Do Large Language Models Recognize and Utilize Non-Mandated Pragmatic Enrichments?

Dingyi Pan (dipan@ucsd.edu)
Department of Linguistics, UC San Diego
La Jolla, CA, 92093

Andrew Kehler (akehler@ucsd.edu)
Department of Linguistics, UC San Diego
La Jolla, CA, 92093

Abstract

Large language models (LLMs), despite being trained primarily on a word prediction task, show remarkable language production and comprehension abilities. Whereas larger and more recent models have achieved partial success on various *pragmatic* tasks, most have only been evaluated on their ability to draw 'mandated' pragmatic inferences (e.g., implicature, presupposition) in which the felicity of a sentence is at stake. In this study, we focus on CONVERSATIONAL ELICITURES (Cohen & Kehler, 2021), a type of non-mandated pragmatic inference that, in the class of cases considered here, involves the potential inference of a causal relation between a proposition denoted by a matrix clause and one derived from a relative clause associated with a direct object (e.g., in sentences like *Melissa detests the children who are arrogant and rude*, the inference that the detesting is a result of the arrogance/rudeness). We investigate whether LLMs are able to draw such inferences and use them in downstream syntactic processing. Our results suggest that larger and more recent models do in fact exhibit these capabilities, at least to some degree.

Keywords: large language models; computational pragmatics

Introduction

Large language models (LLMs) consistently produce coherent and meaningful sentences and discourses, hence demonstrating impressive linguistic abilities. In light of this, various studies have closely examined these abilities to assess the extent to which they parallel known properties of human language interpretation (Linzen, Dupoux, & Goldberg, 2016; Warstadt et al., 2020, *inter alia*). Whereas much of this research has focused on evaluating their syntactic and semantic abilities, fewer studies have examined their skills in the domain of pragmatics. Problems in pragmatics pose unique challenges to LLMs due to their heavy dependence on inference, world knowledge, and context (Chang & Bergen, 2024).

Whereas early LLMs were found to be lacking in certain pragmatic capabilities, more recent ones have seen improvements. For instance, early transformer models such as GPT-2 and DIALOGPT showed mixed results in their abilities to detect and evaluate discourse and dialog coherence (Beyer, Loáiciga, & Schlangen, 2021). Similarly, the InstructGPT model in the GPT-3 family failed to correctly infer the implied meaning in scalar implicatures and presupposition (Cong, 2022). However, a recent study by Hu, Floyd, Jouravlev, Fedorenko, and Gibson (2023) tested a range of models with different sizes and structures on a set of pragmatic tasks in a multiple-choice setup. Most models, except GPT-2 and an instruction-tuned GPT-3 model, achieved performance better than chance in their coherence task, which requires the models to assess whether there is a coherence relation between a pair of sentences. Moreover, large-scale models, such as FlanT5 XL and another GPT-3 model (text-davinci-002), also achieved high accuracy in pragmatic tasks that involve the inference of non-literal meaning, including indirect speech acts, metaphor, and irony. In fact, the performance of these models improves as the number of parameters increases, and in cases where the models did not correctly choose the option that matched the pragmatic interpretation, they were more likely to choose the literal interpretations than options based on lexical similarity.

In this work, we evaluate LLMs on a novel type of pragmatic enrichment that Cohen & Kehler (2021) term CONVERSATIONAL ELICITURE. Cohen & Kehler argue that, unlike more commonly studied types of pragmatic enrichment (implicature, presupposition), the inference of elicitures is not triggered by any threat of communicative failure. For example, in a typical context, sentence (1a) invites the addressee to infer that the speaker intends to convey that not only are the children detested by Melissa and are arrogant and rude, but that they are detested by Melissa because they are arrogant and rude. Note that this inference is not triggered by any syntactic relationship or other type of linguistic felicity requirement that applies to the sentence, and thus the inference is "non-mandated" in nature. This can be seen in (1b), which is likewise perfectly felicitous despite the fact that it will typically not convey an analogous eliciture that causally relates Melissa's detesting to the place where the children live.

1. a) Melissa detests the children who are arrogant and rude.
   b) Melissa detests the children who live in La Jolla.

