Abstract Representations of Binomial Ordering Preferences in Large Language Models

Anonymous ACL submission

Abstract

To what extent do large language models learn abstract representations as opposed to more superficial aspects of their very large training corpora? We examine this question in the context of binomial ordering preferences involving two conjoined nouns in English. When choosing a binomial ordering (*radios and televisions* vs *televisions and radios*), humans rely on more than simply the observed frequency of each option. Humans also rely on abstract preferences beyond the specific binomials themselves (e.g., preferences for short words before long words). We investigate whether large language models simply rely on the observed preference in their training data, or whether they are capable of learning the generative preferences (i.e., abstract representations) that humans rely on. Our results suggest that smaller models are simply copying their training data, but larger models are capable of learning generative preferences. Further, similar to humans, the larger models rely more on generative preferences for low frequency items than high frequency items. Our study provides further insight into large language models' abilities to learn abstract representations of language structure and use.

1 Introduction

Large language models have progressed at an incredible rate in the last few years. Their rise in popularity and sometimes surprising capabilities have raised many questions about what exactly these models learn and how they represent linguistic knowledge. One interesting question that has been examined is whether certain capabilities emerge once models reach a certain size. These capabilities may be related to areas like language understanding, reasoning and math (Wei et al., 2022; Schaeffer et al., 2023), but may instead reflect more structural and grammatical aspects of language use (Teehan et al., 2022). Although models of different sizes appear to generate fluent language, it is unclear to what extent different models rely on superficial characteristics of their immense training corpora, such as word frequency and co-occurrences, and to what extent they learn abstract representations that generalize in ways that are similar to what humans do with far less linguistic input.

In order to investigate large language models' ability to learn abstract representations, it is useful to compare them to human Psycholinguistic data. Unlike large language models, humans don't have access to corpora with a trillion tokens. Despite this, humans' capacity for language is unparalleled, in part due to our incredible ability to learn abstract representations (Berko, 1958; Kapatsinski, 2018; Kapatsinski et al.).

One illustration of this comes from the literature on binomial constructions, where there are two conjoined nouns (e.g., *cats and dogs,* Morgan and Levy, 2015, 2016a,b; Benor and Levy, 2006). Binomial constructions often convey the same meaning regardless of the order (e.g., *radios and televisions* vs *televisions and radios*). Despite this, however, humans sometimes have very strong preferences for one order over the other (e.g., *bread and butter* overwhelmingly prefered over *butter and bread*). For some of these constructions, human ordering preferences are driven by the **observed preferences** in corpus data (i.e., the proportion of one binomial ordering to the other in corpus data, Morgan and Levy, 2016a). However, ordering preferences are sometimes driven by **generative preferences**, which are abstract preferences based on various constraints (e.g., a preference for short words before long words, or a preference for male-coded words before female-coded words, Benor and Levy, 2006).

In order to capture the ordering preferences of humans across binomial constructions, Morgan and Levy (2016a) developed a model to quantify the generative preference of a given binomial in English. The model predicts the probability that a

binomial expression is realized as *A and B* (the alphabetical form was used as a relatively unbiased baseline). This prediction is made as a function of constraints that have been shown to influence binomial ordering preferences in humans (e.g., short words before long words, Benor and Levy, 2006). Morgan and Levy demonstrated that the model's predicted generative ordering preferences are not the same as the observed preferences in corpus data. In other words, the ordering preference for a specific binomial cannot be predicted purely from the proportion of occurrences in the alphabetical order to the occurrences in nonalphabetical. This suggests that humans are not simply reproducing their input, but learning preferences from the data.

Additionally, Morgan and Levy demonstrated in a two-alternative forced choice paradigm that human ordering preferences for low frequency items are primarily driven by generative ordering preferences, and their preferences for high frequency items are driven primarily by the observed preferences in corpus data. They operationalized frequency using the **overall frequency** of a binomial, i.e. the total frequency in both possible orders. This provides a measure of expression frequency that is not confounded with the frequency of a specific order.

Since human ordering preferences deviate from the observed preferences (i.e., humans aren't simply reproducing binomials in the same order that they heard it in), ordering preferences thus present a useful test case for large language models. If large language models learn representations beyond simply memorizing the training dataset or superficially reproducing word co-occurrences, they may learn generative preferences similar to humans, and this will be reflected in their ordering preferences.

