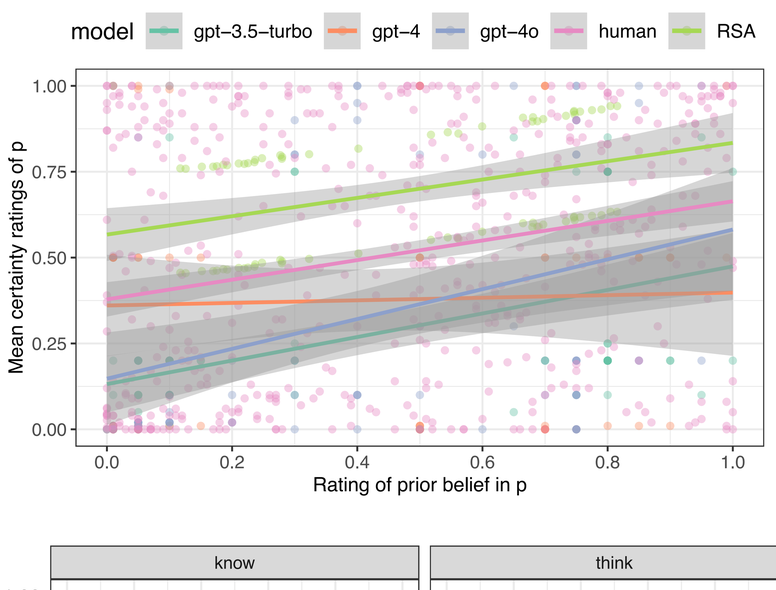
Task
Provide a rating between 0 and 1.



Results

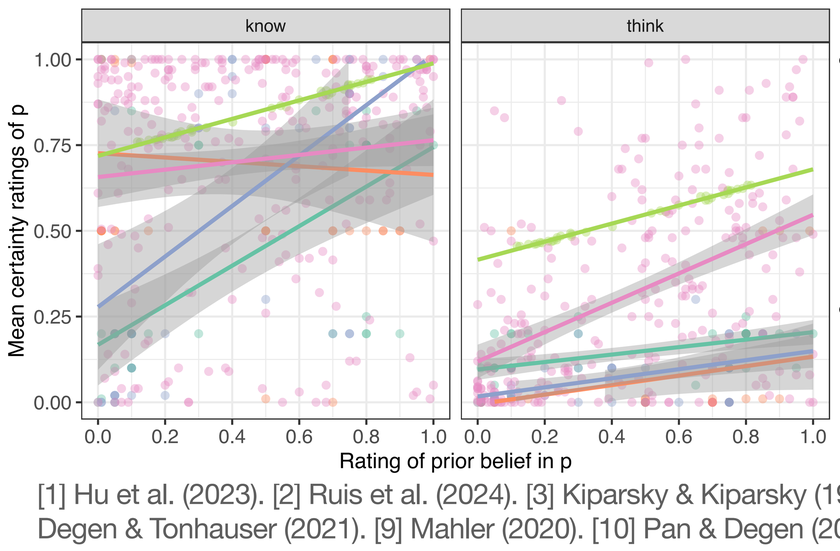
- GPT-3.5-turbo and GPT-4o show the effect of prior on certainty ratings, but it is mostly driven by the uniformly low ratings in the low prior condition.
- GPT-4 shows a smaller effect of prior on certainty ratings, and the effect varies across verbs.

LLMs and RSA vs. human results



RSA/LLM base: human ~ RSA/LLM + (1|participant) + (1|item)
full model: human ~ RSA + LLM + (1|participant) + (1|item)

Results
AIC of RSA and LLM base models: RSA (221.56) < GPT-4 (309.29) < GPT-4o (341.99) < GPT-3.5-turbo (380.44)
→ RSA better fits the human data.
LLM base vs. full: having each of the LLM predictions does **not** significantly improve the model fit
GPT-4 (χ^2=1.46), GPT-4o (χ^2=0.02) , GPT-3.5-turbo (χ^2=0.21)
→ LLMs do not capture additional variances in the human data in comparison to the RSA model.
RSA base vs. full: having RSA predictions as an additional predictor significantly improves the model fit
GPT-4 (χ^2=89.91), GPT-4o (χ^2=121.04) , GPT-3.5-turbo (χ^2=159.81)
→ there is variance that is not captured by the predictions of LLMs but is explained by the RSA model.



- For "know", GPT-4 most closely aligned with the human data, whereas the RSA model and GPT-4o and GPT-3.5-turbo models overestimate the effect of prior belief.
- For "think", all GPT models underestimate the effect of prior on certainty ratings, whereas the RSA model tracks the human data well.

Discussion

- These attested LLMs can capture the world knowledge and are sensitive to factors that affect projection inferences in humans by various degrees, but they might use world knowledge in a more coarse-grained way and do not incorporate it into inference in the same way that humans do.
- There might be additional information or cognitive processes needed to be captured in projection inferences, beyond distributional information in the LLMs.

[1] Hu et al. (2023). [2] Ruis et al. (2024). [3] Kiparsky & Kiparsky (1970). [4] Degen & Tonhauser (2022). [5] Tonhauser et al. (2018). [6] Beaver et al. [2017]. [7] Tonhauser & Degen (under review). [8] Degen & Tonhauser (2021). [9] Mahler (2020). [10] Pan & Degen (2023). [11] Pan (2023). [12] Frank & Goodman (2012). [13] Goodman & Frank (2016). [14] Lorson et al. (2023)