Previous studies with human participants show that addressees not only draw conversational elicitures, but put them to use in sentence processing tasks such as relative clause (RC) attachment (Rohde, Levy, & Kehler, 2011; Hoek, Rohde, Evers-Vermeul, & Sanders, 2021) and pronoun interpretation (Kehler & Rohde, 2019). In light of their non-mandated status, one might wonder whether LLMs acquire the ability to recognize elicitures if for no other reason than their potential to improve word prediction. Addressing this question is the goal of this study. We ask 1) do LLMs have the ability to recognize elicitures, and 2) can LLMs utilize the potential for eliciture in downstream linguistic tasks?

We present two experiments, the design of both of which make use of minimal sentence pairs that contrast object-biased implicit causality (IC) verbs, e.g., *detest* in (2a), with nonIC verbs, e.g., *babysit* in (2b).

2. a) Melissa detests the children.
   b) Melissa babysits the children.

IC verbs are so-called because they are said to impute causality to one of the participants associated with the eventuality they denote, which in turn creates a strong bias toward mentioning that participant in an ensuing clause that offers an explanation (i.e., a cause or a reason) for the occurrence of that eventuality (Garvey & Caramazza, 1974; Brown & Fish, 1983, *inter alia*). For IC verbs that are object-biased such as *detest*, comprehenders hence expect to hear the object mentioned again in an explanation: If Melissa detests the children, then the cause is likely to originate from a property of the children. In contrast, nonIC verbs like *babysits* are associated with weaker and less consistent biases.

A second bias that differentiates IC verbs from others is that they create different sets of expectations for what type of continuation will ensue. Specifically, Kehler, Kertz, Rohde, and Elman (2008) found that IC verbs yield far more explanation continuations (˜60%) than do context sentences with nonIC verbs (˜24%). At an intuitive level, the lexical semantics of verbs like *detest* appear to lead the addressee to ask *Why?* in a way that verbs like *babysit* do not.

Experiment 1 utilizes sentence frames like (1) to examine whether the LLMs under scrutiny are able to detect elicitures in those cases in which one exists. Experiment 2 then examines whether the models are able to put them to use in making predictions about syntactic processing and word prediction.

Experiment 1: Detecting elicitures

Background

A prerequisite to examining whether LLMs put elicitures to use in downstream processing is showing that they can infer elicitures in the first place. This is the goal of Experiment 1.

Methods

Models We evaluated the performance of three closed-source models from the GPT family: GPT-3.5-turbo, GPT-4 (OpenAI, 2023), and GPT-4o (OpenAI, 2024). All three models are trained with reinforcement learning with human feedback (RLHF) to align with human users.

As these models do not provide access to their underlying probability distributions, we evaluate their abilities via prompting tasks. Since prompting results have been shown to not consistently align with underlying probabilities (Hu & Levy, 2023), we also examine five open-source models, including GPT-2 (Radford et al., 2019) and four Llama-3.2 models that include the base models, Llama-3.2-1B (1.23B parameters) and Llama-3.2-3B (3.21B parameters), as well as the instruction-tuned version of these base models. The two base models minimally differ in terms of their number of parameters, both of which are larger than the GPT-2 model (124M parameters). The two instruction-tuned models use supervised fine-tuning and RLHF on the corresponding base models. The responses from closed-source models were obtained through the OpenAI API, and all open-source models were accessed through Hugging Face.

Stimuli Sixty sets of items were used in four conditions that vary in terms of RC type (explanation vs. no-explanation) and the two verb types (IC vs. nonIC). An example stimulus set is shown in (3).

3. a) Melissa detests the children who are arrogant and rude. [IC, ExplRC]
   b) Melissa detests the children who live in La Jolla. [IC, noExplRC]
   c) Melissa babysits the children who are arrogant and rude. [nonIC, ExplRC]
   d) Melissa babysits the children who live in La Jolla. [nonIC, noExplRC]

ExplRC indicates that the RC is intended to provide a plausible explanation for the eventuality denoted by the sentence containing an IC verb (e.g., "who are arrogant and rude" in 3a), whereas noExplRC refers to one that is not (e.g., "who live in La Jolla" in 3b). Importantly, because elicitures in such cases arise from the co-occurrence of the RC and the matrix but not either individually, ExplRCs that give rise to an eliciture in the IC variants are not intended to do so in their corresponding nonIC variants (3c). Thus, of the four conditions, only the IC/ExplRC condition is predicted to give rise to an eliciture. We included 20 verbs per verb type, and each verb was paired with three items for each RC type, adapted from the stimuli used in the self-paced reading task in Rohde et al. (2011).

