Are explicit belief representations necessary? A comparison between Large Language Models and Bayesian probabilistic models

Dingyi Pan
Department of Linguistics
University of California, San Diego
dipan@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
University of California, San Diego
bkbergen@ucsd.edu

Abstract

Large language models (LLMs) have exhibited certain indirect pragmatic capabilities, including interpreting indirect requests and non-literal meanings. Yet, it is unclear whether the success of LLMs on pragmatic tasks generalizes to phenomena that directly probe inferences about the beliefs of others. Indeed, LLMs' performance on Theory of Mind (ToM) tasks is mixed. To date, the most successful computationally explicit approach to making inferences about others' beliefs is the Rational Speech Act (RSA) framework, a Bayesian probabilistic model that encodes explicit representations of beliefs. In the present study, we ask whether LLMs outperform RSA in predicting human belief inferences, even though they do not explicitly encode belief representations. We focus specifically on projection inferences, a type of inference that directly probes belief attribution. We find that some LLMs are sensitive to factors that affect the inference process similarly to humans, yet there remains variance in human behavior not fully captured by LLMs. The RSA model, on the other hand, outperforms LLMs in capturing the variances in human data, suggesting that explicit belief representations might be necessary to construct human-like projection inferences.

1 Introduction

What is required to make inferences about the beliefs of others? Is it necessary to explicitly encode beliefs about others' beliefs? Or does statistical learning over language produce equally accurate predictions? These questions are as relevant to human language comprehension as they are to Natural Language Understanding. Large Language Models (LLMs) serve as an important test case for statistical learning approaches. Although they lack explicit belief state representations, larger and more recent language models appear to perform well in certain pragmatic reasoning and non-literal language understanding tasks (Hu et al., 2023a; Ruis et al., 2024). Yet, recent studies have produced conflicting results in tasks that involve the beliefs of others, including those testing their Theory of Mind (ToM) abilities (Ullman, 2023; Sap et al., 2022; Kosinski, 2024). On the one hand, conflicting results in pragmatic tasks and ToM tasks suggest that it is possible to accomplish pragmatic reasoning through low-level processes that do not require ToM-like reasoning about the interlocutor's mental state. On the other hand, it is possible that explicit belief representations emerge from exposure to language and are then used to perform pragmatic tasks (Hu et al., 2023a).

Compared to LLMs, where the inference capacities and processes are relatively more controversial, it is widely recognized that human comprehenders infer both the state of the world and other people's mental states from language. For instance, when the speaker (i.e., Paul), utters the sentence in Example (1a), the listener can infer that the attitude holder (i.e., John), believes the embedded proposition to be true, i.e., *it is raining*, since it is asserted by the sentence. More importantly, the listener can also infer that it is in fact raining and the speaker also believes so. This inference persists even when the sentence is in an interrogative form as in Example (1b). In this case, what is questioned is what the attitude holder believes, yet the speaker is taken to be committed to the truth of the embedded proposition. Hence, this inference is considered to project through the entailment-canceling environment (Kiparsky and Kiparsky, 1970). This type of inferences about speaker commitment is commonly referred to as PROJECTION INFERENCES.

1. (a) Paul said: John knows that it is raining.
   (b) Paul asked: Does John know that it is raining?

All data, materials, and analysis scripts can be accessed at https://github.com/pennydy/llm_belief

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11483–11498

For humans, pragmatic inferences often rely on Theory of Mind (ToM) abilities (Apperly, 2010). In particular, listeners infer the intended meaning of the speaker by reasoning about the speaker's beliefs and communicative goals (Grice, 1975). This reasoning process has been effectively modeled using the Rational Speech Act (RSA) framework (Frank and Goodman, 2012; Goodman and Frank, 2016; Degen, 2023), which represents the inference process as updating explicit beliefs between a speaker and a listener in a conversation. Among its many applications, RSA models qualitatively reproduce prior human findings in projection inferences through the effects of prior knowledge, semantics, and the at-issueness of the predicate (Pan and Degen, 2023; Pan, 2023). Nonetheless, the RSA framework is expressed at the computational level in Marr's level of analysis (Marr, 1982), where it includes explicit computational representations of beliefs. It remains unclear whether this explicit belief representation is required during pragmatic inference to account for belief attribution during projection inferences.

In contrast to Bayesian accounts, LLMs do not explicitly encode belief representations, which allows them to serve as a test case to gauge the necessity of explicit belief representations in pragmatic inference. Various studies have found that language models are able to capture the systematic variability in human data in scalar implicature (Hu et al., 2023b). The first goal of this study is thus to investigate whether LLMs are sensitive to the above-mentioned factors in projection inferences, as RSA is.

Assuming that LLMs capture some variability in human pragmatic inferences, a second question compares this predictive power to that of Bayesian, RSA-based accounts. Recent studies have evaluated the performance of language models to RSA, which suggest that LLMs do not behave like pragmatic speakers (Jian and Siddharth, 2024). Yet, on the comprehension side, Carenini et al. (2023) show that the predictions of GPT2-XL closely resemble and can be simulated by a pragmatic listener in an RSA framework on the task of interpreting metaphors. Motivated by this approach, and in light of the fact that comparisons between LLMs and RSA on other pragmatic inference tasks are still lacking, we directly compare the results from LLMs and RSA in terms of how well they predict human results in the case of projection inferences. Since RSA explicitly models belief states of the interlocutors while LLMs lack such representations, if LLMs can explain more variance in human behavior, this would disconfirm the fundamental assumption of RSA about the need of a recursive Bayesian process and shed light on the debate about whether belief representations are needed in pragmatic reasoning (Sperber and Wilson, 2002).

Hence, the goals of the current study are twofold: First, we evaluate the performance of LLMs on a particular pragmatic task and whether they are sensitive to factors that modulate human pragmatic inferences. Second, we compare whether Bayesian probabilistic models or LLMs better capture human performances, in order to provide insight into whether the explicit representation of mental states is needed to model human pragmatic inferences.

2 Related work

2.1 Projection inferences

Projection inferences are not a monolithic phenomenon but can be modulated by various factors. Results from experimental studies of human pragmatic inference suggest that the projectivity of the embedded content varies across predicates (Degen and Tonhauser, 2022), which supports the gradient view of factivity and contrasts with the categorical view that there is a clearly defined class of factive verbs that trigger the projection inference. In addition, the inference process is modulated by various factors, including the identity of the predicate (Kiparsky and Kiparsky, 1970), the at-issueness of the embedded content (Tonhauser et al., 2018; Stevens et al., 2017), prosodic focus (Djärv and Bacovcin, 2020), and prior beliefs about the likelihood of the embedded content (Mahler, 2020; Degen and Tonhauser, 2021; Lorson, 2021). In the current study, we focus primarily on the effects of predicates and prior beliefs. For instance, going back to Example 1, if the speaker, Paul, and the attitude holder, John, live in a city that rarely rains, it would be odd for Paul to say "John knows it is raining", and thus even though the factive verb "know" is used, Paul is less likely to be taken to believe that *it is raining*.

2.2 Large language models and pragmatics

Previous work has attempted to test the inference abilities of transformer-based models on inference tasks with various presupposition triggers. For instance, Jiang and de Marneffe (2021) investigate BERT's performance on event factivity, and their

results suggest that its strong performance on a few factuality datasets is due to the statistical regularities in the data instead of its pragmatic reasoning ability. Moreover, in the NOPE (Parrish et al., 2021) and PROPOSE (Asami and Sugawara, 2023) benchmarks, transformer-based models are evaluated on the accuracy of predicting the semantic relations between a sentence with a presupposition trigger and the presupposed content.

