A large-scale investigation of pronoun interpretation biases in LLMs

Joshua K. Hartshorne
MGH Institute of Health Professions

Motivation

Humans have strong intuitions about pronoun interpretation even in the absence of much information:

1. Albert frightened Bart because he… [he = Albert]
2. Albert feared Bart because he… [he = Bart]
3. Albert frightened Bart so he… [he = Bart]
4. Albert feared Bart so he… [he = Albert]

These are biases that can be overturned:

5. Albert frightened Bart because he is is the kind of person Albert frightens.
6. Albert feared Bart because he fears everyone, not just Bart.

Are pronoun biases:

- a phenomenon in their own right (e.g., heurstics to provide early guesses?)
- a byproduct of general pronoun processing mechanisms?

LLMs match human performance at pronoun interpretation (Kocijan et al., 2023).

If pronoun biases are inherent to good pronoun interpretation, LLMs should also show these biases.

Implicit Causality & Consequentiality

Two of the best-understood biases are

- implicit causality (1-2)
- implicit consequentiality (3-4)

Complex interaction between verb and connective (verbs vary in strength of effect, connective doesn't always reverse, etc.) (Hartshorne et al., 2015).

Tests of LLMs on implicit causality have mixed results (Davis & van Schijndel, 2020; Kankowski et al., 2025; Kementchedjhieva et al., 2021; Lam et al., 2023; Upadhye et al., 2020)

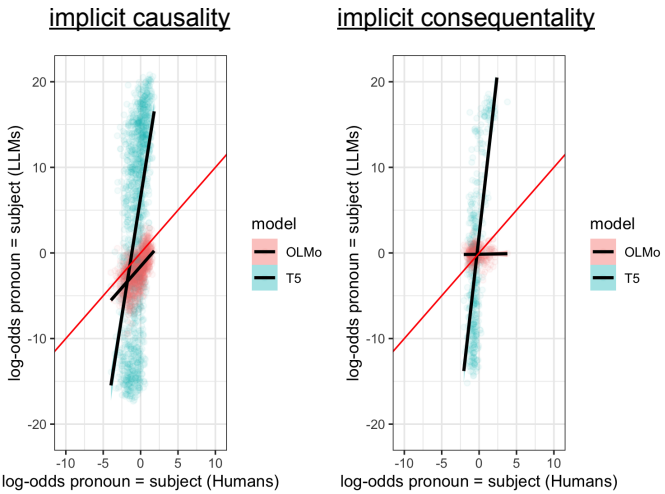
Limitations:

- Limited number of stimuli
- Most test older LLMs (typically GPT-2)
- No studies of implicit consequentiality

Study Design

- Two public datasets of human judgments (Hartshorne & Snedeker, 2013; Hartshorne et al., 2015)
  - **1,484 implicit causality sentences** (e.g., 3-4)
  - **501 implicit consquentiality sentenes** (e.g., 5-6)
- approx. 30 - 300 judgments/sentence
- Two LLMs
  - T5 11B
    - Fine-tuned for pronoun interpretation
  - OLMo-2-1124-13B-Instruct
    - Zero-shot, in-context learner
  - Unlikely either was trained on our stimuli (unlike GPT 3+)

Implicit Causality & Consequentiality



- noise ceiling: r ≅ .91
- T5: r = .77, p < .001
- OLMo: r = .77, p < .001

- noise ceiling: r ≅ .68
- T5: r = .68, p < .001
- OLMo: r = .02, p = .71

Similar results for Gemma!

Discussion & Future Directions

- T5 did well on both tasks (human-level for consequentality)
- OLMo failed completely at consequentiality
- Fits common LLM mix of good performance and catastrophic failure
  - Note: OLMo performs very well on standard LLM tests of pronoun interpretation (e.g., 68% on Winograde).
- Results suggest that good pronoun interpretation does not entail human-like pronoun biases.
- However, the two may be linked in humans.
- Suggests caution in using LLMs as replacements for human subjects.
- Open questions
  - What are differences (if any) in how LLMs and humans interpret pronouns? Does this explain differences in biases?

Bibliography

- Davis, F., & van Schijndel, M. (2020). Discourse structure interacts with reference but not syntax in neural language models. In Proceedings of the 24th Conference on Computational Natural Language Learning (pp. 396-407).
- Hartshorne, J. K., & Snedeker, J. (2013). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. Language and Cognitive Processes, 28(10), 1474–1508.
- Hartshorne, J. K., O'Donnell, T. J., & Tenenbaum, J. B. (2015). The causes and consequences explicit in verbs. Language, cognition and neuroscience, 30(6), 716-734.
- Kankowski, F., Solstad, T., Zarriess, S., & Bott, O. (2025). Implicit causality-biases in humans and llms as a tool for benchmarking llm discourse capabilities. arXiv preprint arXiv:2501.12980.
- Kementchedjhieva, Y., Anderson, M., & Søgaard, A. (2021). John praised mary because he? implicit causality bias and its interaction with explicit cues in lms. arXiv preprint arXiv:2106.01060.
- Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The defeat of the winograd schema challenge. Artificial Intelligence, 325, 103971.
- Lam, S.-Y., Zeng, Q., Zhang, K., You, C., & Voigt, R. (2023). Large language models are partially primed in pronoun interpretation. Findings of the Association for Computational Linguistics: ACL 2023, 9493– 9506.
- Upadhye, S., Bergen, L., & Kehler, A. (2020). Predicting reference: What do language models learn about discourse models? Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 977–982