LLM Storage Writeup

Zachary Nicholas Houghton

Introduction

In the last few years large language models have surged in popularity and have remained in the center of both the media and the recent research. With their surge in popularity has come many debates about to what extent they constitute as effective models of human language (e.g., Bender et al. 2021; Piantadosi and Hill 2022; Piantadosi 2023). These questions have stemmed from clear differences in terms of both the training that the models receive as well as the performance of these models on language processing tasks. For example, common criticisms include their insanely large training size (sometimes being trained on upwards of 15 billion tokens), the potentially unrealistic nature of their tokenization (e.g., Chat GPT tokenizes kite as [*k*, *ite*][1]), and their poor performance on tasks that are trivial to humans (e.g., counting the number of *r*'s in *strawberry*[2]).

Many of these debates are centered around the extend to which large language models are actually learning something abstract from the data and to what extent they are simply regurgitating their training data. However, despite a large body of research, the results have been quite mixed with respect to the extent that they are learning something about the abstract linguistic structure as opposed to simply copying their training data Li and Wisniewski (2021). For example, Haley (2020) demonstrated that many BERT models are not able to reliably determine the correct plural form for novel words. Similarly, Li and Wisniewski (2021) demonstrated that BERT tends to rely on memorization from its training data when producing the correct tense of novel words.

In contrast, recent research has demonstrated BERT's ability to generalize well to novel subject-verb pairs (Lasri et al. 2022) and to use abstract knowledge to predict object-past participle agreements in French (Li, Wisniewski, and Crabbé 2023). Further, McCoy et al. (2023) demonstrated that while GPT-2 copies extensively, it also produces both novel words as well as novel syntactic structures.

Additionally, in an attempt to address criticism about the unrealistic size of the training data for large language models, an interesting line of research has demonstrated that even smaller

[1] https://tiktokenizer.vercel.app/
[2] https://community.openai.com/t/incorrect-count-of-r-characters-in-the-word-strawberry/829618

language models trained on an amount of data comparable to humans seem to be able to learn abstract linguistic knowledge from the data (Misra and Mahowald, n.d.; Yao et al., n.d.). For example, Misra and Mahowald (n.d.) examined whether a language model trained on a similar amount of data as humans could learn article-adjective-numeral-noun expressions (AANNs, e.g., *a beautiful five days*). They found that even after removing AANNs from the training data, language models are still able to learn which AANNs are appropriate and which ones are not (e.g., **a blue five days*). Additionally, Yao et al. (n.d.) examined whether a similar language model can learn the length and animacy preferences for the dative alternation (give X Y vs give Y to X). They found that a language model trained on a comparable amount of data as humans is able to learn these preferences. These results together demonstrate the ability for language models to learn general knowledge about the language.

Given the evidence that large language models show evidence of both learning abstract knowledge as well as copying extensively from their training data, it's unclear in what contexts they are leveraging their stored knowledge as opposed to leveraging their more abstract knowledge.

Stored Representations in Humans

There's a great deal of evidence that many multi-morphemic words and multi-word phrases are stored holistically (Bybee and Scheibman 1999; Bybee 2003; Joseph P. Stemberger and MacWhinney 2004; Joseph Paul Stemberger and MacWhinney 1986). For example, high-frequency phrases containing *don't* (e.g., *I don't know*) are more likely to be phonetically reduced than lower frequency words containing *don't* (Bybee and Scheibman 1999). This suggests that higher-frequency phrases are represented holistically because the phonetic reduction cannot simply be attributed to phonetic reduction at the individual word-level.

