Pragmatic Competence in LLMs: The Case of Eliciture

Anonymous SCIL submission

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.

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 Mary and are arrogant and rude, but that they are detested by Mary because they are arrogant and rude. Note this inference is not triggered by any syntactic relationship nor other type of linguistic felicity requirement that applies to the sentence. This can be seen in (1b), which is perfectly felicitous despite the fact that it will not typically convey an eliciture that casually relates Mary's detesting to where the children live.

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

In this paper, we describe two experiments that examine whether LLMs are able to infer such elicitures and put them to use for word prediction.

1 Experiment 1: Detecting Elicitures

Models. We evaluated the performance of three closed-source models: GPT-3.5-turbo, GPT-4, and GPT-4o, and five open-source models: GPT-2 and four Llama-3.2 models, including Llama-3.2-1B and Llama-3.2-3B and the instruction-tuned version of these base models.

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 (*detest* in (1)) or non-IC verb (substitute *babysit* for *detest* in (1)), and whether the relative clause (RC) conveys a causal eliciture in the IC condition (1a) or not (1b).

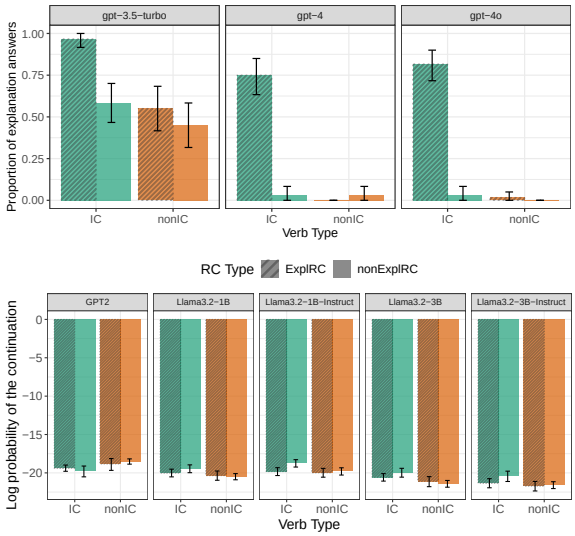

Figure 1: Results of Exp. 1 from the closed-source (a) and open-source models (b).

Tasks. For the closed-source models, we presented each model with the target sentence and explicitly 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 if the model infers the RC answers the *why*-question via causal eliciture. We summed the log probability of each token in the continuation, including punctuation.

Results. The results are shown in Fig. 1. Although GPT-3.5-turbo overgenerated elicitures, all closed-source GPT models inferred that sentences with IC verbs and paired explanation RCs provided answers to the *why*-question, demonstrating the inference of eliciture. The results from the Llama models reveal a predicted significant effect of RC type for

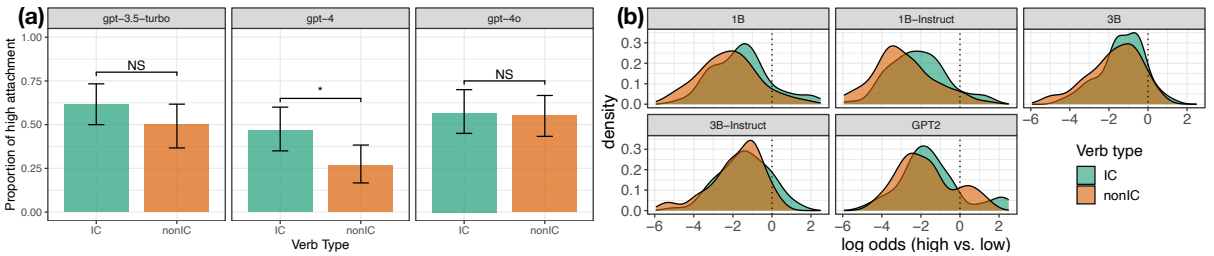

Figure 2: Results of Exp. 2 from the closed source (a) and open-source (b) models.

IC verbs, whereby the continuation has a lower log probability when an eliciture exists that already answers the *why*-question as in the IC/ExplRC condition. GPT-2 showed no predicted effects.

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 (2).

2. (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 (2a), which agrees in number with the lower NP, will be read faster than *are*, which agrees with the higher NP. However, Rohde et al. predicted that this preference would shift toward high attachment for (2b), 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.

Stimuli. We modified the 60 stimulus sets from Exp. 1 to take the form of (2).

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 positions of the NPs were balanced across items.

For the open-source models, we obtained the raw probability of each of the auxiliaries and calculated the log-odds ratio by taking the difference, i.e., $\log(p_{high}) - \log(p_{low})$. Higher log-odds ratios indicate that the model prefers 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 triggered by the IC verb. The log-odds ratio obtained from all Llama models was higher for IC sentences than nonIC ones, as predicted. GPT-2 again revealed no effect.

3 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. At the same time, our findings contribute to the positive evidence of the pragmatic abilities of more recent LLMs. For instance, the three closed-source models were all able to detect eliciture in the IC/ExplRC condition of Exp. 1, although GPT-3.5-turbo overgenerated elicitures to varying extents in the other three conditions. In Exp. 2, however, 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.

All four Llama models revealed the predicted interaction whereby the surprisal of the continuation *", and I don't know why."* was higher in the IC/ExplRC condition than the others. The results of Exp. 2 suggest that they were also able to make predictions about ensuing elicitures, which enabled them to make predictions about a syntactic attachment, and ultimately a specific word (i.e., auxiliary). 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.

References

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

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