The Role of Abstract Representations and Observed Preferences in the Ordering of Binomials in Large Language Models

Zachary Nicholas Houghton (znhoughton@ucdavis.edu) , Kenji Sagae, and Emily Morgan
University of California, Davis

Introduction

- Are LLMs learning abstract representations or simply copying from their training?
- Humans show abstract preferences for binomial ordering (e.g., *bishops and seamstresses vs seamstresses and bishops;* [2]).
  - e.g. Humans prefer short words first (e.g., *bread and butter*).
  - These preferences are not identical to the distribution in corpus data (i.e., human ordering preferences are not equal to the number of times each order occurs in corpus data).

Present Study

- Are LLMs' binomial ordering preferences similarly driven by abstract preferences?
- Or are their preferences driven exclusively by the proportion that each order occurs?

Methods

- LLM ordering preferences for 594 binomials [1],
  - from low-frequency (e.g., *b*ouquets *and wreathes*) to high-frequency (e.g., *black and white*)
- Coded for:
  - observed preferences, or the proportion of occurrences in a given ordering
  - abstract preferences (estimated from the model in [1]).
  - total number of occurrences in either ordering (frequency).
- Data analyzed using Bayesian regression model:

$$LogOdds(AandB) \sim + AbsPref + ObservedPref + Freq + Freq{:}AbsPref + Freq{:}ObservedPref$$

Language Model Predictions

- Obtained predictions for 8 language models:
  - GPT-2 (124M paramters), OLMo 1B (1B parameters), GPT-2 XL (1.5B parameters), Llama-2 7B (7B parameters), OLMo 7B (7B parameters), Llama-3 8B (8B parameters), Llama-2 13B (13B parameters), and Llama-3 70B (70B parameters)
- Language model predictions were calculated as the log odds of the probability of the alphabetical ordering to the probability of the nonalphabetical ordering:

$$P_{alphabetical} = P(A\ |Next\ Item{:}) * P(and\ |Next\ Item{:}\ A) * P(B\ |Next\ Item{:}\ A\ and\ )$$

$$P_{nonalphabetical} = P(B\ |Next\ Item{:}) * P(and\ |Next\ Item{:}\ B) * P(A\ |Next\ Item{:}\ B\ and\ )$$

$$LogOdds(AandB) = \log(\frac{P_{alphabetical}}{P_{nonalphabetical}})$$

- A larger value of *LogOdds(AandB)* indicates a stronger preference by the LLM for the alphabetical ordering. (Alphabetical order is used as a neutral referece order.)

Results

- LLMs do not show sensitivity to abstract preferences for the binomials in the corpus.
- Instead, they simply reproduce the binomials in proportion to the number of times they occurred in each order in their training data.

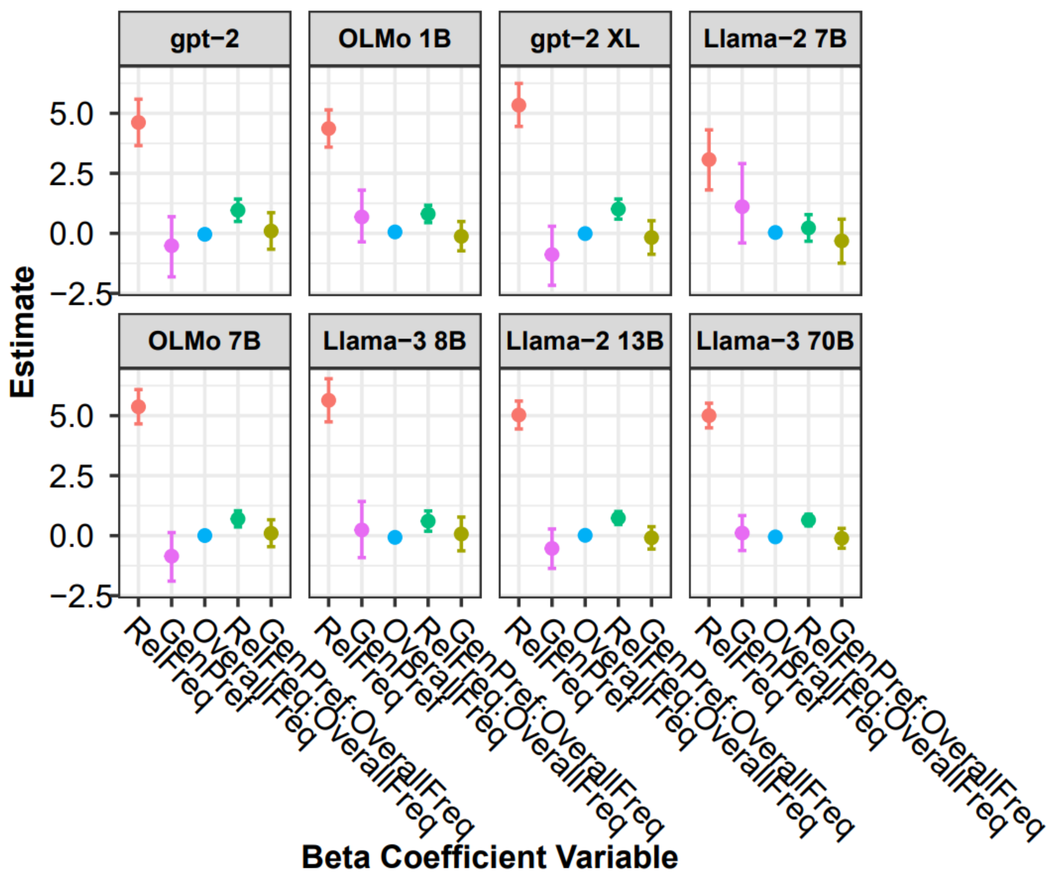

Fig 1. Results for each beta coefficient estimate from each model. Models are arranged from smallest to largest from left to right. The x-axis contains each coefficient and the y-axis contains the predicted beta coefficient of the respective model. Error bars indicate 95% credible intervals.

Summary

- LLMs' ordering preferences are driven by their experience, not abstract preferences.
- This is the case even for low-frequency binomials.
  - Though note that LLM binomial preferences are even more strongly driven by frequency for higher overall frequency binomials.
- Human binomial ordering preferences are driven by abstract preferences [2].
- Caveat: These LLMs have probably experienced these binomials before.
  - Results could instead indicate that LLMs don't rely on abstract preferences if they have experience with the binomial.
  - Still in contrast to humans, who rely on abstract preferences even for high-frequency binomials [3].

Future Work

- Examining ordering preferences for novel binomials (we have data for this already!)
  - Interestingly, LLMs do seem to use abstract preferences when ordering novel binomials.

References

- [1] Morgan, E., & Levy, R. (2015). Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 37).
- [2] Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384-402.
- [3] Morgan, E., & Levy, R. (2024). Productive knowledge and item-specific knowledge trade off as a function of frequency in multiword expression processing. *Language*, *100*(4), e195-e224.