The Role of Abstract Representations and Observed Preferences in the Ordering of Binomials 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 (*radio and television* vs *television and radio*), humans rely on more than simply the observed frequency of each option. Humans also rely on abstract ordering preferences (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 abstract ordering preferences (i.e., abstract representations) that humans rely on. Our results suggest that both smaller and larger models' ordering preferences are driven exclusively by their experience with that item in the training data. Our study provides further insights into differences between how large language models represent and use language and how humans do it, particularly with respect to the use of abstract representations versus observed preferences.

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. 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. For example, in addition to learning that some binomial orderings are more frequent than others (e.g., *bread and butter* is more frequent than *butter and bread*), humans also learn abstract abstract ordering preferences (e.g., short words before long words; Morgan and Levy, 2016a).

In the present study we examine binomial ordering preferences in English in eight large language models with number of parameters ranging from 124M to 70B. 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 abstract ordering preferences. Further, we examine whether large language models, similar to humans, show 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 abstract ordering preferences, and only see effects of observed ordering preferences. On the other hand, if language models are doing more than just memorization, then we may see effects of abstract ordering preferences in addition to effects of observed ordering preferences, and these may change as a function of the binomial's frequency.

Our specific contribution is an investigation of how large language models use abstract knowledge vs. observed preferences through a binomial ordering preference task, along with a discussion about how this differs from language use by humans. We show that language models rely more on the surface-level statistics of their input (e.g, n-gram frequency) than humans do, adding to our understanding of how large language models represent and generate language.

1.1 Evidence for Abstractions in LLMs

Large language models have demonstrated incredible breakthroughs in the last few years, showing impressive capabilities across a wide variety of

tasks. Despite this, previous research has demonstrated mixed results with respect to their abilities to learn abstract representations (e.g., McCoy et al., 2023; LeBrun et al., 2022). Specifically, it remains unclear to what extent large language models are simply copying their training data as opposed to learning something more abstract. For example, Haley (2020) demonstrated that many of the BERT models are not able to reliably determine the plurality of novel words at the same level as humans.

On the other hand, Wei et al. (2021) demonstrated that BERT can generalize well to novel subject-verb pairs. Specifically, they tested BERT's subject-verb agreement ability on novel sentences that it's never seen before and found that BERT seems to learn abstract representations of subject-verb agreement (as evidenced by the fact that it performs well on items it wasn't trained on).

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 trillions of tokens. Despite this, humans' capacity for language is unparalleled, in part due to our incredible ability to learn abstract representations (Berko, 1958; Kapatsinski, 2018).

1.2 Evidence for Abstractions in Humans

Humans are remarkable in our ability to learn and produce language, often producing and processing sentences that we've never encountered before. This is largely enabled by our unique ability to not simply memorize language, but to learn more abstract generalizations. For example, humans develop abstract ordering preferences for how to linearize the message we want to convey (i.e., deciding on which order to say the words that convey the meaning we want to express). 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., *radio and television* vs *television and radio*). Despite this, however, humans sometimes have very strong preferences for one order over the other (e.g., *bread and butter* overwhelmingly preferred over *butter and bread*).

While these preferences are driven in part by experience with the binomial (i.e., which binomial ordering is encountered more often), there are also other factors, such as phonological or semantic constraints, that affect ordering preferences. In other words, human ordering preferences are driven in part by **observed preferences** in corpus data (i.e., the observed preference in their previous language experience, Morgan and Levy, 2016a) and in part driven by **abstract ordering preferences** based on abstract 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 abstract ordering preferences of humans across binomial constructions, Morgan and Levy (2016a) developed a model to quantify the abstract ordering preference of a given binomial in English. They demonstrated that the model's predicted abstract ordering preferences are not the same as the observed preferences in corpus data. They further demonstrated that human ordering preferences for low-frequency items are primarily driven by abstract 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 them; Morgan and Levy, In press), 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 abstract ordering preferences similar to humans, and this may be reflected in their binomial ordering preferences.

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 (a relatively unbiased reference form), 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 frequency estimates.