Additionally, there is evidence from the psycholinguistics literature that high-frequency multi-word phrases have holistic representations (Siyanova-Chanturia, Conklin, and Heuven 2011; Morgan and Levy 2016a, 2016b, 2015, 2024; O'Donnell 2016). For example, Siyanova-Chanturia, Conklin, and Heuven (2011) demonstrated that binomials (e.g., *bread and butter*) are read faster in their frequent ordering than in their infrequent ordering. This suggests that reading times for multi-word phrases can't be reduced to the reading times of the individual words. However, it is possible that these faster reading times were driven by knowledge of generative preferences, such as a preference for short words before long words. Thus to investigate this, Morgan and Levy (2016a) examined whether human reading times for binomials is driven by generative preferences (e.g., a preference for more culturally central words first) or by experience with the binomial. Specifically, they examined whether human ordering preferences were driven simply by the relative frequency of the binomial. For example, *bread and butter* is vastly preferred over *butter and bread*. Is this driven by the fact that *bread and butter* is more frequent than *butter and bread* or driven by more abstract constraints, such as a preference to place the shorter word first? Interestingly, they found that high-frequency binomial ordering preferences are driven primarily by experience (i.e.,

the more frequent ordering is preferred) while low-frequency binomial ordering preferences are driven primarily by generative preferences (e.g., a preference for short words before long words). This suggests that humans are relying on generative knowledge for lower-frequency items, but relying on item-specific knowledge for high-frequency items, suggesting that high-frequency items (e.g., *bread and butter*) are stored holistically.

Given that humans rely on generative preferences for some binomials and item-specific preferences for other binomials, binomials present a good test case for examining this same trade off in large language models. Specifically, since humans holistically store high-frequency binomials holistically and rely more on generative knowledge for low-frequency binomials, if large language models are learning similarly, high-frequency binomials may be represented differently in each order, while compositional binomials may be represented similarly regardless of ordering.

Present Study

Since humans rely more on abstract knowledge for lower frequency items and rely more on their experience for high-frequency binomials (Morgan and Levy 2024), a natural consequence of this is that they have learned separate representations for high-frequency binomials. If large language models are learning similarly to humans, they may also learn separate representations for high-frequency binomials but not for lower-frequency binomials.

The present study addresses this question by examining the semantic representations of binomials varying in relative frequency (the proportion of occurrences in one ordering to the other ordering) and overall frequency (the overall frequency of the binomial, regardless of ordering). We examine the embeddings for both ordering of binomials in a sentence context, as well as examine the embeddings for a compositional form of the binomial (which we will elaborate on in the methods section). We hypothesize that the representations of the more frequent form (higher relative frequency form) for binomials with a high overall frequency may diverge more from the compositional representation than the less frequent ordering (lower relative frequency form) does for the same binomial. That is, for high-frequency binomials, the representation for the more frequent ordering may be more different from the compositional representation than the less-frequent ordering. For lower-frequency binomials, large language models may not learn different representations for the different orderings of the same form, regardless of the relative frequency.

In Experiment 1, we examine the representations of different binomials across different large language models and in Experiment 2 we examine the timecourse of these representations across each hidden layer in OLMo's 1B model (Groeneveld et al. 2024).

Experiment 1

In Experiment 1 we examine the representations of binomials for GPT-2, GPT-2 XL (Radford et al. 2019), OLMo-1B, OLMo-7B (Groeneveld et al. 2024), and Llama2-7B (Touvron et al., n.d.). We examine the representations for different binomials in sentence contexts as well as the compositional representations of those same binomials. We explain these metrics in detail below.

Methods

Dataset

Our dataset consists of 784 sentences containing binomials. The sentences have been annotated for both relative frequency and overall frequency. Relative frequency is operationalized as the proportion of occurrences in alphabetical order (a neutral reference order) to occurrences in nonalphabetical order. Overall frequency is operationalized as the count of *A and B* plus the count of *B and A*. Counts were obtained using the Google *n*-grams corpus (Lin et al. 2012).

Semantic Embeddings

In order to examine the semantic compositionality of binomials, we examined the semantic embeddings of five different large language models: GPT-2, GPT-2 XL (Radford et al. 2019), Llama-2 7B (Touvron et al., n.d.), OLMo 1B and OLMo 7B (Groeneveld et al. 2024).[3]

For each LLM we examined the semantic embeddings of the binomials in a sentence context. We accomplished this by passing the sentence through each large language model and extracting the second-to-last hidden layer for each of the words in the binomial. Since LLMs generate an embedding for each word, we computed the mean of these embeddings to represent the semantic embedding of the entire binomial in a sentence context (hereafter referred to as holistic embeddings). Next, we obtained the embedding for each word in the binomial individually, outside of a sentence context. We then computed the mean of these embeddings to represent the semantic embedding of the compositional form of the binomial (hereafter referred to as the compositional embeddings).