Although these benchmarks consider different sentence structures, such as negation and interrogatives, they do not explicitly control and test for the effect of world knowledge and the other modulating factors of projection inferences. In addition, projection inference is framed as a Natural Language Inference (NLI) task, where models are evaluated on the accuracy of predicting the label that categorizes the relationship between the sentence with a presupposition trigger and the projected content. As discussed in the previous section, projection inference is a general phenomenon that is not limited to factive verbs that presuppose the truth of the embedded content, and different clause-embedding verbs exhibit gradience in the projection inference patterns. Therefore, using a classification task is not sufficient in capturing the nuances in the projection pattern.

On the other hand, LLMs are able to succeed in certain tasks that involve pragmatic reasoning. For instance, results from (Ruis et al., 2024) suggest that instruction-tuned LLMs, including OpenAI's text-<engine>-001-series, ChatGPT, and GPT-4, demonstrate certain pragmatic reasoning abilities for implicature resolution. Especially when it is combined with few-shot and chain-of-thought prompting techniques, GPT-4 is able to achieve average human-level performance. Furthermore, Hu et al. (2023a) find that LLMs demonstrate pragmatic abilities, including understanding non-literal language in cases that involve explicit reasoning about the intent of the speaker, such as polite deceit and irony. The errors that LLMs make are similar to those that humans make, such that they lean towards literal interpretations of sentences, instead of other heuristics, such as word similarity. All LLMs tested are distributional learners trained on text, which lack explicit representations of the mental states of the interlocutors. It is thus unclear what existing results mean about whether mental state representation is needed for pragmatic inference. As mentioned in the Introduction, LLMs fail at certain ToM tasks, so it is possible that their success on pragmatic tasks might be due to low-level linguistic heuristics.

2.3 Bayesian models

One way to model the belief update between the interlocutors in a conversation is by explicitly modeling the inference process as in the Rational Speech Act (RSA) framework (Frank and Goodman, 2012; Goodman and Frank, 2016; Degen, 2023). Under this framework, language production and interpretation are formalized as recursive reasoning between speaker and listener. Specifically, upon observing an utterance, the *pragmatic listener* updates their prior beliefs about the world by reasoning about an utterance produced by a *pragmatic speaker*. Both interlocutors are assumed to be rational and soft-maximize the utility of utterances and their interpretations. In the probabilistic pragmatic literature, RSA has been used to model various pragmatic phenomena, such as scalar implicature (Bergen et al., 2012) non-literal interpretation (Kao et al., 2014), and polite speech (Yoon et al., 2020).

RSA has also been adapted to projection inference, where it models the effects of prior and at-issueness of the predicate on inference patterns, by taking the inferred speaker belief (b_{SP}) given the utterance (u) to model the projectivity of the content (Pan and Degen, 2023; Pan, 2023). Here, we briefly summarize the core structure of the mix-RSA model proposed in Pan (2023).

One key component in the RSA framework is the utterance space, where each utterance is considered to be an alternative to the other. In particular, this mix-RSA model focuses on the projection inference pattern of the factive verb "know" and the non-factive verb "think" in interrogatives, and each of the two verbs can be combined with the affirmative (*p*) and negated embedded clause (*not p*). In addition, the unembedded polar interrogative "p?" is also considered as a possible alternative. Taken together, the utterance space includes five possible utterances: "p?", "know p", "know not p", "think p", and "think not p".

The *literal listener* (L_0) reasons about the literal semantics of each utterance and has a uniform prior over the belief. The utterance is felicitous if the belief in *p* exceeds the threshold θ_u associated with the verb that is used, as defined in Equation 1.

$$P_{L_0}(b_{SP}|u) \propto \begin{cases} 1 & \text{if } b_{SP} > \theta_u \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

On the other hand, the *pragmatic speaker* (S_1) produces an utterance proportional to its utility determined by the optimality parameter α. This is defined in Equation 2.

$$P_{S_1}(u|b_{SP}) \propto \exp(\alpha \cdot U(u; b_{SP})) \quad (2)$$

The utility is defined as balancing the trade-off between the informativeness and the cost of that utterance (Eq. 3), where the informativeness is defined as the probability that a literal listener would correctly infer the intended meaning based on the lexical semantics of that utterance. Both the negation and the presence of an embedded clause are assumed to contribute to the costs of an utterance, formulated as (C_{Neg}) and (C_{Embed}).

$$U(u; b_{SP}) = \ln P_{L_0}(b_{SP}|u) - C(u), \quad (3)$$
$$\text{where } C(u) = C_{\text{Neg}}(u) + C_{\text{Embed}}(u)$$

At the *pragmatic listener* (L_1) level, projectivity is modeled as the degree of speaker belief given the utterance, $P(b_{SP}|u)$. Instead of updating prior beliefs based on their internal model of the speaker following Bayes' rule in the standard RSA framework, the pragmatic listener in this model combines the prior belief distribution and the inferred belief distribution based on the speaker production distribution, as shown in Equation 4. Specifically, the pragmatic listener probabilistically considers the expected speaker production distribution or defaults back to their prior beliefs, proportional to the at-issueness of the embedded content as determined by the predicate in the utterance, ($P(q_{CC}|u)$), and the non-at-issueness, ($P(q_{MC}|u)$), respectively.

$$P_{L_1}(b_{SP}|u) \propto \underbrace{P_{S_1}(u|b_{SP})}_{\text{speaker model}} \cdot P(q_{MC}|u)$$
$$+ \underbrace{P(b_{SP})}_{\text{prior belief}} \cdot P(q_{CC}|u), \quad (4)$$
$$\text{where } P(q_{MC}|u) + P(q_{CC}|u) = 1$$

The values of the three free parameters in the model (the optimality parameter α and the two cost terms) are estimated using Bayesian Data Analysis with the projection rating data from Pan and Degen (2023) and the prior rating data from Degen and Tonhauser (2021). The predictions of the model qualitatively capture the patterns in human data with respect to the effect of prior and the difference between predicates, especially in the case of "p?", "know p", and "think not p" (Pan, 2023).

In sum, with the explicitly defined inference process and alternative sets, RSA models serve as a state-of-the-art computational theory of inference. As such, they serve as a useful comparison with LLMs, which lack clearly defined belief representations but also seem to accomplish certain pragmatic tasks.

3 Methods

3.1 Models

We test three GPT models via the OpenAI API: GPT-3.5-turbo, GPT-4, and GPT-4o. These models are fine-tuned with instruction following and human feedback, and results from previous studies suggest that they demonstrate certain pragmatic reasoning abilities and are able to interpret language in context with further fine-tuning (Ruis et al., 2024). For the RSA model, we use the mix-RSA model mentioned in the previous section.

3.2 Materials and procedure

To ensure a fair comparison between human results and model responses, we adopted the setup used in human experiments as closely as possible by reusing the stimuli. Specifically, the materials and procedure mirror Experiment 1 in Degen and Tonhauser (2021). In the current study, each model is tested on two separate tasks: the first prior task elicits the prior belief rating of embedded content given the two facts, and the second projection task investigates the effect of prior belief on projection by embedding the content with different clause-embedding predicates.