Tasks The prompts used to elicit responses from the closed-source models consisted of two parts: A system prompt that introduced the task and a main prompt that contained the instruction and stimulus. The target sentence was introduced after "Sentence:" followed by the comprehension question prompted by "Question:". The model was asked to give the response after "Answer:", as shown below.

> *Sentence*: Melissa detests the children who are generally arrogant and rude.
>
> *Question*: Does this sentence explain why Melissa detests the children? If yes, please provide an explanation. If not, just say no and you don't need to provide an explanation.
>
> *Answer*:

For the open-source models, we used the same 240 sentences with the sluiced continuation *", and I don't know why."* appended to the end of each sentence:

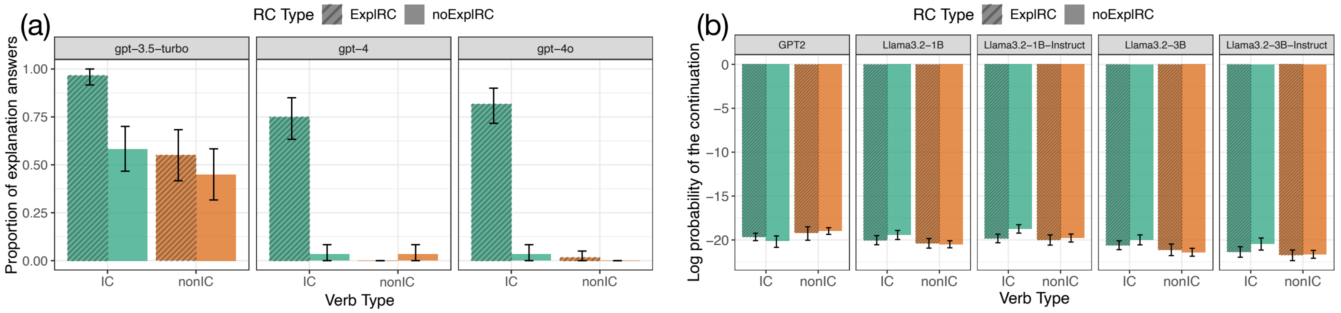

Figure 1: Proportion of explanation answers given by the three closed-source models (a) and the log probabilities of the continuation assigned by the three open-source models (b). The error bars represent 95% confidence intervals.

Sentence: Melissa detests/babysits the children who are generally arrogant and rude, and I don't know why.

Since the continuation explicitly expresses the ignorance of the cause of the event denoted by the matrix sentence, one would expect it to be less natural (and hence be less probable) in cases in which the model inferred that the RC provides the cause compared to those cases in which it does not.[1]

Evaluation For the closed-source models, we measured the number of responses in which the model answered "yes" and provided an explanation that matched the content of the RC, indicating that the model considered the RC to be the answer to the "why" question. Both "no" responses and "yes" responses with answers different from the content of the RC were coded as the model not giving an explanation answer.

For the open-source models, we first elicited the log probability of each token in the continuation *", and I don't know why."*, including the comma at the beginning and the final period. Since the continuation contains multiple tokens and is of the same length for all sentences, we summed the token-level log probabilities and then compared the aggregated log probability across the four conditions.

Results

Fig. 1(a) shows the results for the three closed-source models. GPT-3.5-turbo overgenerated explanation answers in conditions other than the IC/ExplRC condition, mostly by taking the content of the RC to be an explanation when it was not intended. However, it still shows the predicted pattern whereby there were more explanation answers in the IC/ExplRC condition than in the other three. GPT-4 and GPT-4o show a much stronger pattern, whereby they almost exclusively produced the explanation answers in the IC/ExplRC condition.

