Investigating the use of pragmatic inferences and the predictive power of language models in sentence processing

Dingyi Pan, Andrew Kehler
Department of Linguistics, {dipan, akehler}@ucsd.edu

UC San Diego

Research questions

1. Can large language models (LLMs) integrate and use pragmatic information in making syntactic relative clause (RC) attachment decisions in English?
   - **Exp. 1: Yes, (most) models can!**
2. How well can models predict the human reading time in this syntactic task that is driven by pragmatic inferences?
   - **Exp. 2: Larger and instruction-tuned models do not always have better predictive power.**

Pragmatic inference and RC attachment

Using pragmatic inference in RC attachment
Pragmatic inferences can shift the default low-attachment biases associated with English RCs toward high attachment [1,2].

a) Melissa babysits the children of the musicians who are arrogant and rude.
b) Melissa detests the children of the musicians who are arrogant and rude.

The reasoning is three-fold:
1. Implicit causality (IC) verbs (e.g., detests) create a strong expectation for an ensuing explanation [3].
2. The explanation can be provided by the immediately-following RC.
3. Object-biased IC verbs create a strong expectation that the explanation will be about the verb's direct object (*i.e., the children*).

This pragmatic inference is not mandated by any syntactic or other linguistic felicity requirement.

c) Melissa detests the children of the musicians who live in La Jolla.

The predictive power of LLMs

Surprisal theory: the processing difficulty of a word is proportional to its surprisal, $-\log P(x_i \mid x_{<i})$, estimated by language models [4,5].

Predictive power: a model's ability to predict human reading time has been used to measure their resemblance to the underlying psychological mechanism of human sentence processing [6-8].

However…
- Models systematically underestimate the magnitude of processing difficulty for different syntactic ambiguous constructions [9].
- Larger models trained on extremely large datasets are not always better in predicting reading time than smaller models, due to their "superhuman" ability in next-word prediction [10,11].
- The predictive power of instruction-tuned LLMs on reading time is worse than that of base LLMs [12].

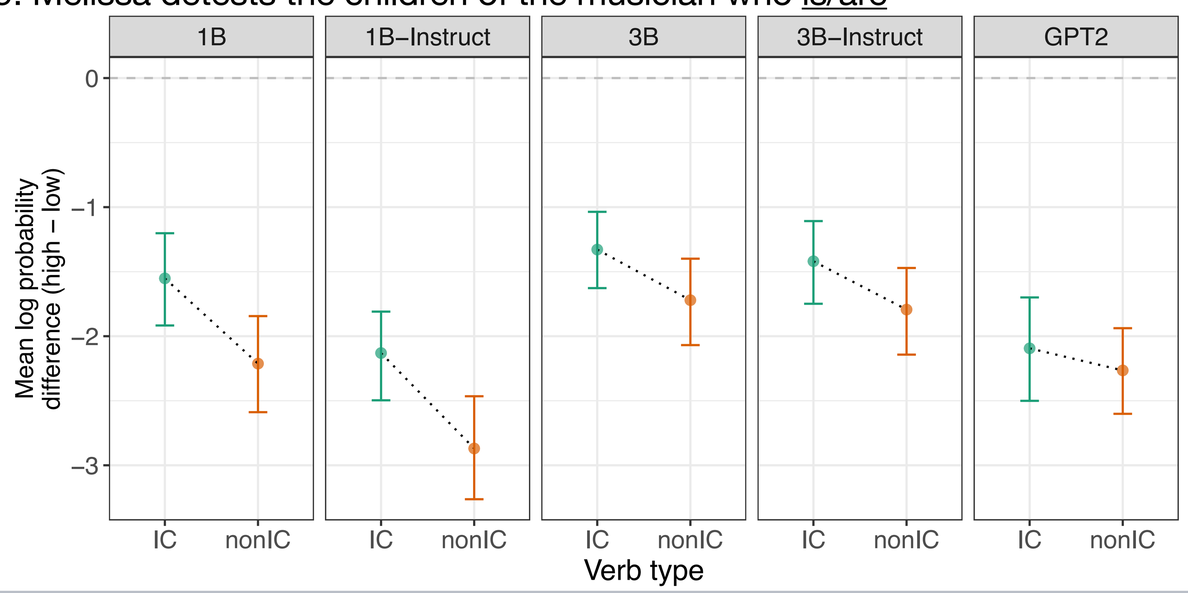
References and links

[1] Rohde, Levy, Kehler (2011). *Cognition.* [2] Hoek et al. (2021). *Cognition.* [3] Garvey & Caramazza (1974). *Linguistic Inquiry.* [4] Hale (2001). *Proceedings of NAACL*. [5] Levy (2008). *Cognition*. [6] Fossum & Levy (2012). *Proceedings of CMCL.* [7] Goodkind & Bicknell (2018). *Proceedings of CMCL.* [9] Wilcox et al. (2020). *Proceedings of Cogsci.* [10] Oh & Schuler (2023). TACL. [11] Shain et al. (2024). *PNAS.* [12] Kuribayashi et al. (2024). *Findings of NAACL.*

Data and analysis: https://github.com/pennydy/llm_eliciture.