In the prior task, 20 critical items from Degen and Tonhauser (2021) are used as the embedded contents. Each of the items is paired with a "high prior" fact and a "low prior" fact that makes the content more or less likely a priori, respectively. For instance, knowing that "Julian is Cuban" (a high prior fact) makes the embedded content *Julian dances salsa* relatively more likely than knowing that "Julian is German" (a low prior fact). The notions of "high" and "low" are used in a relative sense, and these labels were confirmed by and drawn from human judgments (Degen and Tonhauser, 2021).

Therefore, similar to the rating task for human participants, the models are prompted to provide a

rating from 0 to 1 to the question that probes the likelihood of the content with the carrier sentence "How likely is that ...?", as illustrated in the example below.

> **Fact**: Julian is German.
>
> **Question**: How likely is it that Julian dances salsa?

Then in the projection task, all items in the prior task are used as the embedded content of the 20 clause-embedding predicates, the same as those included in Degen and Tonhauser (2021). The sentence is presented as a question asked by a speaker, and each sentence is paired with a fact. Then we use the "certain that" diagnostic (Djärv and Bacovcin, 2020; Tonhauser et al., 2018, *inter alia*), where the model is asked to provide a rating from 0 to 1 to the question with the structure "Is SPEAKER certain that...?", as shown below, and the response is taken to be the degree of the ascribed speaker belief.[1]

> **Fact**: Julian is German.
>
> **Sentence**: Paul asks: Does John know that Julian dances salsa?
>
> **Question**: Is Paul certain that Julian dances salsa?

For both tasks, the model is instructed to predict a continuous rating, which differs from previous work where the task of the model was to predict the label of the relationship between the target sentence with the presupposition trigger and the presupposed content (Parrish et al., 2021; Asami and Sugawara, 2023). The present approach closely matches human experimental approaches and avoids the potential confounds of the lexical overlap between the target sentence and the presupposed content when predicting the NLI labels.

3.3 Prompt structure

The prompt consists of two parts: the system prompt, which introduces the tasks and instructs the model to provide a numerical number between 0 and 1, and the main prompt, which contains the critical experimental item as described in the previous section.

3.4 Analysis

To test whether models can capture the effect of prior in the first task, we fit a linear regression model predicting the elicited prior belief rating from the prior type with by-item random intercepts and prior type by-item slope. To test the effect of prior type on projection ratings in the second projection task, we fit another linear regression model to predict the elicited projection rating from a fixed effect of the prior type and the by-item random intercept and random slope for the prior type.

On the other hand, to compare the LLMs' predictions and the RSA predictions to the human behavioral results in Experiment 1 from Degen and Tonhauser (2021), we narrowed to the two verbs "think" and "know" since the RSA model considers them as alternatives in the utterance space and the model parameters are estimated with the human belief ratings for these two verbs. For the RSA model and each LLM, we first fit a linear mixed-effects regression model predicting the human certainty ratings from the model predictions and the by-item random intercept (the base models). We compared the fit of each model using the Akaike Information Criteria (AIC).

Then, to quantitatively evaluate how well each GPT model captures the human data in comparison to the RSA model, we fit another linear mixed-effects model with the RSA model predictions as an additional predictor to predict the human certainty ratings from the model predictions (the full models). Following analyses in Jones et al. (2023), we used a Chi-square (χ^2) test to compare the full model of each GPT model to the corresponding base model. If adding RSA prediction significantly improves the model fit, then it indicates that the RSA model captures additional variance in the human behavior that is not predicted by the LLM. On the flip side, to quantitatively measure whether RSA captures additional variance than each GPT model, we compared each of the full models to the base RSA model. If having the prediction of the LLMs as the additional factor improves the model fit, then LLMs capture additional variance in human judgments that are not explained by the RSA model. Structures of all statistical models and results are summarized in Appendix A.

[1] We also tested the models with the prompt "Does SPEAKER believe that ...?", which is used in Pan and Degen (2023); Pan (2023) to directly elicit belief ratings among human participants and calibrate the RSA predictions. With this belief rating prompt, GPT-4 does not capture the variance among predicates, but its belief ratings against the prior ratings are more in line with the human results. See Appendix B for the results and discussion.

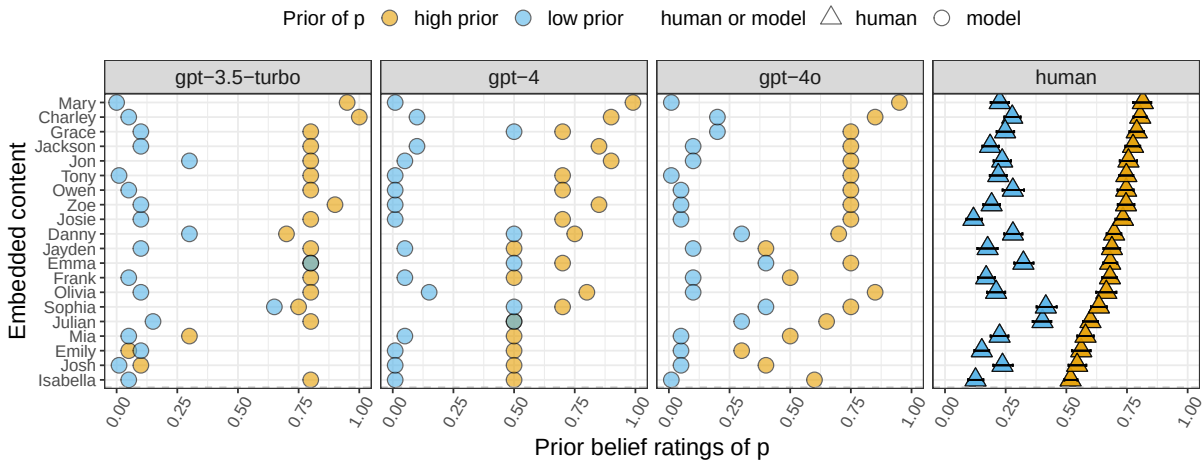


Figure 1: The mean prior likelihood of each embedded content, by models and humans. The contents and the human results are drawn from Degen and Tonhauser (2021). Each piece of content is labeled by a person's name as it describes a scenario or a feature of that person. Each dot in the three model facets represents the rating of items, each dot in the human facet presents the mean rating for each item, and the error bars represent bootstrapped 95% confidence intervals.

4 Results

4.1 Prior knowledge

Figure 1 shows the model and human predictions of the prior belief rating for each item. For all three models, the prior ratings in the high prior condition are higher than those in the low prior condition, although compared to human results, there is less variance in the ratings across items. This observation was borne out statistically: There is a significant main effect of prior type (GPT-3.5-turbo: $\beta = 0.56, t = 7.96, p < .001$; GPT-4: $\beta = 0.53, t = 9.03, p < .001$; GPT-4o: $\beta = 0.54, t = 12.67, p < .001$). This suggests that 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.

4.2 Effect of prior on projection inferences

Figure 2 shows the mean certainty ratings of the embedded content across 20 items in the two prior conditions by predicate. GPT-3.5-turbo and GPT-4o show the effect of prior on the projection ratings, such that the mean ratings in the high prior condition are higher than those in the low prior condition. The pattern in the predictions of GPT-4 is less clear, where the mean ratings for verbs like "know" are similar in the two prior conditions, while the mean ratings for "acknowledge" are relatively more distinct. These observations were borne out statistically in the linear mixed-effects model predicting the certainty ratings of each LLM from the fixed effect of prior type. The results show a significant main effect of prior type (GPT-3.5-turbo: $\beta = 0.18, t = 4.69, p < .001$; GPT-4:$\beta = 0.08, t = 2.20, p = .0403$; GPT-4o: $\beta = 0.20, t = 6.25, p < .001$), suggesting that all three models capture the differences between the two types of prior belief.