These observations were borne out statistically. We conducted a Bayesian mixed-effects logistic regression to predict the model response, i.e., whether the sentence provides an explanation as to why the event conveyed by the matrix clause occurred, from the dummy-coded main effects of verb type (reference level: IC) and RC type (reference level: ExplRC) as well as the maximal random effects structure that allowed the model to converge, including the by-item random intercept and slopes for the effects of verb type and RC type.[2] Weakly informative priors were included in the model.[3]

We considered an effect significant if 0 is not included in the credible interval. Across the closed-source models, there was a significant main effect of verb type (GPT-3.5-turbo: $\beta = -3.58$, $CrI = [-5.72, -1.94]$; GPT-4: $\beta = -5.87$, $CrI = [-8.41, -3.96]$; GPT-4o: $\beta = -6.00$, $CrI = [-8.59, -4.12]$), suggesting the models were more likely to judge the sentences in ExplRC conditions to provide an explanation when IC verbs were used vs. nonIC verbs. In addition, the models were less likely to consider IC sentences to provide an explanation in the noExplRC condition compared to the ExplRC condition (GPT-3.5-turbo: $\beta = -3.40$, $CrI = [-20.52, -4.69]$; GPT-4: $\beta = -5.40$, $CrI = [-8.12, -3.47]$; GPT-4o: $\beta = -5.51$, $CrI = [-8.03, -3.76]$). Lastly, there was also a significant interaction between verb type and RC across the models (GPT-3.5-turbo: $\beta = 2.86$, $CrI = [1.02, 5.09]$; GPT-4: $\beta = 4.30$, $CrI = [1.34, 7.24]$).[4]

Fig. 1(b) shows the mean log probabilities of the continuation assigned by the open-source models. The data was analyzed using a Bayesian mixed-effects linear regression to predict the log probability of the continuation from the dummy-coded main effects of verb type (reference level: IC) and RC type (reference level: noExplRC), as well as the maximal random effect structure that allowed the model to converge, which includes the by-item random intercept and slopes for the effects of verb type and RC type.

For GPT-2, there was no significant main effect of RC

[1] This reasoning assumes that the antecedent of the sluice is interpreted to be the matrix sentence. Since sluicing has been argued to refer to at-issue content (AnderBois, 2010), and the denotations of restrictive RCs are not at-issue, interpretations with RC antecedents are expected to be unlikely.

[2] Models were run using the `brms` package (Bürkner, 2021) in R (R Core Team, 2022).

[3] $Normal(0,3)$ was used for fixed effects in the models analyzing GPT-4 and GPT-4o results due to the sparsity of data. All other reported models used the flat prior for fixed effects.

[4] Although Fig. 1(a) shows a strong interaction between verb type and RC type for GPT-4o, the credible interval includes 0 ($\beta = 0.02$, $CrI = [-5.26, 4.48]$), possibly due to zero observations in the nonIC/noExplRC condition. Comparing the full model to a reduced model without the interaction term yields a Bayes Factor of 23.2, providing strong evidence for the interaction effect.

type ($\beta = 0.45$, $CrI = [-0.24, 1.14]$) and no significant interaction between verb type and RC type ($\beta = -0.62$, $CrI = [-1.40, 0.17]$). Although the effect of verb type was significant ($\beta = 1.12$, $CrI = [0.46, 1.79]$), it was in the opposite direction, whereby the continuation that expresses ignorance regarding the cause is more likely when a nonIC verb is used than when an IC verb is used.

On the other hand, all Llama models revealed an effect of verb type in the noExplRC conditions, suggesting that the continuation is less likely for nonIC verbs than for IC verbs when the RC does not provide an explanation (Llama-3.2-1B: $\beta = -1.08$, $CrI = [-1.68, -0.46]$; Llama-3.2-1B-Instruct: $\beta = -1.05$, $CrI = [-1.58, -0.51]$; Llama-3.2-3B: $\beta = -1.44$, $CrI = [-2.02, -0.87]$; Llama-3.2-3B-Instruct: $\beta = -1.20$, $CrI = [-1.76, -0.64]$). This is expected in light of the aforementioned finding of Kehler et al. (2008), whereby IC verbs create a stronger expectation for an ensuing explanation than nonIC verbs do. Since non-IC verbs are less likely to raise the question *Why?*, explicitly addressing the question with "I don't know why" is predicted to be more surprising in the nonIC/noExplRC condition than in the IC/noExplRC condition. In addition, the main effect of RC type was also significant, suggesting that given an IC verb, the continuation is less likely when the RC provides an explanation than when the RC does not (Llama-3.2-1B: $\beta = -0.59$, $CrI = [-1.14, -0.06]$; Llama-3.2-1B-Instruct: $\beta = -1.08$, $CrI = [-1.61, -0.54]$; Llama-3.2-3B: $\beta = -0.62$, $CrI = [-1.22, -0.02]$; Llama-3.2-3B-Instruct: $\beta = -0.92$, $CrI = [-1.52, -0.3]$). Finally, a significant positive interaction between verb type and RC type suggests that the type of RC affected IC verbs more than the nonIC verbs (Llama-3.2-1B: $\beta = 0.73$ $CrI = [0.08, 1.39]$; Llama-3.2-1B-Instruct: $\beta = 0.89$, $CrI = [0.35, 1.42]$; Llama-3.2-3B: $\beta = 0.92$, $CrI = [0.23, 1.60]$; Llama-3.2-3B-Instruct: $\beta = 0.83$, $CrI = [0.17, 1.48]$).