In the present study we examine binomial ordering preferences in English in four large language models with number of parameters ranging from 124M to 13B. Specifically, we ask whether ordering preferences in these models are determined entirely by the observed preferences of binomials in corpus data, or whether the language models also learn generative preferences. Further, we examine whether large language models, similarly to humans, show stronger effects of generative preferences in lower frequency items, and stronger effects of observed ordering preferences in high frequency items. If large language models are just reproducing superficial characteristics of the training data, we should see no effects of generative preferences, and only see effects of observed ordering preferences. On the other hand, if language models are doing more than just memorization, then we should see effects of generative preferences in addition to effects of observed ordering preferences, and these may change as a function of overall frequency.

2 Methods

2.1 Dataset

In order to examine the ordering preferences of binomial constructions in large language models, we use a corpus of binomials from Morgan and Levy (2015). The corpus contains 594 binomial expressions which have been annotated for various phonological, semantic, and lexical constraints that are known to affect binomial ordering preferences. The corpus also includes:

1. The estimated generative preferences for each binomial representing the ordering preference for the alphabetical ordering, estimated from the above constraints (independent of frequency). The generative preferences take a value between 0 and 1, with 0 being a stronger preference for the nonalphabetical form, and 1 being a stronger preference for the alphabetical form. The generative constraints were calculated using Morgan and Levy (2015)'s model.

2. The observed binomial orderings which are the proportion of binomial orderings that are in alphabetical order for a given binomial, gathered from the Google *n*-grams corpus (Lin et al., 2012). The Google *n*-grams corpus is magnitudes larger than the language experience of an individual speaker and thus provides reliable freuqency estimates.

3. The overall frequency of a binomial expression. Overall frequencies were also obtained from the Google *n*-grams corpus (Lin et al., 2012).

2.2 Language Model Predictions

In order to derive predictions for large language models, we used the following models from the GPT-2 (Radford et al., 2019) family and the Llama-2 (Touvron et al., 2023) family, from smallest to largest in number of parameters: GPT-2 (124M paramters), GPT-2 XL (1.5B parameters), Llama-2 7B (7B parameters), and Llama-2 13B (13B parameters). For each model, we calculated the ordering

preferences of the alphabetical form for each binomial in the dataset. The predicted probability of the alphabetical form was calculated as the product of the model's predicted probability of each word in the binomial. In order to accurately calculate the probability of the first word in the binomial, each binomial was given the prefix "Next item: ". Thus the probability of the alphabetical form, *A and B* is:

$$\begin{aligned} P_{alphabetical} &= P(A|Next\,item:) \\ &\times P(and|Next\,item:A) \\ &\times P(B|Next\,item:A\,and) \end{aligned} \quad (1)$$

where *A* is the alphabetically first word in the binomial and *B* is the other word. Additionally, the probability of the nonalphabetical form, *B and A* is:

$$\begin{aligned} P_{nonalphabetical} &= P(B|Next\,item:) \\ &\times P(and|Next\,item:B) \\ &\times P(A|Next\,item:B\,and) \end{aligned} \quad (2)$$

Finally, to get an overall ordering preference for the alphabetical form, we calculated the (log) odds ratio of the probability of the alphabetical form to the probability of the nonalphabetical form:

$$LogOdds(AandB) = log(\frac{P_{alphabetical}}{P_{nonalphabetical}}) \quad (3)$$

2.3 Analysis

The data was analyzed using Bayesian linear regression models, implemented in *brms* (Bürkner, 2017) with weak, uninformative priors. For every model, the dependent variable was the log odds of the alphabetical form to the nonalphabetical form. The fixed-effects were generative preference (represented as $GenPref$ below), observed preference ($ObservedPref$), overall frequency ($Freq$), and an interaction between overall frequency and generative preference ($Freq:GenPref$), and an interaction between overall frequency and observed preference ($Freq:ObservedPref$). The model equation is presented below:

$$\begin{aligned} LogOdds(AandB) \sim\ & GenPref \\ &+ ObservedPref \\ &+ Freq \\ &+ Freq:GenPref \\ &+ Freq:ObservedPref \end{aligned} \quad (4)$$

Frequency was logged and centered, and generative preference and observed preference were centered such that they ranged from -0.5 to 0.5 (instead of from 0 to 1). Note that since generative preference and observed preference are on the same scale, we can directly draw comparisons between the beta coefficients for these fixed-effects in our regression model.