However, even though the results for GPT-4 were not significant given the statistical model, GPT-4 qualitatively captures the gradience of the ratings among predicates. On the other hand, although GPT-3.5-turbo and GPT-4o also capture the gradient patterns in the high prior condition, both seem to overestimate the effect of prior. In particular, except for the ratings for canonically factive verbs like "annoyed" and "know," the certainty ratings in the low prior condition are more uniform among predicates than those in the high prior condition, which suggests that the effect of the low prior might dominate the effect of the predicate during the inference.

Moreover, across the models, the difference between the certainty ratings in the two prior conditions was smaller for verbs like "acknowledge" and "inform" than for verbs like "pretend" and "think." For GPT-3.5-turbo and GPT-4o, this difference in magnitude seems to be driven by the uniformly low ratings of verbs in the low prior condition since the prior condition capture the by-predicate variances. Another possibility is that the "(optionally)

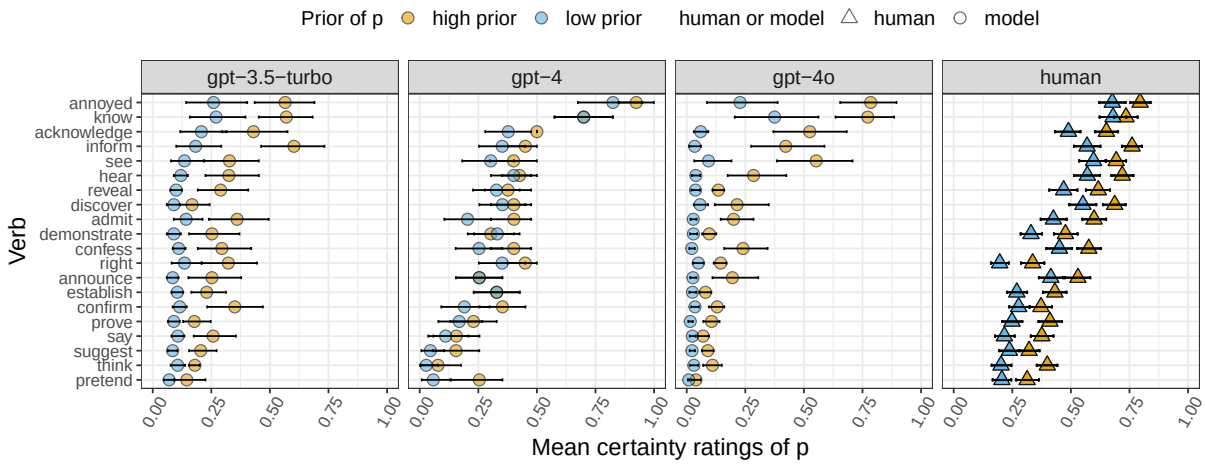


Figure 2: The mean certainty rating against prior ratings, by predicate and by model. The human results are drawn from Degen and Tonhauser (2021). Each dot represents the mean belief rating for each verb for models, and the grand mean across items and across participants for humans. Error bars represent the 95% confidence intervals.

factive" verbs are more robust to the effect of prior than "non-factive" ones. Yet, this observation does not apply to all verbs, such as "see," which is an "optionally factive," but have similar ratings in the two conditions. This demonstrates even LLMs that are trained on texts and can capture the statistical patterns in language do not show a clearly defined class of factive verbs.

4.3 LLMs and RSA compared to human results

Figure 3 shows the mean certainty ratings of GPT models, the RSA model, and the human results against mean prior ratings of the embedded content.[2] Qualitatively, both RSA and two LLMs, GPT-3.5-turbo and GPT-4o, capture the effect of prior on the certainty rating, where the speaker is considered to be more certain about the embedded content p when p is more likely a priori. Yet, none of these three models fully capture the human results: the predictions of the RSA model are higher than humans do, whereas both GPT-3.5-turbo and GPT-4o predict the certainty rating to be lower than humans. On the other hand, GPT-4 does not seem to capture the linear relationship between the prior belief and the certainty ratings. The observation was borne out statistically: the RSA base model has the lowest AIC value in comparison to the LLMs (RSA: $AIC = 221.56$; GPT-3.5-turbo: $AIC = 380.44$; GPT-4: $AIC = 309.29$; GPT-4o: $AIC = 341.99$), suggesting that the RSA model fits the human data better than the LLMs.

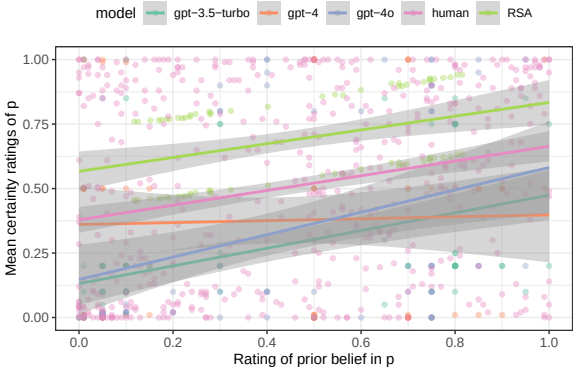


Figure 3: The mean certainty ratings against the prior rating of the embedded content of humans, RSA, and LLMs. Each dot represents the mean ratings of items for each verb, and the ribbons represent bootstrapped 95% confidence intervals.

As described in the Analysis section, for each LLM, we fit two linear mixed-effects models, one without the RSA predictions as a predictor (the base model) and one with the RSA predictions (the full model) to test whether the RSA model explains the variance in the human data that is not captured the LLMs. Across the regression models for all three LLMs, having RSA predictions as an additional predictor significantly improves the model fit (GPT-3.5-turbo: $\chi^2(2) = 159.81, p < .001$; GPT-4: $\chi^2(2) = 89.91, p < .001$; GPT-4o: $\chi^2(2) = 121.04, p < .001$), suggesting that there is variance that is not captured by the predictions of LLMs but is explained by the RSA model.

[2]For the linear relationship between certainty ratings and the prior ratings of each model for all 20 predicates, see Figure 6 in Appendix C.

Likewise, we compared the base model of RSA to each of the three full models that include the LLM predictions as a predictor to evaluate whether there is variance in the human judgment that is not modeled by RSA but is explained by the LLM. Overall, having each of the LLM predictions does not significantly improve the model fit (GPT-3.5-turbo: $\chi^2(2) = 0.21, p = 0.6393$; GPT-4: $\chi^2(2) = 1.46, p = 0.2274$; GPT-4o: $\chi^2(2) = 0.02, p = 0.8951$), suggesting that LLMs do not capture additional variances in the human data in comparison to the RSA model.

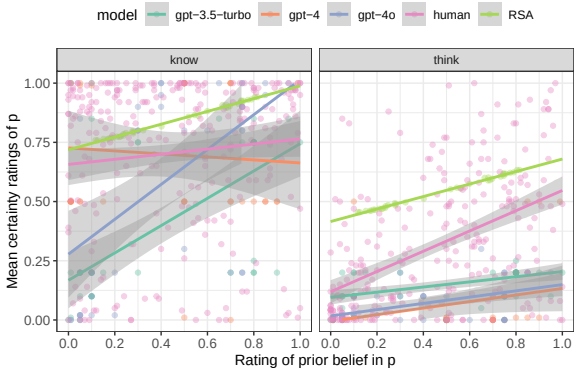


Figure 4: The mean certainty ratings for "know" and "think", against the prior rating of the embedded content. Each dot represents the mean ratings of items for each verb, and the ribbons represent bootstrapped 95% confidence intervals.