Discussion

All closed-source models provided more explanation responses in the IC/ExplRC condition than in the other three. Since the sentences in the IC/ExplRC and IC/noExplRC conditions contain the same verbs but different RCs, the observed difference in the model response cannot be attributed solely to properties of the IC verbs. Likewise, the contrast between the model answers in the IC/ExplRC and nonIC/ExplRC conditions suggests that the large proportion of explanation judgments in the IC/ExplRC condition were not solely driven by the RC. Taken together, these results indicate that closed-source models have the ability to draw elicitures.

Among the open-source models, the Llama models all show the effects of verb type and the content of the RC as well as their interaction on the log probability of the continuation that expresses ignorance of the potential cause. This suggests that the models are able to draw eliciture inferences, regardless of the model size and the use of additional instruction-tuning. In contrast, GPT-2 shows no such evidence. Although the tested models have a small range of model parameters, the difference between GPT-2 and the Llama models supports the idea that model size may coarsely affect pragmatic capabilities, and specifically that there is a large improvement in performance beyond 1B parameters (Hu et al., 2023). Moreover, since the effect of pragmatic enrichment on downstream word prediction is exhibited by both the base Llama models and the instruction-tuned Llama ones, the potential (dis)advantages of instruction-tuning and RLHF are inconclusive.

These findings raise the question of whether the models can use their knowledge of eliciture to guide syntactic processing. In Experiment 2, we use ambiguous RC attachment as a test case.

Experiment 2: Using elicitures in syntactic processing

Background

Since all models except GPT-2 demonstrated the ability to detect elicitures and, in the case of the Llama models, leverage them in making word predictions, we now examine whether this pragmatic inference also affects the downstream syntactic processing of ambiguous RCs.

The studies reported on by Rohde et al. (2011) serve as the inspiration for this experiment. Rohde et al. examined sentence fragments of the sort shown in (4), in which an RC that follows the relative pronoun *who* could attach to one of two NPs, one of which is singular and one plural.

4. a) Melissa babysits the children of the musician who ____
   b) Melissa detests the children of the musician who _____

English exhibits a well-documented low-attachment bias for RCs (Frazier & Clifton, 1996; Carreiras & Clifton, 1999, *inter alia*), whereby an ensuing RC will preferentially attach to *the musician* in (4). However, Rohde et al. predicted that whereas the bias for low attachment should hold for nonIC cases (4a), the bias may shift toward high attachment in their IC variants (4b). Their reasoning runs as follows. First, as we have discussed, IC verbs create a strong expectation for an ensuing explanation. Second, we have seen that such an explanation could potentially be conveyed by the speaker via eliciture with an immediately-ensuing RC. Finally, recall also from the introduction that object-biased IC verbs create a strong expectation that any ensuing explanation will re-mention the verb's direct object. As a result, we would expect an RC that conveys an explanation eliciture to be about that object, *which is the high attachment point for the relative clause*. Therefore, on the assumption that addressees are able to integrate these three types of pragmatic information on-line and use them to inform an incremental syntactic processing decision, we would expect a greater bias toward high attachment in object-biased IC contexts than in nonIC contexts, the latter of which create neither a strong expectation for an upcoming explanation nor a strong expectation that any such explanation would be about the direct object. These predictions were confirmed in an off-line sentence completion study and an on-line reading time study (Rohde et al., 2011).