We then measured the cosine similarity between the holistic embeddings and the compositional embeddings for the alphabetical and nonalphabetical forms of each binomial. This is presented mathematically in Equation 1 and Equation 2, where cos_{α} is the cosine similarity between the holistic embeddings of the alphabetical form of the binomial and the compositional form, $cos_{\neg\alpha}$ is the cosine similarity between the embeddings of the nonalphabetical form of the binomial and the compositional form, h_{α} and $h_{\neg\alpha}$ are the embeddings of the holistic form of the binomial

[3]All of our code can be found publicly available at https://github.com/znhoughton/LLM-Storage.

in alphabetical and nonalphabetical forms respectively (in a sentence context), and c is the embeddings of the compositional form. Since c represents the mean of the embeddings for each word in the binomial out of context, order does not matter. Cosine similarity ranges from -1 to 1 where 1 indicates two extremely similar vectors and -1 indicates two extremely dissimilar vectors.

$$\cos \alpha = \frac{\mathbf{h} \cdot \mathbf{c}}{\|\mathbf{h}\| \|\mathbf{c}\|} \tag{1}$$

$$\cos \neg\alpha = \frac{\mathbf{h}_{\neg} \cdot \mathbf{c}}{\|\mathbf{h}_{\neg}\| \|\mathbf{c}\|} \tag{2}$$

For each binomial, we then calculated *LogCosSim* which is the logged quotient of cos_{α} and $cos_{\neg\alpha}$ (Equation 3). A larger positive value indicates a greater degree of similarity between the holistic embeddings for the alphabetical form and the embeddings of the compositional form (i.e., the holistic embeddings of the alphabetical form are more similar to the embeddings of the compositional form than the holistic embeddings of the nonalphabetical form are) and a larger negative value represents the opposite.

$$LogCosSim = log(\frac{cos_{\alpha}}{cos_{\neg\alpha}}) \tag{3}$$

Analysis

We used a Bayesian mixed-effects model to examine how the semantic similarity between the holistic embeddings and the compositional embeddings trade off as a function of relative and overall frequency. Specifically, we modeled *LogCosSim* as a function of overall frequency, which was centered and logged, *RelFreq* which ranged from -0.5 to 0.5 (with 0.5 representing a binomial that appears only in the alphabetical form, and -0.5 representing a binomial that appears only in the nonalphabetical form), and their interaction. Our model is presented below in Equation 4.

$$LogCosSim \sim OverallFreq * RelFreq \tag{4}$$

Results

Our results for each model are presented in Table 1 and visualized in Figure 1. Following (Houghton et al. 2024) we also report the percentage of posterior samples greater than zero. Since we are using Bayesian mixed-effects models, we are not forced into a binary of significant or non-significant. By reporting the percentage of posterior samples greater than zero, we present a more nuanced picture of our results.

Table 1: Bayesian linear mixed-effects model results of each model

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
GPT-2					
Intercept	0.00	0.00	0.00	0.00	72.44
OverallFreq	0.00	0.00	0.00	0.00	14.22
RelFreq	-0.04	0.01	-0.05	-0.02	0.00
OverallFreq:RelFreq	0.00	0.00	-0.01	0.00	0.04
GPT-2 XL					
Intercept	0.00	0.00	0.00	0.00	14.77
OverallFreq	0.00	0.00	0.00	0.00	99.67
RelFreq	0.00	0.01	-0.01	0.01	68.45
OverallFreq:RelFreq	0.00	0.00	0.00	0.01	97.82
OLMo-1B					
Intercept	0.00	0.01	-0.01	0.02	56.38
OverallFreq	0.00	0.00	-0.01	0.00	33.00
RelFreq	-0.15	0.03	-0.21	-0.10	0.00
OverallFreq:RelFreq	-0.02	0.01	-0.03	0.00	0.43
OLMo-7B					
Intercept	0.00	0.01	-0.01	0.02	55.77
OverallFreq	0.00	0.00	-0.01	0.00	32.45
RelFreq	-0.15	0.03	-0.21	-0.09	0.00
OverallFreq:RelFreq	-0.02	0.01	-0.03	0.00	0.58
Llama2-7B					
Intercept	0.00	0.00	-0.01	0.00	9.25
OverallFreq	0.00	0.00	0.00	0.00	3.40
RelFreq	-0.01	0.01	-0.03	0.01	8.06
OverallFreq:RelFreq	0.00	0.00	-0.01	0.00	3.17