As an exploratory analysis, since studies with human participants suggest that the certainty ratings vary across verbs, we analyzed the results for "think" and "know" separately. Figure 4 shows the mean certainty ratings against the mean-centered prior belief ratings when p is embedded under each verb.[3] For "know", GPT-4 is most closely aligned with the human data, whereas the RSA model and the other two GPT models overestimate the effect of prior belief. We combined the certainty ratings from both humans and the four models and fit a linear mixed-effects model predicting the combined certainty ratings from the main effect of the model type (including four levels: human, RSA, GPT-3.5-turbo, GPT-4, GPT-4o; Reference level: human) and the mean-centered prior belief rating as well as the by-item random intercept. There is a main effect of prior belief ($\beta = 0.21, t = 4.44, p < .001$), such that the certainty ratings of p are higher when it is more likely a priori. In terms of the model type, the ratings of GPT-3.5-turbo are significantly lower than human results ($\beta = -0.28, t = -5.34, p < .001$), and the predictions of GPT-4o and the RSA model are marginally higher than the human results (GPT-4o: $\beta = -0.12, t = -2.29, p = .0226$; RSA: $\beta = 0.14, t = 2.46, p = .0145$). Yet, there is no significant difference between the ratings of GPT-4 and human results ($\beta = 0.00, t = 0.01, p = .9871$).

On the other hand, for "think", all GPT models underestimate the effect of prior on certainty ratings, whereas the RSA model tracks the human data well. We fit another linear mixed-effects model as the one for "know", and the results show that there is a significant main effect of prior ratings ($\beta = 0.33, t = 10.29, p < .001$). Crucially, the ratings of each of the LLMs are significantly lower than the human results (GPT-3.5-turbo: $\beta = -0.16, t = -4.58, p < .001$; GPT-4: $\beta = -0.25, t = -6.97, p < .001$; GPT-4o: $\beta = -0.22, t = -6.30, p < .001$), whereas the predictions of the RSA are significantly higher than the human results ($\beta = 0.22, t = 6.01, p < .001$).

5 Discussion

This study evaluated the performance of three LLMs and one RSA model on projection inferences to determine how well each predicts previously reported human data. First, we found that LLMs can capture the world knowledge that either makes the embedded content more or less likely a priori. This type of world knowledge serves as the prior belief that affects projection inference of the embedded content in humans. In addition, we also showed that these attested LLMs are sensitive to factors that affect projection inferences in humans by various degrees. Specifically, all three models are sensitive to the effect of prior on projection inferences, similar to humans. GPT-4 shows the gradient projection patterns across predicates as humans do. Nonetheless, although GPT-3.5-turbo and GPT-4o capture the gradience in the high prior condition, both overestimate the effect of the low prior, where they show little variance among predicates. It is possible that although these models show the effect of world knowledge, they do so in a more coarse-grained way and do not incorporate it into inference in the same way that humans do.

In addition, for both prior and projection tasks,

[3]We also tested using the categorical prior type distinctions (i.e., "high" vs. "low" prior), and the same result patterns still hold. The results of "know" and "think" are summarized in Tables 7 and 8 in Appendix A.5, respectively.

some variance in the human data cannot be fully explained by the LLMs predictions, suggesting that there might be additional information or cognitive processes needed to capture world knowledge and projection inference, beyond distributional information and fine-tuning from prompting and human feedback as in the LLMs.

Specifically, in terms of their abilities to approximate human judgments in projection inferences between these two types of computational models, RSA outperforms the three attested LLMs as measured by AIC. The results are more nuanced when we analyze the results of each verb individually. For "know", GPT-4 closely matches the projection inference patterns in human results, while the other models overestimate the effect of prior and either predict higher certainty ratings (RSA) or lower certainty ratings (GPT-3.5-turbo and GPT-4o) than the human results. Yet, both LLMs and RSA models fail to capture the projection inference patterns of "think" in the human data, where all three models underestimated the effect of prior and RSA overestimated it. This might be because "think" is used more frequently and pervasively across different scenarios (Pan and Degen, 2023). For instance, "John thinks Julian dances salsa" might be interpreted as an indirect answer to the question "Who should we invite to give a performance", where "think" has a parenthetical reading and does not contribute to the meaning of the utterance. Nonetheless, the same sentence can be used to express incredulity about John's belief if everyone in the conversation knows that Julian doesn't dance salsa.

Moreover, the likelihood ratio test reveals that adding RSA predictions improved predictions of human judgments over LLM predictions alone. Thus, there is some variance in the human data that is not fully captured by LLMs. On the flip side, having the LLM predictions does not improve the fit of the RSA predictions on the human data. This suggests that models with explicit belief representations are able to more accurately mirror human pragmatic inference. Contrary to the findings that GPT2-XL behaves similarly to the pragmatic listener in the RSA model for metaphor interpretations (Carenini et al., 2023), LLMs in the present study fail to capture some of the factors that affect projection inferences and cannot explain nuances in the human data.

Lastly, although we did not directly test Theroy of Mind (ToM) in this study, the results seem to suggest reasoning about other people's belief in a ToM-like way is needed at least in the case of inferring the interlocutors' belief. Even if language models can approximate the communicative intents of the interlocutors by only having accessing texts and can have partial representations of beliefs to guide the generation of subsequent texts, as argued in Andreas (2022), it is not enough to fully predict human pragmatic inference abilities. Explicit belief representations might be needed to infer the beliefs of others.

Taken together, 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. Granted, it is possible that projection inference is complex and requires additional epistemic reasoning not only about the speaker but also about the person whose belief is being reported, which might make the belief representation prominent, and future studies can adopt a similar methodology to compare different types of models and investigate the need for belief representation in other pragmatic phenomena.

Limitations

The conclusions are contingent on the structure of this particular mix-RSA model. It is possible that there might exist more optional RSA models to capture projection inferences, and these might more closely match the human performance than the LLMs.

In addition, the current study used prompting as a way to elicit the model response given that it is not possible to obtain the probability distribution of these GPT models. However, results from prompting do not always align with the raw log probabilities, especially in complex tasks that are less similar to next-word predictions Hu and Levy (2023). Therefore, as pointed out by one reviewer, the results are limited to the choice of prompting, and future studies should explore other open-source models for a more direct comparison between RSA and LLMs in terms of their predictive power.

Another limitation of the study is that both the human results and the model predictions of LLMs and RSA are English only. Thus, future studies should investigate what projection inferences are like in other languages, especially in those that have complex evidential markings or different types of verbs, and whether LLMs can capture the inference

patterns in those languages. Especially for low-resource languages with limited LLM availability, it is possible that explicit belief attribution might be better at capturing human results.

Moreover, the current study uses projection inference as a test case and compares the predictive power of LLMs to RSAs to investigate whether explicit belief representations are needed during the inference process. Future studies can expand the range of pragmatic inferences, both in English and across other languages. As suggested by one reviewer, the RSA framework has been adapted cross-linguistically in the case of manner implicatures in Mandarin Chinese (Cong, 2021) and pronoun resolution in French (Schulz et al., 2021). Thus, comparing the predictions of LLMs and RSA cross-linguistically in different pragmatic tasks will better inform the pragmatic theories in terms of the role of explicit representations of interlocutors' mental states during conversations.