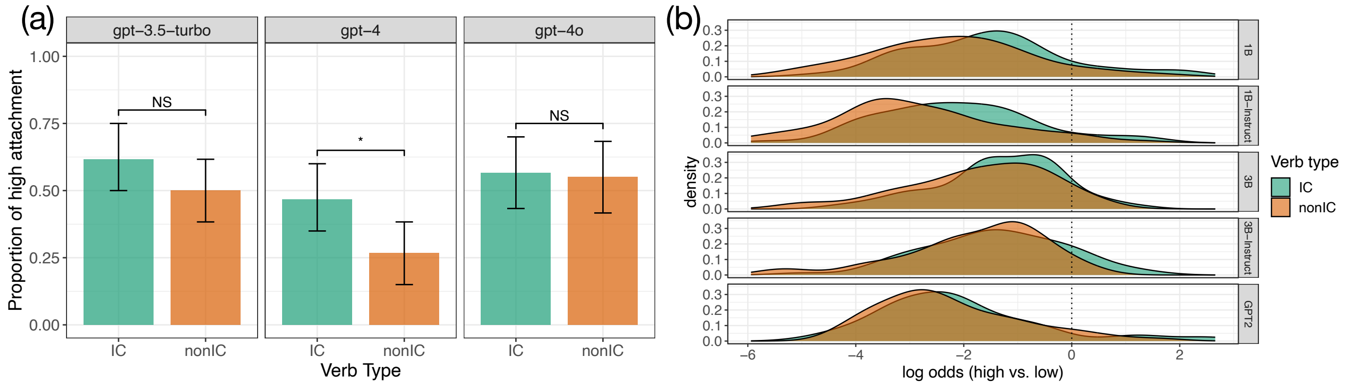

Figure 2: The proportion of responses that show high attachment bias in the three closed-source models (a) and the log-odds ratio between the probability of the critical word that reflects either the high attachment bias or the low attachment bias in three open-source models (b). The error bars represent 95% confidence intervals.

With respect to LLMs, Davis and van Schijndel (2020) found that GPT-2 XL, the largest model in the GPT-2 family, failed to show the high attachment RC bias with IC verbs. Here, we examine whether more recent LLMs show evidence of being able to integrate and use pragmatic enrichments in making this sort of syntactic attachment decision.

Methods

Models The same two sets of models as those examined in Experiment 1 were used.

Stimuli Sixty stimulus sets following the format of (4) were created, using the same pairs of verbs (IC and nonIC) used in Experiment 1, which in turn were drawn from Study 2 in Rohde et al. (2011). As in (4), the direct object of the main verb is always a complex NP containing a singular NP and a plural NP, both of which are the possible attachment sites for the RC, followed by the relative pronoun *who*. This follows the setup of the self-paced reading task in Rohde et al. (2011), where human participants anticipated which NP the RC would modify based on the verb type without seeing the full RC. The high attachment site of half of the items have a plural NP and half have a singular NP.

Tasks For the closed-source models, the stimulus was presented with two possible next words – e.g., the auxiliaries *"is"* and *"are"* – in a two-alternative forced-choice (2AFC) task:

> *Sentence*: Melissa detests/babysits the children of the musician who ____.
>
> *Options*: 1) is, 2) are

Because the two candidate NP attachment sites and the two auxiliaries differ in number, a model's choice of auxiliary reveals its attachment bias. Since models are sensitive to the order of options (Pezeshkpour & Hruschka, 2024), we randomized the order of the two options between items, so that half of the items have the option that agrees in number with the high NP appearing first, while the other half present the option that agrees with the low NP first.

For the open-source models, we obtained the raw probability of each of the auxiliary verbs (i.e., *"is"* and *"are"* in the example below) as a measure of their attachment expectations.

> *Sentence*: Melissa detests/babysits the children of the musician who is/are

As in the self-paced reading task in Rohde et al. (2011), this design directly probes whether the models leverage elicitur to predict the next word via predicting the attachment decision before an ensuing RC is even seen.

Evaluation For the closed-source models, we recorded the model choice of auxiliary, revealing the preference for either the high attachment or low attachment site.

For the open-source models, we calculated the log-odds ratio of each sentence with the two auxiliary verbs by subtracting the log probability of the auxiliary that agrees with the second NP in number, which is the low attachment site, from the log probability of the auxiliary that agrees with the first NP in number, which is the high attachment site, i.e., $\log(p_{high}) - \log(p_{low})$. Higher log-odds ratios indicate larger model preferences for high attachment.

