Pragmatic Competence in LLMs: The Case of Eliciture

Dingyi Pan and **Andrew Kehler**
Department of Linguistics, UC San Diego
La Jolla, CA, 92093
{dipan|akehler}@ucsd.edu

Large language models (LLMs) consistently produce coherent and meaningful sentences and discourses, hence demonstrating impressive linguistic abilities. Various studies have examined these abilities to assess the extent to which they parallel known properties of human language interpretation. 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 and Bergen, 2024), and indeed results of previous studies have been mixed. On the one hand, early transformer models like GPT-2 struggle with scalar implicatures and presupposition (Cong, 2022) and fail at detecting and evaluating discourse coherence (Beyer et al., 2021). On the other hand, Hu et al. (2023) found that more recent large-scale language models achieved high accuracy in pragmatic tasks that involve reasoning about the intended meaning of the speaker.

In this paper, we evaluate LLMs on a novel type of pragmatic enrichment that Cohen & Kehler (2021) term CONVERSATIONAL ELICITURE. Consider (1a), which invites the addressee to infer that not only are the children detested by Melissa <u>and</u> are arrogant and rude, but that they are detested by Melissa <u>because</u> they are arrogant and rude.

1. (a) Melissa detests the children who are arrogant and rude. [IC, ExplRC]
   (b) Melissa detests the children who live in La Jolla. [IC, noExplRC]

Note that this inference is not triggered by any syntactic relationship or other type of linguistic felicity requirement that applies to the sentence. Thus, unlike other more commonly studied pragmatic inferences where sentence felicity is at stake (e.g., implicature, presupposition), elicitures are non-mandated. This can be seen in (1b), which is perfectly felicitous despite the fact that it will not typically convey an eliciture that casually relates Melissa's detesting to where the children live.

Previous psycholinguistic studies have demonstrated that people use eliciture inferences in sentence processing tasks such as relative clause (RC) attachment (Rohde et al., 2011; Hoek et al., 2021) and pronoun interpretation (Kehler and Rohde, 2019). Here, we ask two questions regarding the pragmatic abilities of LLMs: Whether LLMs draw elicitures (Exp. 1), and whether LLMs are able to leverage elicitures to guide downstream syntactic processing (Exp. 2).

1 Experiment 1: Detecting Elicitures

Models. We evaluated the performance of eight LLMs: three closed-source models (GPT-3.5-turbo, GPT-4, and GPT-4o) and five open-source models (GPT-2, Llama-3.2-1B, Llama-3.2-3B, and the instruction-tuned versions of the latter two models). The pragmatic abilities of the closed-source models are evaluated via prompting. Since results yielded by prompting might not be an accurate reflection of the underlying linguistic abilities of interest (Hu and Levy, 2023), we evaluate the inferential behavior of the five open-source models by measuring the log probability of a continuation (described below).
Stimuli. We used 60 sets of items in a 2x2 design varying whether the verb in the matrix sentence is an implicit causality (IC) verb (e.g., *detest* in (1)) or non-IC verb (e.g., *babysit* in (2)), and whether the relative clause (RC) conveys a causal eliciture in the IC condition (ExplRC, e.g., *"who are arrogant and rude"* in (1a)) or not (noExplRC, e.g., *"who live in La Jolla"* in (1b)). Since both the IC verb and the explanation RC are required to draw the eliciture inference, the ExplRCs that give rise to an eliciture in the IC variants are not intended to do so in their corresponding non-IC variants (2a).