3 Results

A visualization of our results are presented in Figure 1. The language models are ordered from left to right in terms of increasing number of parameters. For each model, the figure shows the values for each of the coefficients from the model in Equation 4, representing how strongly each language model relies on observed preference, generative preference, and overall frequency. The last two coefficients in each model additionally show interactions between overall frequency and generative preference, and overall frequency and observed preference. Our full model results are included in Appendix A.

We find no effect of generative preference in GPT-2. We also don't find an interaction effect between generative preference and overall frequency in GPT-2. We do, unsurprisingly, find an effect of observed preference in GPT-2, suggesting that the model is mostly reproducing the ordering preferences found in its training set. We also find an interaction effect between observed preference and overall frequency, suggesting that the effect of observed frequency is even stronger for high frequency items.

The effects found in GPT-2 XL are similar to the effects found in GPT-2. That is, we find an effect of observed preference and an interaction effect between observed preference and overall frequency, but no effects of generative preference. These results suggest that GPT-2 XL, similar to GPT-2, seems to mostly be reproducing the same ordering preferences as the input, though more so for high frequency items.

Llama-2 7B patterns somewhat similarly to GPT-2 XL, however the effects of generative preference are slightly stronger. While the credibility interval for generative preferences crosses zero, 92% of the posterior samples were greater than zero, suggesting a small effect of generative preference. Additionally, there is a strong effect of observed frequency, as with the two (smaller) GPT-2 models.

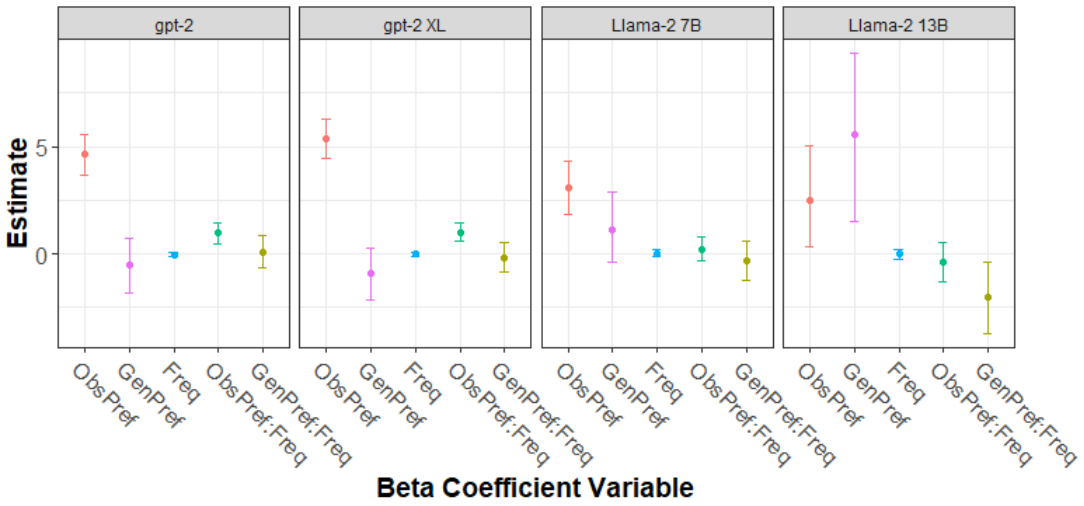

Figure 1: Results for each beta coefficient estimate from each model. Models are arranged from smallest to largest from left to right. The x-axis contains each coefficient (for generative preference, observed preference, overall frequency, interaction between overall frequency and generative preference, and interaction between overall frequency and observed preference) and the y-axis contains the predicted beta coefficient of the respective model. Error bars indicate 95% credible intervals.

Finally, for Llama-2 13B we see quite a strong effect of generative preference, and an interaction effect between generative preference and overall frequency, suggesting that the model is relying to some extent on generative preferences and is relying on them more for low frequency items than for high frequency items. This is similar to the pattern seen in humans.

Notably, for Llama-2 13B, for an item with mean overall frequency, generative preferences are a stronger predictor of the model's ordering preference than observed preferences. Note that since generative preference and observed preference are on the same scale, we can directly compare their coefficient values. The general trends are that the strength of generative preference appears to increase with model size, while the strength of observed frequency is higher for the two smaller models and decreases, while remaining positive, for the two larger models.