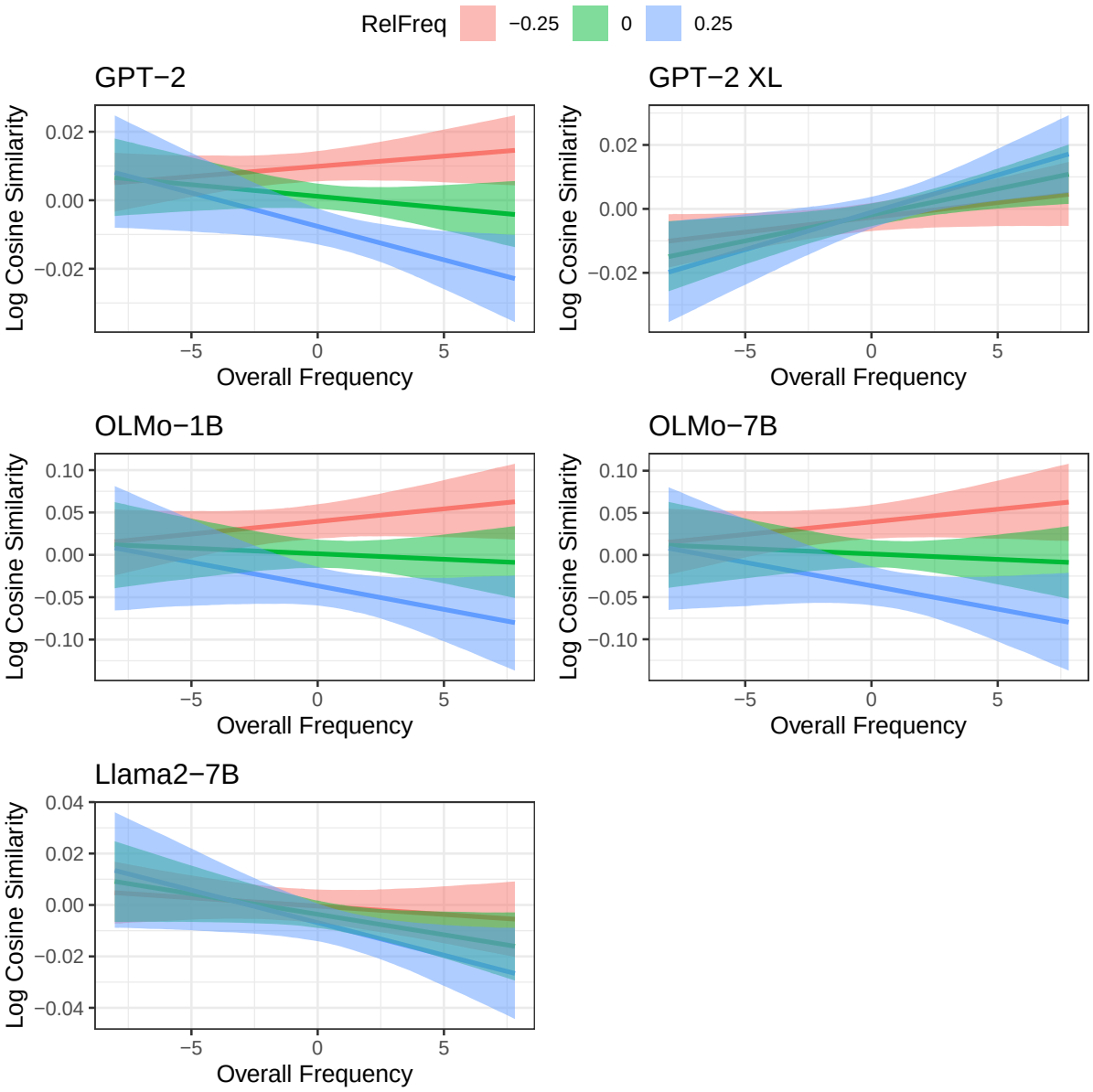


Figure 1: Visualization of the effects of overall frequency and relative frequency on the cosine similarity between embeddings.