2. (a) Melissa babysits the children who are arrogant and rude. [nonIC, ExplRC]
   (b) Melissa babysits the children who live in La Jolla. [nonIC, noExplRC]

Tasks. For the closed-source models, we presented each model with the target sentence and explicitly

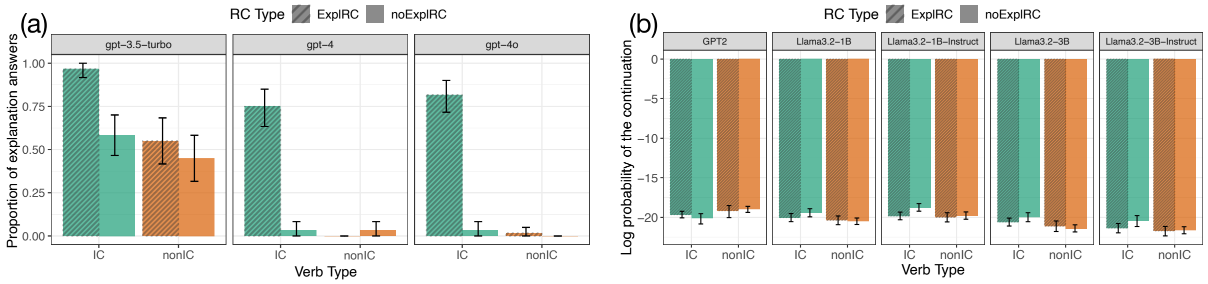

Figure 1: Proportion of explanation answers given by the three closed-source models (a) and the log probabilities of the continuation assigned by the three open-source models (b). The error bars represent 95% confidence intervals.

asked it if the sentence contains an answer to why the event in the matrix clause occurred. We measured the number of "yes" responses to the question and confirmed that the explanation provided by the model matches the content of the RC.

For the open source models, we used the same 240 sentences with the continuation *", and I don't know why."* appended to the end. This continuation should have a lower log probability (i.e., higher surprisal) if the model has inferred that the RC answers the *why*-question via causal elicitur. We summed the log probability of each token in the continuation, including punctuation.

Results. The results are shown in Fig. 1. All closed-source models revealed evidence of inferring elicitures. Although GPT-3.5-turbo overgenerated elicitures, it was still more likely to infer that sentences with IC verbs and paired explanation RCs provided answers to the *why*-question. GPT-4 and GPT-4o show a much stronger pattern, whereby they almost exclusively produced explanation answers in the IC/ExplRC condition.

Turning to the open-source models, results from the Llama models revealed that in IC contexts, the continuation was less likely in sentences in which the RC provides an explanation than when it does not. Further, there is a reliable effect of verb type in the noExplRC conditions, suggesting that when the RC does not provide an explanation, the continuation is less likely for nonIC verbs than for IC verbs. This result is expected since non-IC verbs are less likely to prompt an expectation for an explanation of the event in the matrix clause (Kehler et al., 2008) and hence raise the question *Why?*. Thus, explicitly stating "I don't know why" is predicted to be more surprising in the nonIC/noExplRC condition than in the IC/noExplRC condition. Lastly, the interaction between verb type and RC type was significant, suggesting that the type of RC affected IC verbs more than nonIC verbs. In contrast, GPT-2 showed none of the predicted effects. In sum, these results suggest that all Llama models were able to draw the eliciture inference, but not GPT-2.

Discussion. All closed-source models provided more explanation responses in the IC/ExplRC condition than in the other three conditions. Further, all Llama models showed the effects of verb and RC content as well as their interaction on the continuation that expresses the ignorance of the cause, suggesting that regardless of the model size and instruction-tuning, these models are able to draw eliciture inferences. In contrast, GPT-2 does not show any patterns that would suggest the inference of eliciture. This result is in line with previous findings of a large improvement in performance on pragmatic tasks for models with greater than 1B parameters (Hu et al., 2023).

One might worry that the expected patterns we observed in the model performance are not due to the inference of eliciture, but are instead driven by the establishment of lower-level (e.g., word) associations. We believe this interpretation is unlikely given our 2x2 design. Specifically, since sentences in the IC/ExplRC and IC/noExplRC conditions minimally differ in the content of the RC, the observed differences in the model responses and log probabilities cannot be attributed solely to the properties of IC verbs. Similarly, sentences in the IC/ExplRC and nonIC/ExplRC conditions have the same RC but different verb types, and thus the differences between conditions cannot be solely driven by the RC either. Taken together, the results suggest that all closed-source models and the Llama models show the ability to draw elicitures.