Results

Fig. 2(a) shows the proportion of responses that indicate a preference for high attachment for each closed-source model. We fit a Bayesian mixed-effects logistic regression predicting the model response from the main effect of verb type (reference level: IC verbs) and the maximal random effect that allowed the model to converge, which includes the by-item random intercept.

Only GPT-4 shows a significant effect of verb type ($\beta = 2.73$, $CrI = [1.03, 5.11]$), whereby the model prefers high attachment more when an IC verb is used than when a nonIC verb is used. The attachment preference was not significantly different between the two verb types in GPT-3.5-turbo ($\beta = 0.57$, $CrI = [-0.22, 1.39]$) and GPT-4o ($\beta = 0.15$, $CrI = [-0.87, 1.20]$), suggesting that neither exhibited a high attachment preference triggered by IC verbs.

For the open-source models, Fig. 2(b) shows the distribution of the log-odds ratio between the probabilities of the critical word revealing a high attachment preference and the probabilities of the critical word revealing a low attachment preference. We fit a Bayesian mixed-effects linear regression predicting the log-odds ratio value from the main effect of verb type and the maximal random effect structure that allowed the model to converge, which includes the by-item random intercept.

There was a significant main effect of verb type for all four Llama models, such that the log-odds ratio was lower when a nonIC verb was used than when an IC verb was used (Llama-3.2-1B: $\beta = -0.66$, $CrI = [-1.06, -0.26]$; Llama-3.2-1B-Instruct: $\beta = -0.74$, $CrI = [-1.07, -0.42]$; Llama-3.2-3B: $\beta = -0.39$, $CrI = [-0.70, -0.10]$; Llama-3.2-3B-Instruct: $\beta = -0.38$, $CrI = [-0.66, -0.08]$). This result suggests that for the Llama models, the high attachment preference is stronger when an IC verb is used than when a nonIC verb is used. However, the effect of verb type was not significant for GPT-2 ($\beta = -0.17$, $CrI = [-0.46, 0.11]$).

Discussion

Among the closed-source models, only GPT-4 shows a higher high-attachment preference for IC verbs than for nonIC verbs, in line with the human results. In contrast, neither GPT-3.5-turbo nor GPT-4o shows a significant difference between the two verb types. One possible explanation is that GPT-4 has more parameters than the other two models. That said, since the number of parameters and the exact model structure are unknown for these models, definitive conclusions cannot be drawn. Moreover, the non-significant results might be a byproduct of the prompting task. Whereas model responses to prompts have been treated in the literature as a proxy for the underlying probability distribution, these results, especially the negative ones, may be due to the task requiring additional metalinguistic knowledge to carry out. As a result, task performance may not align with raw probabilities that reflect linguistic abilities, with this misalignment becoming more pronounced as the task diverges from next-word prediction (Hu & Levy, 2023).

On the other hand, the results show that all Llama models have a higher bias toward the high attachment site when an IC verb is used as compared to when a nonIC one is. These results suggest that not only can models infer elicitures, but indeed anticipate them as a source of information when performing word prediction, since the full RC itself is not presented to them. However, GPT-2 again does not show the intended behavior, in line with the findings in Davis and van Schijndel (2020). This suggests that GPT-2 cannot use pragmatic inference to make RC attachment decisions which, in light of the results of Experiment 1, is likely because it failed to draw elicitures in the first place. Taken together, these results suggest that larger models can use pragmatic inference to guide downstream syntactic processing in ways that pattern with the behavior of human participants, whereas smaller models are less capable in this regard.

General discussion

The results of the experiments presented here suggest that LLMs have the ability to make non-mandated pragmatic enrichments in the form of conversational elicitures, with larger and more recent models demonstrating sensitivity to the influence of pragmatic inferences on syntactic processing. Overall, our findings contribute to the positive evidence of the pragmatic abilities of LLMs and their ability to leverage pragmatic inference in guiding downstream processing tasks.