3. The overall frequency of a binomial expression (the number of times the binomial occurs in either alphabetical or non-alphabetical order). 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, the Llama-2 (Touvron et al., 2023) family, Llama-3 family (https://github.com/meta-llama/llama3), and the Olmo (Groeneveld et al., 2024) family. From smallest to largest in number of parameters: 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). 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 prepended with 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 each model, the dependent variable was the log odds of the alphabetical form to the nonalphabetical form. The fixed-effects were abstract ordering preference (represented as *AbsPref* below), observed preference (*ObservedPref*), overall frequency (*Freq*), an interaction between overall frequency and abstract ordering preference (*Freq:AbsPref*), and an interaction between overall frequency and observed preference (*Freq:ObservedPref*). The model equation is presented below:

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

Frequency was logged and centered, and abstract ordering 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 abstract ordering preference and observed preference are on the same scale, we can directly draw comparisons between the coefficient estimates for these fixed-effects in our regression model.

3 Results

Our full model results are presented in the appendix (Table 1) and visualized in Figure 1. For each model, the figure shows the values for each of the

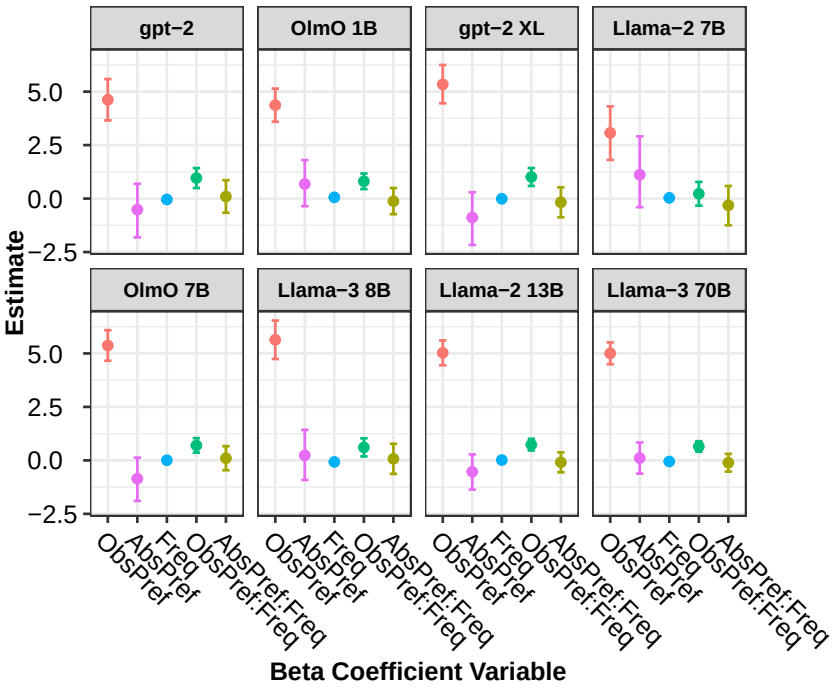

Figure 1: Results for each beta coefficient estimate from each model. Models are arranged from smallest to largest from left to right. The x-axis contains each coefficient and the y-axis contains the predicted beta coefficient of the respective model. Error bars indicate 95% credible intervals.

coefficients from the model in Equation 4, representing how strongly each language model relies on observed preference, abstract ordering preference, overall frequency, the interaction between abstract ordering preference and overall frequency, and the interaction between observed preference and overall frequency.

Our results are similar across all the large language models we tested. Specifically, we find no effect of abstract ordering preferences and no interaction effect between abstract ordering preference and overall frequency. We do find an effect of observed preference suggesting that the models are mostly reproducing the ordering preferences found in their training. We also find an interaction effect between observed preference and overall frequency, suggesting that the effect of observed frequency is stronger for high-frequency items.

4 Conclusion

In the present study we examined the extent to which abstract ordering preferences and observed preferences drive binomial ordering preferences in large language models. We find that their ordering preferences are driven primarily by the observed preferences. Further, they rely more on observed preferences for higher frequency items than lower frequency items. Finally, they don't seem to be using abstract ordering preferences at all in their ordering of binomials.

Our results give us insight into the differences between humans and large language models with respect to the ways in which they trade off between abstract and observed preferences. For example, our dataset contains low-frequency binomials (e.g. *alibis and excuses*), including binomials that a college-age speaker would have heard only once in their life. Due to their low frequency, humans rely substantially on abstract ordering preferences to process these lower frequency items (Morgan and Levy, In press). This is not the case, however, for large language models, which rely exclusively on observed preferences for these items. This is true even for the smallest models we tested, such as GPT-2. We conclude that, although large language models can produce human-like language, they accomplish this in a quantitatively different way than humans do: they rely on observed statistics from the input in at least some cases when humans would rely on abstract representations.