Since all of the models besides GPT-2 show evidence of being able to draw elicitures, our findings raise the question of whether these models can leverage them to guide syntactic processing. In Exp. 2, we examine the effect of eliciture in a case study using ambiguous RC attachment.

2 Experiment 2: Anticipating Elicitures

Background. Rohde et al. (2011) reported on an experiment using examples like those in Exp. 1, except where the direct object of the main verb is a complex NP containing singular and plural NPs as possible attachment sites for an ensuing RC (3).

3. (a) Melissa babysits the children of the musician who is/are ...
   (b) Melissa detests the children of the musician who is/are ...

The well-documented low-attachment bias in English predicts that the auxiliary *is* in (3a), which agrees in number with the lower NP, will be read faster than *are*, which agrees with the higher NP (Frazier, 1978; Carreiras and Clifton, 1999, *inter alia*). However, Rohde et al. (2011) predicted that this bias would shift toward high attachment for (3b), due to (i) IC verbs creating a high expectation that an explanation will ensue, (ii) that an ensuing RC might provide one through eliciture, and (iii) any such explanation would be about the direct object of the matrix verb, which is the high attachment option for the RC. Their predictions were confirmed. Here we examine whether LLMs show evidence of the same behavior.

Models. The behaviors of the same models examined in Exp. 1 were evaluated.

Stimuli. We modified the 60 stimulus sets from Exp. 1 to take the form of (3). We counterbalanced and randomized the order of the two noun phrases, such that half of the items have the plural NP as the high attachment site, and half have the singular NP as the high attachment site.

Tasks. For the closed-source models, we presented each of the two auxiliaries as possible continuations and asked the model to select between them. The order of the answer choices, reflecting either the high or low attachment bias, was balanced across items.

For the open-source models, we obtained the raw probability of each auxiliary and calculated the log-odds ratio by taking the difference, i.e., $\log(p_{high}) - \log(p_{low})$. Higher log-odds ratios indicate a greater model bias toward high attachment.

Results. The results are shown in Fig. 2. There was a significant effect of verb type for GPT-4, showing a greater high attachment preference with IC verbs than with nonIC verbs. Neither GPT-3.5-turbo nor GPT-4o showed the expected high attachment preference for IC verbs. For the open-source models, the log-odds ratio obtained from all Llama models was higher for IC sentences than nonIC ones, suggesting that the high attachment preference is stronger with IC contexts. GPT-2 did not show a difference in attachment preference between the two verb types.

Discussion. Among the three closed-source models, only GPT-4 shows an increase in the high-attachment preference when an IC verb is used than when a non-IC verb is used. Even though GPT-3.5-turbo and GPT-4o exhibited evidence of drawing elicitures when explicitly prompted in Exp. 1, neither showed a significant difference in the attachment preference between the two verb types. A possible reason for this finding is that GPT-4's performance is enabled by having more parameters than the other two models. This hypothesis remains speculative, however, since the number of parameters and the specifications of the model architectures have not been made public. In addition, the non-significant results might be a by-product of the prompting task, since prompting may require additional metalinguistic knowledge, and hence model performance may not always align with raw probabilities that reflect linguistic abilities (Hu and Levy, 2023).

On the other hand, among the five open-source models, all Llama models showed a stronger bias for the high attachment site when an IC verb is used than when a nonIC verb is used. Together with the results in Exp. 1, this suggests that these models can not only infer elicitures but also anticipate them as a source of information when processing the RC. In contrast, GPT-2 does not show the expected pattern, suggesting that it lacks the ability to use pragmatic inferences in RC attachment decisions. This result is likely due to its failure to draw elicitures in the first place, as demonstrated in Exp. 1.