Overall we find a negative effect of relative frequency for GPT-2, OLMo-1B, OLMo-7B, and Llama-2 7B. These results suggest that in general for those models, the embeddings for ordering of the binomial that is more frequent (i.e., higher relative frequency) are less similar to the compositional embeddings than the embeddings for the ordering of the binomial that is less frequent are. Additionally, for these models there is a negative interaction effect. This suggests that for high-frequency binomials there is a stronger effect of relative frequency than for low-frequency binomials. Specifically, the difference between the embeddings for high relative frequency binomials and the compositional embeddings is even greater in high-frequency binomials than low-frequency ones.

In general, we find mixed results for overall frequency, with GPT-2 XL showing a strong effect of overall frequency such that the embeddings for high-frequency binomials (ignoring relative frequency) are more similar to the compositional embeddings than the embeddings for low-frequency binomials are.

Finally, the interaction effect between relative frequency and overall frequency are also in the opposite direction for GPT-2 XL compared to the others. Specifically, for GPT-2 XL, as overall frequency increases, the embeddings for the more frequent ordering of the binomials are *more* similar to the compositional embeddings than the embeddings for the less frequent ordering are.

Discussion

Overall we find that for GPT-2, OLMo-1B, OLMo-7B, and Llama2-7B, the representations for more the more frequent form of the binomial diverges more from the representation of the compositional form as a function of overall frequency. That is, for low-frequency binomials, there is not much of a difference between the alphabetical and nonalphabetical orderings of binomials, however for high-frequency binomials, there is a larger difference between the representation of the more frequent ordering and the representation of the compositional ordering.

Interestingly, this is not the case for GPT-2 XL, where the opposite pattern is observed: as overall frequency increases, the representations of the more frequent ordering of the binomial become even more similar to the compositional representations. This suggests that not all large language models are learning the same preferences and further that these preferences may not be completely necessary to generate human-like text.

In summary, our results suggest that for higher frequency binomials in most large language models, the semantic representation for the more frequent form of the binomial diverges more from the representation of the compositional form. This suggests that some large language models tend to learn different representations for high-frequency binomials, similar to what has been argued that humans do (Morgan and Levy 2016a). However, it's unclear on what timescale this emerges and at what hidden layers this result holds for. For example, does this difference emerge early in training or does it take a large amount of training for these different representations to emerge? Further, since different layers have been proposed to correspond to

different functions [e.g., earlier layers may represent more phonological knowledge while later layers may represent more semantic knowledge; Tenney (2019)], it is possible that these results may be vary across different layers. In Experiment 2 we examine both of these questions.

Experiment 2

Experiment 2 is an exploratory analysis examining how representations for binomials emerge throughout training across different hidden layers. Specifically, since OLMo (Groeneveld et al. 2024) released the model's checkpoints at various stages in the training we can examine how our results in Experiment 1 emerge throughout training. Further, since the model is open access we can also examine the different hidden-layers of the model.