5 Limitations

There are a few important limitations in our study. The first limitation is that we don't know exactly how many times each of the large language models has seen each binomial tested. We can approximate the binomial's frequency using corpus data, which gives us an indication of the frequency of the binomial in a language model's training set, but it is possible that the large language models saw the binomials more than we expect. Thus, the current study can't differentiate between a model that has learned abstract ordering preferences but doesn't use it for binomials that it has seen, and a model that simply hasn't learned abstract ordering preferences. Although, there is some hope with the recent development of open access large language models, such as OlmO (Groeneveld et al., 2024), where the training data is publicly available. We have future plans to examine the ordering preferences of novel binomials in the OlmO series of models to determine whether LLMs have learned ordering preferences at all.

Additionally, the binomials tested here are only 3 words and relatively fixed in the sense that variations such as *bread and also butter* are not very common. Thus these are potentially easier for the large language models to memorize compared to longer or less-fixed strings, which could be tested in future work.

Further, while we examined language models of various sizes and determined that the number of parameters does not seem to play a role in whether these models employ abstract ordering preferences for binomials, our analysis was not designed to investigate the effect of training set size.

Finally, our experiments deal only with binomials in English.

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.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Coleman Haley. 2020. This is a bert. now there are several of them. can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341.

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

Benjamin LeBrun, Alessandro Sordoni, and Timothy J O'Donnell. 2022. Evaluating distributional distortion in neural language modeling. *arXiv preprint arXiv:2203.12788*.

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.

R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670.

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)*.

Emily Morgan and Roger Levy. In press. Productive knowledge and item-specific knowledge trade off as a function of frequency in multiword expression processing. *Language*.

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

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, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. *arXiv preprint arXiv:2109.07020*.

GPT-2					**GPT-2XL**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	-0.10	0.10	-0.30	0.10	0.05	0.09	-0.13	0.23
AbsPref	-0.52	0.64	-1.81	0.69	-0.89	0.63	-2.17	0.29
Observed	**4.62**	0.50	3.66	5.59	**5.34**	0.46	4.45	6.25
Freq	-0.04	0.06	-0.15	0.07	-0.01	0.05	-0.11	0.09
AbsPref:Freq	0.10	0.39	-0.66	0.86	-0.17	0.36	-0.87	0.53
Observed:Freq	**0.96**	0.24	0.49	1.43	**1.01**	0.21	0.59	1.43
Llama-2 7B					**Llama-2 13B**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	0.22	0.13	-0.03	0.47	0.12	0.08	-0.04	0.27
AbsPref	1.11	0.84	-0.40	2.91	0.32	0.54	-0.72	1.38
Observed	**3.07**	0.64	1.81	4.31	**5.25**	0.40	4.46	6.05
Freq	0.04	0.07	-0.10	0.17	-0.08	0.04	-0.16	0.01
AbsPref:Freq	-0.32	0.47	-1.24	0.59	-0.02	0.32	-0.64	0.60
Observed:Freq	0.23	0.28	-0.33	0.78	**0.72**	0.19	0.34	1.09
Llama-3 8B					**Llama-3 70B**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	0.15	0.09	-0.03	0.33	0.04	0.05	-0.06	0.14
AbsPref	0.23	0.59	-0.92	1.42	0.10	0.38	-0.63	0.85
Observed	**5.64**	0.46	4.75	6.54	**5.00**	0.27	4.49	5.52
Freq	-0.07	0.05	-0.17	0.03	-0.05	0.03	-0.11	0.00
AbsPref:Freq	0.07	0.36	-0.63	0.78	-0.11	0.21	-0.52	0.30
Observed:Freq	**0.60**	0.22	0.18	1.03	**0.65**	0.12	0.41	0.89
OlmO 1B					**OlmO 7B**			
	Est.	**Err.**	**2.5**	**97.5**	**Est.**	**Err.**	**2.5**	**97.5**
Intercept	0.06	0.08	-0.09	0.22	0.04	0.07	-0.10	0.18
AbsPref	0.69	0.54	-0.33	1.79	-0.86	0.51	-1.88	0.11
Observed	**4.36**	0.39	3.58	5.12	**5.37**	0.36	4.67	6.08
Freq	0.06	0.04	-0.02	0.14	0.01	0.04	-0.07	0.08
AbsPref:Freq	-0.12	0.31	-0.73	0.47	0.10	0.28	-0.47	0.64
Observed:Freq	**0.81**	0.19	0.44	1.17	**0.70**	0.17	0.37	1.04

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. AbsPref is the abstract ordering preferences, Observed is the observed preference in corpus data, and Freq is the overall frequency of the binomial.