Experiment 1: Deciding relative clause attachment site

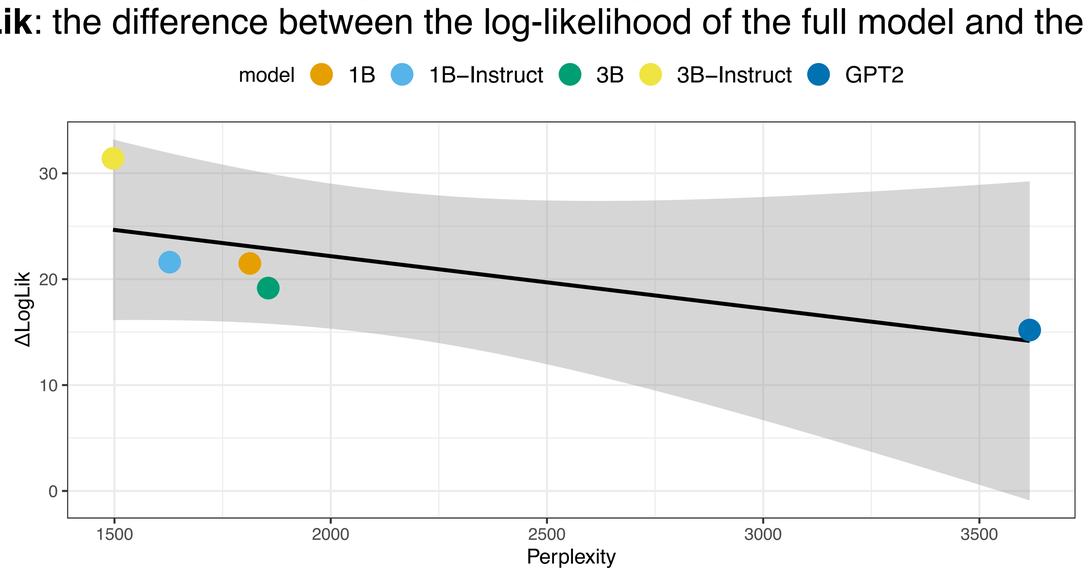
Models
GPT-2, Llama-3.2-1B, Llama-3.2-3B, Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct
Stimuli (60 sentences in each condition)
Melissa detests/babysits the children of the musician who ___ [IC/nonIC]
Prompt
Sentence: Melissa detests the children of the musician who is/are
Results



All Llama models have a higher bias toward the high attachment cite for IC verbs than for nonIC verbs, suggesting that they anticipate an explanation continuation and use it as a source of information when predicting the next word. GPT-2 does not show this behavior.

Experiment 2: Predicting reading time

Measures of critical regions and the two preceding words (for spillover effects)
Full GAM model: reading time ~ **surprisal** + word length + frequency
Baseline GAM model: reading time ~ word length + frequency
ΔLogLik: the difference between the log-likelihood of the full model and the baseline model.



The relationship between perplexity and ΔLogLik is negative, suggesting that the better the model predicts the next word (i.e., the lower the perplexity is), the better it models reading time (i.e., the larger the ΔLogLik is).

Root mean squared error (RMSE): the difference between the predicted reading time of the critical region based on a GAM model that was fit to the measures in non-critical regions and the actual reading time.

Model	RMSE
GPT-2	107.81
Llama-3.2-1B	89.81
Llama-3.2-3B	93.06
Llama-3.2-1B-Instruct	99.36
Llama-3.2-3B-Instruct	102.47

GPT-2 has the largest RMSE value, followed by the two instruction-tuned models and the two base models. For models with the same training objectives, smaller models have lower RMSE values than the larger models.
Hence, compared to larger models and instruction-tuned models, smaller models more effectively generalized the relationship between surprisal and reading time in non-critical regions to the critical regions.

Discussion

The results of the experiments presented here suggest that LLMs could generate expectations about ensuing pragmatic inferences, with larger and more recent models demonstrating sensitivity to the influence of pragmatic inferences on syntactic processing. Models lacking pragmatic inference abilities, which tend to be smaller models like GPT-2, also exhibit reduced psychometric validity in modeling human reading behavior in a sentence processing task that is modulated by pragmatic inferences. In contrast, models that appear to possess such pragmatic abilities show mixed results in their ability to predict reading times.