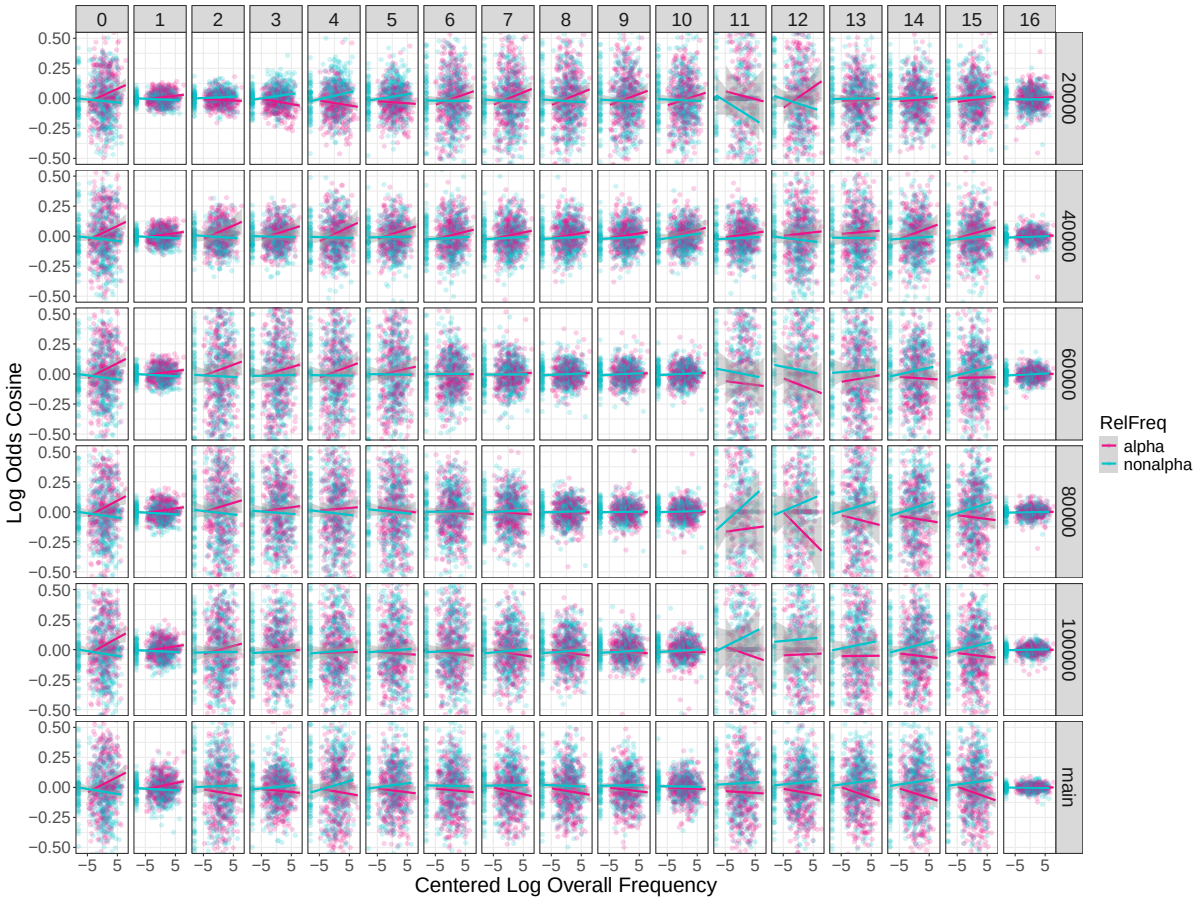
Methods

The methods in Experiment 2 were almost identical to those used in Experiment 1, with two main exceptions: first, rather than examining several different large language models, we instead examined a single large language model: OlmO 1B. OlmO 1B has released checkpoints at different stages in learning. As such, we can examine the representations of binomials at different stages of learning. Second, we also examined the representations at each hidden layer in the model in order to examine how the representation changes across layers.

For the present study, we examine the embeddings for our sentences from Experiment 1 at each hidden layer at multiple different steps in the training. In addition to examining the model after being trained, we also examine the embeddings after being trained for 20000 (84B tokens), 40000 (168B tokens), 60000 (252B tokens), 80000 (336B tokens), and 100000 (419B tokens) steps.

Results

A visualization of the embeddings at different layers and different checkpoints is included below:



There are two trends that are notable. First, the earlier layers tend to display an opposite pattern from the later layers, with the more frequent embeddings being more similar to the compositional form. Second, the holistic representation of the frequent ordering diverges from the compositional representation at about the 80000th step (336B tokens). We will discuss the implications of both of these in the discussion section.

Discussion

Our results demonstrate that from early on in the training the frequency difference is reflected in the embeddings in the early layers. Interestingly, however, this is not reflected in the representation at later layers. Instead, the differences in representations emerge in later layers over time.

These results are consistent with the general idea that later layers encode more semantic information, since the pattern of the more frequent embeddings diverging is seen in the later layers, while the earlier layers show the opposite pattern.