Acknowledgments

We would like to thank Catherine Arnett, Judith Degen, Andrew Kehler, and Zachary Houghton for helpful comments and discussion.

References

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Daiki Asami and Saku Sugawara. 2023. Propres: Investigating the projectivity of presupposition with various triggers and environments. *arXiv preprint arXiv:2312.08755*.

Leon Bergen, Noah Goodman, and Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, pages 120–125.

Gaia Carenini, Louis Bodot, Luca Bischetti, Walter Schaeken, and Valentina Bambini. 2023. Large language models behave (almost) as rational speech actors: Insights from metaphor understanding. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*.

Yan Cong. 2021. *Competition in Natural Language Meaning: The Case of Adjective Constructions in Mandarin Chinese and Beyond*. Michigan State University.

Judith Degen. 2023. The Rational Speech Act Framework. *Annual Review of Linguistics*, 9:519–540.

Judith Degen and Judith Tonhauser. 2021. Prior beliefs modulate projection. *Open Mind*, 5:59–70.

Judith Degen and Judith Tonhauser. 2022. Are there factive predicates? an empirical investigation. *Language*, 98(3):552–591.

Kajsa Djärv and Hezekiah Akiva Bacovcin. 2020. Prosodic effects on factive presupposition projection. *Journal of Pragmatics*, 169:61–85.

Kajsa Djärv and Hezekiah Akiva Bacovcin. 2020. Prosodic effects on factive presupposition projection. *Journal of Pragmatics*, 169:61–85.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023a. 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.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023b. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, 11:885–901.

Mingyue Jian and N. Siddharth. 2024. Are llms good pragmatic speakers? *Preprint*, arXiv:2411.01562.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: Bert for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

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*.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Paul Kiparsky and Carol Kiparsky. 1970. *FACT*, pages 143–173. De Gruyter Mouton, Berlin, Boston.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45).

Alexandra Lorson. 2021. The influence of world knowledge on projectivity. Unpublished master's thesis, The University of Potsdam.

Taylor Mahler. 2020. The social component of the projection behavior of clausal complement contents. *Proceedings of the Linguistic Society of America*, 5(1):777–791.

David Marr. 1982. Vision: A computational investigation into the human representation and processing of visual information.

Dingyi Pan. 2023. Projection inferences with clause-embedding predicates in rsa models. stanford digital repository. Unpublished master's thesis, Stanford University.

Dingyi Pan and Judith Degen. 2023. Towards a computational account of projection inferences in polar interrogatives with clause-embedding predicates. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R Bowman, and Tal Linzen. 2021. Nope: A corpus of naturally-occurring presuppositions in english. *arXiv preprint arXiv:2109.06987*.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Miriam Schulz, Heather Burnett, and Barbara Hemforth. 2021. Corpus, experimental and modeling investigations of cross-linguistic differences in pronoun resolution preferences. *Glossa: a journal of general linguistics*, 6(1).

Dan Sperber and Deirdre Wilson. 2002. Pragmatics, modularity and mind-reading. *Mind & language*, 17(1-2):3–23.

Jon Stevens, Marie-Catherine de Marneffe, Shari R Speer, and Judith Tonhauser. 2017. Rational use of prosody predicts projection in manner adverb utterances. In *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society*.

Judith Tonhauser, David I Beaver, and Judith Degen. 2018. How projective is projective content? gradience in projectivity and at-issueness. *Journal of Semantics*, 35(3):495–542.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. 2020. Polite speech emerges from competing social goals. *Open Mind*, 4:71–87.

A Summary of the statistical results

A.1 Prior knowledge

$model\_prior\_predictions \sim prior\_type + (1|item)$

Model	Coefficient β	Estimate	Std. Error	df	t value	Pr(>|t|)
gpt-3.5-turbo	(Intercept)	0.16	0.05	37.57	3.02	0.00455 **
	prior_typehigh_prior	0.56	0.07	19.00	7.96	<.001 ***
gpt-4	(Intercept)	0.16	0.04	38.00	3.77	<.001 ***
	prior_typehigh_prior	0.53	0.06	38.00	9.03	<.001 ***
gpt-4o	(Intercept)	0.13	0.03	36.79	3.94	<.001 ***
	prior_typehigh_prior	0.54	0.04	19.00	12.67	<.001 ***

Table 1: Combined regression results from three models testing whether the model captures the high vs. low prior distinctions.

A.2 Prior on projection inferences

$model\_certainty\_ratings \sim prior\_type + (1 + prior\_type|item)$

Model	Coefficient β	Estimate	Std. Error	df	t value	Pr(>|t|)
gpt-3.5-turbo	(Intercept)	0.13	0.02	19.00	6.99	<.001 ***
	prior_typehigh_prior	0.18	0.04	19.00	4.69	<.001 ***
gpt-4	(Intercept)	0.30	0.04	19.00	8.35	<.001 ***
	prior_typehigh_prior	0.08	0.04	19.00	2.20	0.0403 *
gpt-4o	(Intercept)	0.06	0.01	19.00	4.63	<.001 ***
	prior_typehigh_prior	0.20	0.03	19.00	6.25	<.001 ***

Table 2: Combined regression results from three models testing the effect of prior on projection inferences.

A.3 Additional predictive power of RSA results for each LLM

Base model (LLM only, without RSA predictions):

$human\_results \sim model\_certainty\_ratings + (1|item) + (1|participant)$

Full model (with RSA predictions):

$human\_results \sim model\_certainty\_ratings + RSA\_predictions + (1|item) + (1|participant)$

Model	Coefficient	npar	AIC	BIC	logLik	deviance	Chisq χ^2	Df	Pr(>Chisq)
gpt-3.5-turbo	gpt35_base	5.00	367.41	388.68	-178.71	357.41			
	gpt35_full	6.00	209.60	235.12	-98.80	197.60	159.81	1.00	<.001 ***
gpt-4	gpt4_base	5.00	296.27	317.54	-143.14	286.27			
	gpt4_full	6.00	208.36	233.88	-98.18	196.36	89.91	1.00	<.001 ***
gpt-4o	gpt4o_base	5.00	328.84	350.11	-159.42	318.84			
	gpt4o_full	6.00	209.80	235.32	-98.90	197.80	121.04	1.00	<.001 ***

Table 3: Combined model comparison results testing whether RSA model explains the variance in human results that is not captured by LLMs.

A.4 Additional predictive power of each LLM's responses for RSA

Base model (RSA only, without LLM responses):

$human\_results \sim RSA\_predictions + (1|item) + (1|participant)$

Full model (with RSA predictions):

$human\_results \sim model\_certainty\_ratings + RSA\_predictions + (1|item) + (1|participant)$

Model	Coefficient	npar	AIC	BIC	logLik	deviance	Chisq χ^2	Df	Pr(>Chisq)
gpt-3.5-turbo	rsa_base	5.00	207.82	229.09	-98.91	197.82			
	gpt3.5_full	6.00	209.60	235.12	-98.80	197.60	0.22	1.00	0.6393
gpt-4	rsa_base	5.00	207.82	229.09	-98.91	197.82			
	gpt4_full	6.00	208.36	233.88	-98.18	196.36	1.46	1.00	0.2274
gpt-4o	rsa_base	5.00	207.82	229.09	-98.91	197.82			
	gpt4o_full	6.00	209.80	235.32	-98.90	197.80	0.02	1.00	0.8951

Table 4: Combined model comparison results, testing whether there is variance in human results that is not modeled by RSA but is explained by LLM responses.