It is remarkable that with enough parameters the model learns this aspect of language structure and use that has been verified in human language processing, but is not apparent in superficial characteristics of corpus data. Although the English text generated by Llama-2 7B certainly appears fluent, this result shows that larger models are not simply more capable in terms of tasks and knowledge, but also in how they capture linguistic structure.

4 Conclusion

The present study examined the ordering preferences of binomial expressions in large language models and found that for smaller large language models, such as GPT-2 and GPT-2 XL, they are mostly replicating the ordering found in their training data, as should be expected. Interestingly, however, our results demonstrate that larger language models, mainly Llama-2 13B, are not simply reproducing orderings in the same proportion as was in their training data, but also learning generative preferences, which generalize over the observed data. Further, they are also relying more on generative preferences for low frequency items than for high frequency items, which is the strategy that has been observed in humans through psycholinguistic experiments (Morgan and Levy, 2016a). This result adds to our understanding of the extent to which large language models capture human cognitive processes underlying our language capabilities, as opposed to simply capturing the resulting language at a more superficial level. This understanding is crucial in determining in what ways language models may help us understand how humans learn and represent language.

5 Limitations

Our study faces a few limitations. The differences in the models tested go beyond simply the num-

ber of parameters in each model. The two smaller models tested (GPT-2 and GPT-2 XL) are from a different family of models than the two larger models (Llama-2 7B and Llama-2 13B), although all of these models are transformer-based autoregressive language models. Additionally, the Llama-2 and GPT-2 models were trained on different data. Thus, some of the differences in our current results could be due to the difference in training data as well.

References

Sarah Bunin Benor and Roger Levy. 2006. The chicken or the egg? a probabilistic analysis of english binomials. *Language*, pages 233–278.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Vsevolod Kapatsinski. 2018. *Changing minds changing tools: From learning theory to language acquisition to language change*. MIT Press.

Vsevolod Kapatsinski, Paul Olejarczuk, and Melissa A Redford. Perceptual learning of intonation: Adults are more narrow-minded than children. *Cognitive Science*.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174.

Emily Morgan and Roger Levy. 2015. Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *CogSci*. Citeseer.

Emily Morgan and Roger Levy. 2016a. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:384–402.

Emily Morgan and Roger Levy. 2016b. Frequency-dependent regularization in iterated learning. In *The Evolution of Language: Proceedings of the 11th international conference (EVOLANG 2016)*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*.

Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 2022. Emergent structures and training dynamics in large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 146–159, virtual+Dublin. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

A Appendix

Our full model results are shown in Table 1.

GPT-2				
	Est.	**Err.**	**2.5**	**97.5**
Intercept	-.10	.10	-.30	.10
GenPref	-.52	.64	-1.81	.69
Observed	**4.62**	.50	3.66	5.59
Freq	-.04	0.06	-.15	0.07
GenPref:Freq	.10	.39	-.66	0.86
Observed:Freq	**.96**	.24	.49	1.43
GPT-2 XL				
	Est.	**Err.**	**2.5**	**97.5**
Intercept	.05	.09	-.130	.23
GenPref	-.89	.63	-2.17	.29
Observed	**5.34**	.46	4.45	6.25
Freq	-.01	0.05	-.11	.09
GenPref:Freq	-.17	.36	-.87	0.53
Observed:Freq	**1.01**	.21	.59	1.43
Llama-2 7B				
	Est.	**Err.**	**2.5**	**97.5**
Intercept	.22	.13	-.03	.47
GenPref	1.11	.84	-.40	2.91
Observed	**3.07**	.64	1.81	4.31
Freq	.04	.07	-.10	.17
GenPref:Freq	-.32	.47	-1.24	.59
Observed:Freq	.23	.28	-.33	.78
Llama-2 13B				
	Est.	**Err.**	**2.5**	**97.5**
Intercept	.01	.22	-.41	.43
GenPref	**5.57**	2.00	1.54	9.32
Observed	**2.51**	1.22	.33	5.04
Freq	-.00	.12	-.23	.23
GenPref:Freq	**-2.01**	.86	-3.75	-.41
Observed:Freq	-.36	.47	-1.30	.55

Table 1: Model results for each language model. The Estimate is given in the "Est." column, the standard deviation of the posterior is given in the "Err." column. The columns labeled 2.5 and 97.5 represent the lower and upper confidence interval boundaries. GenPref is the generative preferences, Observed is the observed preference in corpus data, and Freq is the overall frequency of the binomial.