Additionally, the embeddings seem to converge in the last layer, suggesting that the different representations may not be important for the task of next-word prediction.

Finally, the fact that the pattern emerges rather slowly (taking several hundred billion tokens of training) suggests that the model must experience the binomial quite a lot in order to learn a separate representation for it. This suggests that the model isn't simply learning a separate representation because the binomial occurs in different semantic contexts (because if that were the case we would see the pattern emerge quite early on), but because it occurs frequently. This is in line with usage-based theories that have argued that holistic representations in humans emerge as a function of usage (e.g., Bybee 2003).

Conclusion

The present study demonstrates that the semantic embeddings for the more frequent ordering of a given binomial become less similar to the compositional embeddings as a function of the overall frequency of the binomial. That is, the embeddings of the more frequent ordering of a high-frequency binomial (e.g., *bread and butter*) are less similar to the compositional embeddings than the less frequent ordering's embedidngs (e.g., *butter and bread*) are. Another way to frame these results is that the same form (i.e., the same words) can give rise to quantitatively (but systematically) different representations in large language models and this varies depending on the overall frequency of that form (in either order).

It may not seem particularly surprising that the more frequent form diverges in semantic representation from the compositional form. After all, by definition a large language model has more experience with the more frequent form, which means the embeddings are being updated more often for the more frequent ordering. This in turn creates more opportunities for those embeddings to diverge from the compositional embeddings. However, what is interesting is how this effect emerges over time: early on in the training, the embeddings for the more frequent form are more similar to the compositional form across both earlier and later layers. Further, as training continues this stays the case for early layers, but undergoes a reversal in later layers.

One possible explanation for our results is that the more frequent form may be occurring in particularly different contexts from the compositional and less frequent forms (e.g., perhaps they are more idiomatic, such as *black and white*[4]). However, if this were the case then we would expect to see the embeddings for the frequent form to diverge from the embeddings of the compositional form quite early in training. Instead, however, we actually see the opposite early in the training: the embeddings for the more frequent form are *more* similar to those of the compositional form and it takes time for these embeddings to diverge.

[4]Although all of our sentences were sentences that encouraged a compositional reading of our binomials, and very few of our binomials had a particularly idiomatic meaning to begin with.

Another possibility is that early in training for high-frequency binomials, the large language model's experience with the individual words may largely overlap with the large language model's experience with the frequent form of the binomial (e.g., the model's experience with contexts containing the binomial *bread and butter* are also contributing to the large language model's experience with the individual words). Thus, initially these embeddings may be similar until the large language model experiences enough data to learn different representations. As the model experiences more sentence contexts with the binomial, the representation for the more frequent ordering has more opportunities to diverge from the representation of the individual words. This process explains why the same form can give rise to different representations.

Finally, our results can also be considered predictions for how humans may learn representations. Future work would do well to examine whether it is also the case that the semantic representations for the more frequent ordering of high-frequency binomials diverge more from the compositional representations in humans. Our results also make predictions about the timescale of learning: for young children, the pattern of results may actually be the opposite from adults, since at earlier checkpoints in our model the embeddings for the more frequent ordering of high-frequency binomials were more similar to the compositional embeddings.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency." In, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Bybee, Joan. 2003. *Phonology and Language Use*. Vol. 94. Cambridge University Press. https://books.google.com/books?hl=en&lr=&id=IM2rnd-Hw14C&oi=fnd&pg=PP15&dq=bybee+2003+phonology&ots=oIS3kjcs8C&sig=xTECvGoj0XT_p39S36F9vIzQE5E.

Bybee, Joan, and Joanne Scheibman. 1999. "The Effect of Usage on Degrees of Constituency: The Reduction of Don't in English." *Linguistics* 37 (4). https://doi.org/10.1515/ling.37.4.575.

Groeneveld, Dirk, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, et al. 2024. "Olmo: Accelerating the Science of Language Models." *arXiv Preprint arXiv:2402.00838*.

Haley, Coleman. 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*, 333–41.