A.5 Comparing human with model results for "know" and "think"

With mean-centered prior ratings

$projection\_rating \sim model + centered\_prior\_rating + (1|item)$, reference level = "human"

Coefficient	Estimate β	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.70	0.02	436.00	37.92	<.001 ***
gpt-3.5-turbo	-0.28	0.05	436.00	-5.34	<.001 ***
gpt-4	0.00	0.05	436.00	0.02	0.9871
gpt-4o	-0.12	0.05	436.00	-2.29	0.0226 *
RSA	0.14	0.06	436.00	2.46	0.0145 *
centered_prior_rating	0.21	0.05	436.00	4.44	<.001 ***

Table 5: Results of the linear mixed-effects regression predicting the certainty ratings given "know" from the model type (human, RSA, or LLM) and prior belief ratings.

Coefficient	Estimate β	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.31	0.01	436.00	24.62	<.001 ***
gpt-3.5-turbo	-0.16	0.04	436.00	-4.58	<.001 ***
gpt-4	-0.25	0.04	436.00	-6.97	<.001 ***
gpt-4o	-0.22	0.04	436.00	-6.30	<.001 ***
RSA	0.22	0.04	436.00	6.01	<.001 ***
centered_prior_rating	0.33	0.03	436.00	10.29	<.001 ***

Table 6: Results of the linear mixed-effects regression predicting the certainty ratings given "think" from the model type (human, RSA, or LLM) and prior belief ratings.

With binary prior type

$projection\_rating \sim model + prior\_type + (1|item)$, reference level = "human'

Coefficient	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.76	0.03	22.87	26.34	<.001 ***
gpt-3.5-turbo	-0.29	0.05	417.96	-5.44	<.001 ***
gpt-4	-0.01	0.05	417.96	-0.14	0.8894
gpt-4o	-0.13	0.05	417.96	-2.53	0.0118 *
RSA	0.14	0.06	418.72	2.45	0.0149 *
low_prior	-0.11	0.04	19.85	-3.03	0.0067 **

Table 7: Results of the linear mixed-effects regression predicting the certainty ratings given "know" from the model type (human, RSA, or LLM) and prior type.

Coefficient	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.38	0.02	24.34	19.24	<.001 ***
modelgpt-3.5-turbo	-0.16	0.04	418.37	-4.27	<.001 ***
modelgpt-4	-0.25	0.04	418.37	-6.71	<.001 ***
modelgpt-4o	-0.23	0.04	418.37	-6.25	<.001 ***
modelRSA	0.24	0.04	420.44	6.11	<.001 ***
prior_typelow_prior	-0.16	0.02	26.53	-6.46	<.001 ***

Table 8: Results of the linear mixed-effects regression predicting the certainty ratings given "think" from the model type (human, RSA, or LLM) and prior type.

B Alternative belief prompt

For the projection task, we tested the model predictions with the belief prompt used in Pan and Degen (2023), as in the structure "Does SPEAKER believe that ...?". One example is shown below.

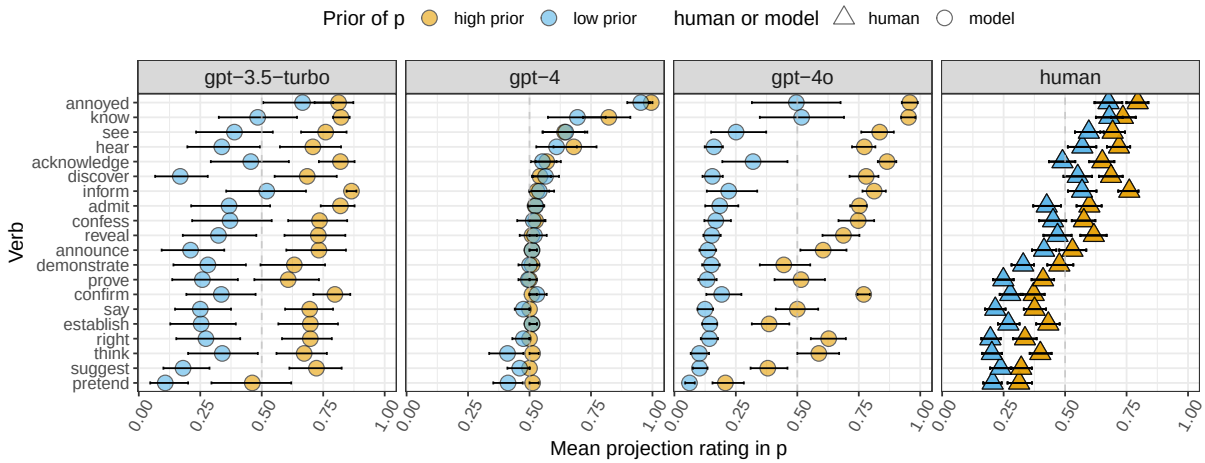
> **Fact**: Julian is German.
>
> **Sentence**: Paul asks: Does John know that Julian dances salsa?
>
> **Question**: Does Paul believe that Julian dances salsa?

Figure 5a shows the mean belief ratings of the embedded content across 20 items by predicate and by model. Both GPT-3.5-turbo and GPT-4o show the effect of prior on the projection ratings, and this observation is statistically borne out (GPT-3.5-turbo: $\beta = 0.40, t = 6.43, p < .001$; GPT-4o: $\beta = 0.46, t = 13.03, p < .001$).

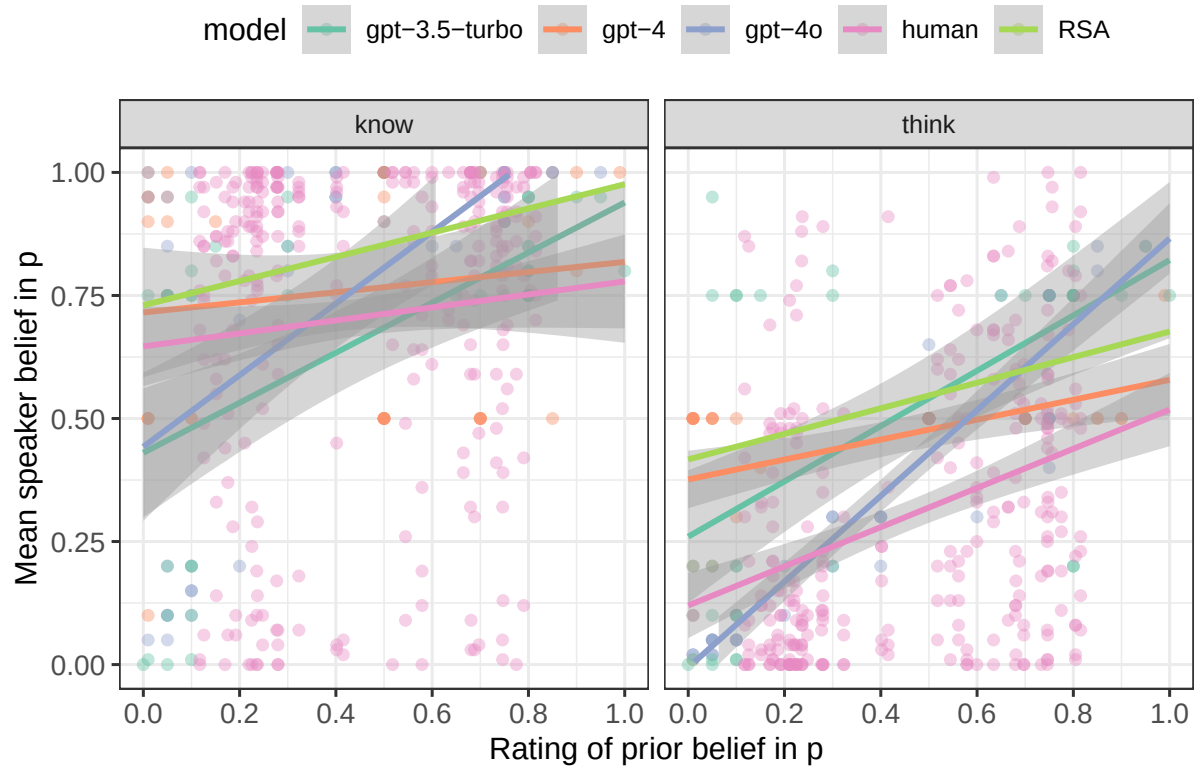
Moreover, GPT-4o qualitatively captures the gradience in the ratings among predicates, similar to the results with human participants. However, both models seem to overestimate the effect of prior. In particular, the ratings in the low prior condition are lower than those in the high prior and are more uniform among predicates, which suggests that the effect of the low prior might dominate the effect of the predicate during the inference.