3 General Discussion

The pattern we observe shows that larger and more recent LLMs demonstrate the greatest sensitivity to the presence of eliciture. On the one hand, the negative results for GPT-2 cast doubt on its ability to draw elicitures, aligning with prior studies showing at-chance performance on other pragmatic tasks (Beyer et al., 2021; Hu et al., 2023). At the same time, our findings contribute to the positive evidence of the pragmatic abilities of more recent LLMs. In Exp. 1, the three closed-source models were all able to detect eliciture in the IC/ExplRC

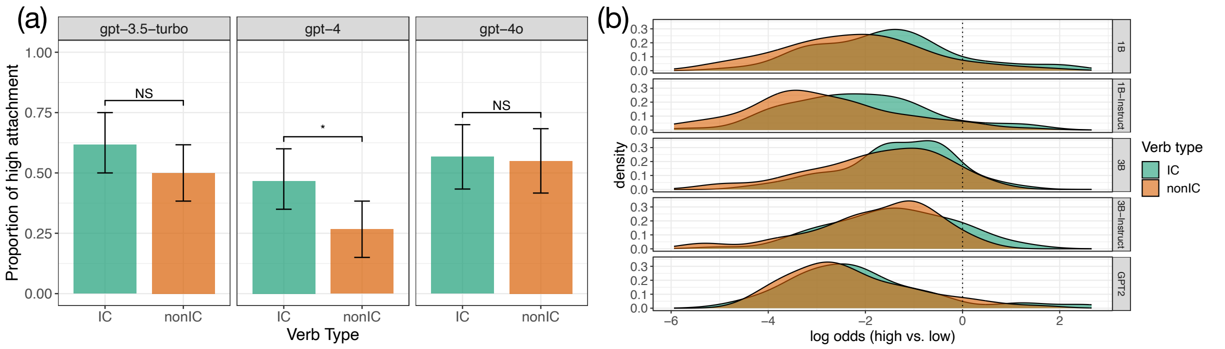

Figure 2: The proportion of responses that show high attachment bias in the three closed-source models (a) and the log-odds ratio between the probability of the critical word that reflects either high or low attachment bias in five open-source models (b). The error bars represent 95% confidence intervals.

condition, although GPT-3.5-turbo overgenerated elicitures to varying extents in the other three conditions. Similarly, all four Llama models revealed the predicted interaction whereby the log probabilities of the continuation *", and I don't know why."* were lower in the IC/ExplRC condition than the others.

In terms of the use of pragmatic inference in syntactic processing, the results of Exp. 2 suggest that the Llama models were also able to make predictions about ensuing elicitures, which in turn enabled them to make predictions about a syntactic attachment decision, as reflected by the relevant preference for a specific word (i.e., auxiliary). Moreover, even though all closed-source models were able to draw the eliciture inference when prompted, only GPT-4 displayed evidence that the anticipation of eliciture impacted the prediction of an auxiliary in the IC condition, reflecting a greater bias toward high attachment compared to the non-IC condition. Further research with other models and larger data sets will be necessary to pin down the properties of LLMs and their training that most contribute to their ability to detect and utilize eliciture.

Acknowledgments

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

References

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? Targeted evaluation of coherence prediction from language models. In *Proceedings of NAACL 2021*, pages 4164–4173.

Manuel Carreiras and Charles Clifton, Jr. 1999. Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory and Cognition*, 27:826–833.

Tyler A Chang and Benjamin K Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Jonathan Cohen and Andrew Kehler. 2021. Conversational eliciture. *Philosophers' Imprint*, 21(12):1–26.

Yan Cong. 2022. Psycholinguistic diagnosis of language models' commonsense reasoning. In *Proceedings of the first workshop on commonsense representation and reasoning (CSRR 2022)*, pages 17–22.

Lyn Frazier. 1978. *On comprehending sentences: Syntactic parsing strategies*. Ph.D. thesis, University of Conneticut.

Jet Hoek, Hannah Rohde, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2021. Expectations from relative clauses: Real-time coherence updates in discourse processing. *Cognition*, 210:104581.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of ACL-2023*, pages 4194–4213.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of semantics*, 25(1):1–44.

Andrew Kehler and Hannah Rohde. 2019. Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics*, 154:63–78.

Hannah Rohde, Roger Levy, and Andrew Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.