Houghton, Zachary, Misaki Kato, Melissa Baese-Berk, and Charlotte Vaughn. 2024. "Task-Dependent Consequences of Disfluency in Perception of Native and Non-Native Speech." *Applied Psycholinguistics*, January, 1–17. https://doi.org/10.1017/S0142716423000486.

Lasri, Karim, Olga Seminck, Alessandro Lenci, and Thierry Poibeau. 2022. "Subject Verb Agreement Error Patterns in Meaningless Sentences: Humans Vs. BERT." *arXiv Preprint arXiv:2209.10538*.

Li, Bingzhi, and Guillaume Wisniewski. 2021. "Are Neural Networks Extracting Linguistic Properties or Memorizing Training Data? An Observation with a Multilingual Probe for Predicting Tense." In. https://shs.hal.science/halshs-03197072/.

Li, Bingzhi, Guillaume Wisniewski, and Benoît Crabbé. 2023. "Assessing the Capacity of Transformer to Abstract Syntactic Representations: A Contrastive Analysis Based on Long-Distance Agreement." *Transactions of the Association for Computational Linguistics* 11 (January): 18–33. https://doi.org/10.1162/tacl_a_00531.

Lin, Yuri, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. "Syntactic Annotations for the Google Books Ngram Corpus." In, 169174. https://aclanthology.org/P12-3029.pdf.

McCoy, R Thomas, 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: 652–70.

Misra, Kanishka, and Kyle Mahowald. n.d. "Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs." https://doi.org/10.48550/arXiv.2403.19827.

Morgan, Emily, and Roger Levy. 2015. "Modeling Idiosyncratic Preferences : How Generative Knowledge and Expression Frequency Jointly Determine Language Structure," 1649–54.

———. 2016a. "Abstract Knowledge Versus Direct Experience in Processing of Binomial Expressions." *Cognition* 157: 384–402. https://doi.org/10.1016/j.cognition.2016.09.011.

———. 2016b. "Frequency-Dependent Regularization in Iterated Learning." *The Evolution*

of Language: Proceedings of the 11th International Conference.

———. 2024. "Productive Knowledge and Item-Specific Knowledge Trade Off as a Function of Frequency in Multiword Expression Processing." *Language* 100 (4): e195–224. https://muse.jhu.edu/pub/24/article/947046.

O'Donnell, Timothy J. 2016. "Productivity and Reuse in Language." *Productivity and Reuse in Language*, 1613–18. https://doi.org/10.7551/mitpress/9780262028844.001.0001.

Piantadosi, Steven T. 2023. "Modern Language Models Refute Chomsky's Approach to Language." *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, 353–414.

Piantadosi, Steven T, and Felix Hill. 2022. "Meaning Without Reference in Large Language Models." *arXiv Preprint arXiv:2208.02957*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1 (8): 9. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

Siyanova-Chanturia, Anna, Kathy Conklin, and Walter J. B. van Heuven. 2011. "Seeing a Phrase " Time and Again" Matters: The Role of Phrasal Frequency in the Processing of Multiword Sequences." *Journal of Experimental Psychology: Learning Memory and Cognition* 37 (3): 776–84. https://doi.org/10.1037/a0022531.

Stemberger, Joseph Paul, and Brian MacWhinney. 1986. "Frequency and the Lexical Storage of Regularly Inflected Forms." *Memory & Cognition* 14 (1): 17–26. https://doi.org/10.3758/BF03209225.

Stemberger, Joseph P., and Brian MacWhinney. 2004. "Are Inflected Forms Stored in the Lexicon." *Morphology: Critical Concepts in Linguistics* 6: 107122. https://books.google.com/books?hl=en&lr=&id=bGl0aKBld3cC&oi=fnd&pg=PA107&dq=stemberger+2004+inflected&ots=RdvzVaC_NS&sig=0DJV8gUVaoZv_COZqcLXOu5_evU.

Tenney, I. 2019. "BERT Rediscovers the Classical NLP Pipeline." *arXiv Preprint arXiv:1905.05950*.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. n.d. "Llama 2: Open Foundation and Fine-Tuned Chat Models."

Yao, Qing, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. n.d. "Both Direct and Indirect Evidence Contribute to Dative Alternation Preferences in Language Models." https://doi.org/10.48550/arXiv.2503.20850.