In contrast, GPT-4 does not seem to capture the effect of prior ($\beta = 0.03, t = 1.30, p = 0.21$), as shown in the top right facet in Figure 5a. Furthermore, except for the canonically factive verbs like "annoyed" and "know," the ratings seem to be random, centered around 0.5. These results are different from those reported above with the "certain that" prompt, which suggests that the prompt might affect the results.

In terms of the comparison between RSA and LLMs with respect to how well they capture the human data, we fit four regression models predicting the human certainty ratings drawn from Degen and Tonhauser (2021) from the predictions of each model as well as the by-item random intercept. According to the AIC, RSA outperforms all LLMs in capturing the human results (RSA: $AIC = 215.95$), whereas GPT-4o captures human judgments better than other LLMs (GPT-3.5-turbo: $AIC = 447.45$; GPT-4: $AIC = 377.30$; GPT-4o: $AIC = 388.34$).



(a) The mean belief rating from humans and LLMs against prior ratings.



(b) The mean belief ratings for "know" and "think", against the prior rating of the embedded content.

Figure 5: Results of the belief rating tasks 5a and the comparison between LLMs responses and RSA predictions 5b. Each dot in 5a represents the mean belief rating for each verb for models and the grand mean across items and across participants for humans. The error bars in 5a and the shaded ribbon in 5b represent the 95% confidence intervals.

Figure 5b shows the mean belief ratings for "think" and "know" against prior belief ratings. The pink line represents the human data, and the other lines represent the prediction of LLMs and the RSA model. We fit the linear mixed-effects model similar to the one reported in the main content for exploratory analysis for "think" and "know." For "know", the RSA predictions are marginally different from the human data ($beta = 0.14, t = 2.53, p = 0.0117$), whereas the predictions of the LLMs are not (GPT-3.5-turbo: $\beta = -0.05, t = -0.96, p = 0.338$; GPT-4: $\beta = 0.06, t = 1.17, p = 0.242$; GPT-4o: $\beta = 0.04, t = 0.79, p = 0.4295$;). For "think", the predictions of RSA, GPT-3.5-turbo, and

GPT-4 (RSA: $\beta = 0.22, t = 5.60, p < .001$; GPT-3.5-turbo: $\beta = 0.20, t = 21.66, p < .001$; GPT-4: $\beta = 0.16, t = 4.30, p < .001$) are significantly different from human results. Interestingly, GPT-4o is not significantly different from human results ($\beta = 0.05, t = 1.45, p = .149$). In sum, this seems to suggest that models seem to be better at capturing the results of "know" but not those of "think."

C Effect of prior on certainty ratings by all 20 predicates

Figure 6 shows the by-predicate mean certainty ratings of all three GPT models and human participants against their own prior belief ratings of each item.

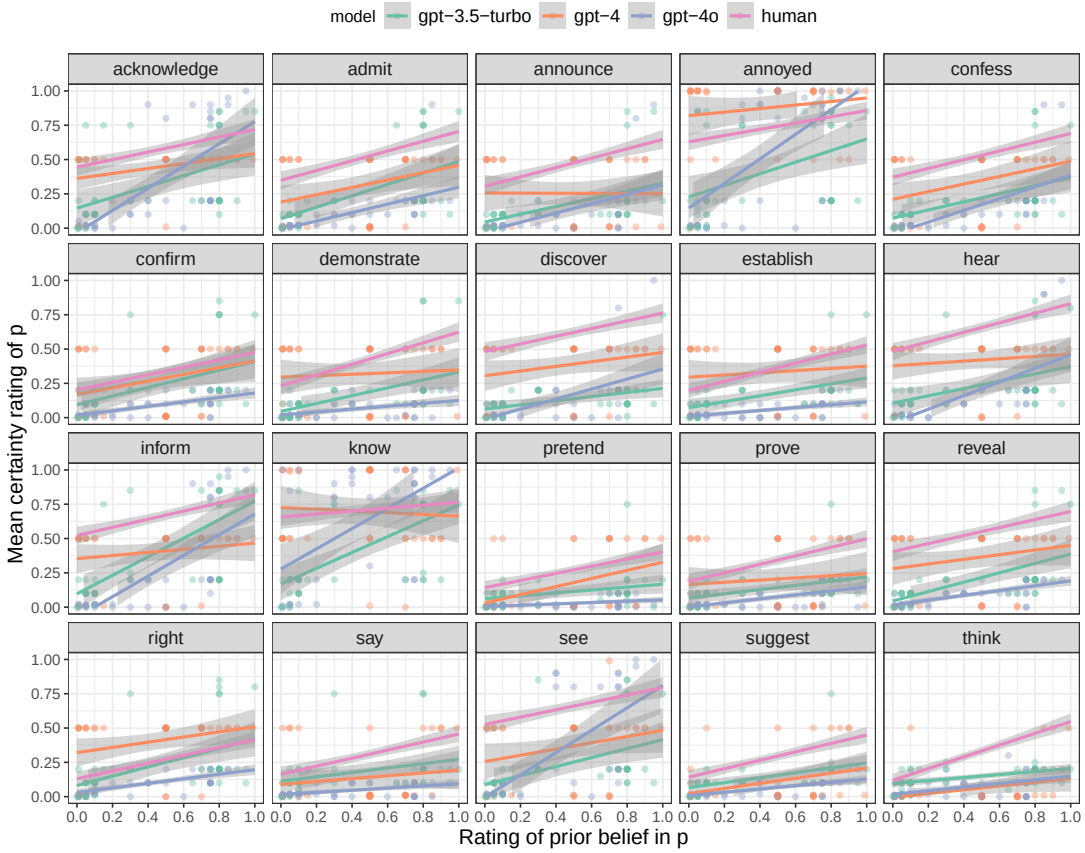


Figure 6: The mean certainty ratings for all 20 verbs against the prior rating of the embedded content. The human data are drawn from (Degen and Tonhauser, 2021). Each dot represents each model's certainty and prior rating of items for each verb, and the ribbons represent bootstrapped 95% confidence intervals.