The results from Experiment 1 suggest that all models except GPT-2 have the ability to draw elicitures either by explicitly answering a comprehension question or by leveraging them during word prediction. Moreover, the results from Experiment 2 indicate that GPT-2 fails to bring the pragmatic factors described herein to bear in predicting the likely attachment site for an ensuing RC, a result that is predicted from the fact that it appears to lack the ability to draw elicitures in the first place. This result in fact aligns with prior studies showing its at-chance performance on other pragmatic tasks (Beyer et al., 2021; Hu et al., 2023). In contrast, the Llama models, regardless of their sizes and whether additional instruction-tuning is used, all demonstrate behavior consistent with their ability to use elicitures in the downstream tasks.

Since GPT-2 and the base Llama models differ in many respects, it is unclear what factors contribute to the improved performance observed in Llama models. Moreover, although both GPT-3.5-turbo and GPT-4 displayed the ability to detect elicitures in Experiment 1, both models failed to show the expected increase in high-attachment bias in the IC verb condition in Experiment 2. Future work will be necessary to evaluate a wider range of models from different model families that vary in model size and training objective.

A limitation of this study is the absence of quantitative comparisons between the model predictions and human performance on identical stimuli. Although the results of the reading time study of Rohde et al. (2011) are available, we modified the stimuli used in our Experiment 2, in part to reduce the possibility that the models had seen these stimuli during training. Hence, the results reported here cannot be compared directly to those of the original study. Furthermore, we have no human data on the tasks used in Experiment 1 nor in the 2AFC design used for the closed-source models in Experiment 2. Performing such comparisons is therefore also a subject for future work.

Acknowledgments

We would like to thank Alex Warstadt for suggesting the sluiced prompt used in Experiment 1 and for other helpful comments, and to participants at the Seventh California Meeting on Psycholinguistics for useful feedback.

References

AnderBois, S. (2010). Sluicing as anaphora to issues. In *Semantics and linguistic theory* (pp. 451–470).

Beyer, A., Loáiciga, S., & Schlangen, D. (2021, June). Is incoherence surprising? Targeted evaluation of coherence prediction from language models. In K. Toutanova et al. (Eds.), *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4164–4173). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.naacl-main.328/` doi: 10.18653/v1/2021.naacl-main.328

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *14*(3), 237–273.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. doi: 10.18637/jss.v100.i05

Carreiras, M., & Clifton, C., Jr. (1999). Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory and Cognition*, *27*, 826-833.

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, *50*(1), 293–350.

Cohen, J., & Kehler, A. (2021). Conversational eliciture. *Philosophers' Imprint*, *21*(12), 1–26.

Cong, Y. (2022, May). Psycholinguistic diagnosis of language models' commonsense reasoning. In A. Bosselut et al. (Eds.), *Proceedings of the first workshop on commonsense representation and reasoning (csrr 2022)* (pp. 17–22). Dublin, Ireland: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2022.csrr-1.3/` doi: 10.18653/v1/2022.csrr-1.3

Davis, F., & van Schijndel, M. (2020, November). Discourse structure interacts with reference but not syntax in neural language models. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th conference on computational natural language learning* (pp. 396–407). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.conll-1.32/` doi: 10.18653/v1/2020.conll-1.32

Frazier, L., & Clifton, C., Jr. (1996). *Construal*. Cambridge, Mass: MIT Press.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*(3), 459–464.

Hoek, J., Rohde, H., Evers-Vermeul, J., & Sanders, T. J. (2021). Expectations from relative clauses: Real-time coherence updates in discourse processing. *Cognition*, *210*, 104581.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023, July). A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4194–4213). Toronto, Canada: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2023.acl-long.230/` doi: 10.18653/v1/2023.acl-long.230

Hu, J., & Levy, R. (2023, December). Prompting is not a substitute for probability measurements in large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 5040–5060). Singapore: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2023.emnlp-main.306/` doi: 10.18653/v1/2023.emnlp-main.306

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44.

Kehler, A., & Rohde, H. (2019). Prominence and coherence in a bayesian theory of pronoun interpretation. *Journal of Pragmatics*, *154*, 63–78.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Pezeshkpour, P., & Hruschka, E. (2024, June). Large language models sensitivity to the order of options in multiple-choice questions. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the association for computational linguistics: Naacl 2024* (pp. 2006–2017). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2024.findings-naacl.130/` doi: 10.18653/v1/2024.findings-naacl.130

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rohde, H., Levy, R., & Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, *118*(3), 339–358.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.