LLM Storage Writeup

Zachary Nicholas Houghton

Introduction

Methods

Dataset

Semantic Embeddings

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

For each LLM we gathered two key measurements: 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. Since LLMs generate an embedding for each word, we computed the mean of these embeddings to represent the semantic embedding of the entire binomial (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 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 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

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

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 the quotient of this value for the alphabetical form of the binomial and the nonalphabetical form of the binomial and subsequently logged this value (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 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 tradeoff 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

The results of our models for each LLM are presented below. For all of our models, following (Houghton et al. 2024) we 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 can present a more nuanced picture of our results.

GPT-2

Our mixed-effects model is presented below in Table 1 and visualized in Figure 1. There was a meaningful main-effect of relative frequency ($\beta = -0.035$), suggesting that as relative frequency increases (i.e., for binomials with an increasing preference for the alphabetical form), the holistic embeddings for the alphabetical form are *less* similar to the compositional embeddings than the nonalphabetical holistic embeddings are. Further, there was a meaningful interaction effect between overall frequency and relative frequency ($\beta = -0.005$), suggesting that for high-frequency binomials there is a stronger effect of relative frequency than for low-frequency binomials. Specifically, for high-frequency binomials, those with a larger relative frequency have a lower $LogCosSim$ value. That is, for high-frequency binomials, the more preferred ordering's holistic embeddings are less similar to the compositional embeddings than the less preferred ordering's holistic embeddings are.

Our results suggest that GPT-2 is learning separate representations for high-frequency binomials, but may not be learning separate representations for low-frequency binomials.

Table 1: Model results for our Bayesian mixed-effects model for GPT-2.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	0.001	0.002	-0.003	0.005	72.4375
OverallFreq	-0.001	0.001	-0.002	0.001	14.2250
RelFreq	-0.035	0.006	-0.047	-0.022	0.0000
OverallFreq:RelFreq	-0.005	0.001	-0.008	-0.002	0.0375

GPT-2 XL

Test

Test

Test

Test

Our mixed-effects model is presented below in Table 2 and visualized in Figure 2. We found a meaningful main-effect for overall frequency ($\beta = 0.002$), though this seems to be driven largely by our interaction effect. We also found a meaningful interaction effect ($\beta = 0.003$) between relative frequency and overall frequency, suggesting that for higher-frequency binomials, the holistic embeddings for the alphabetical form were *more* similar to the compositional form than the holistic embeddings for the nonalphabetical form were.

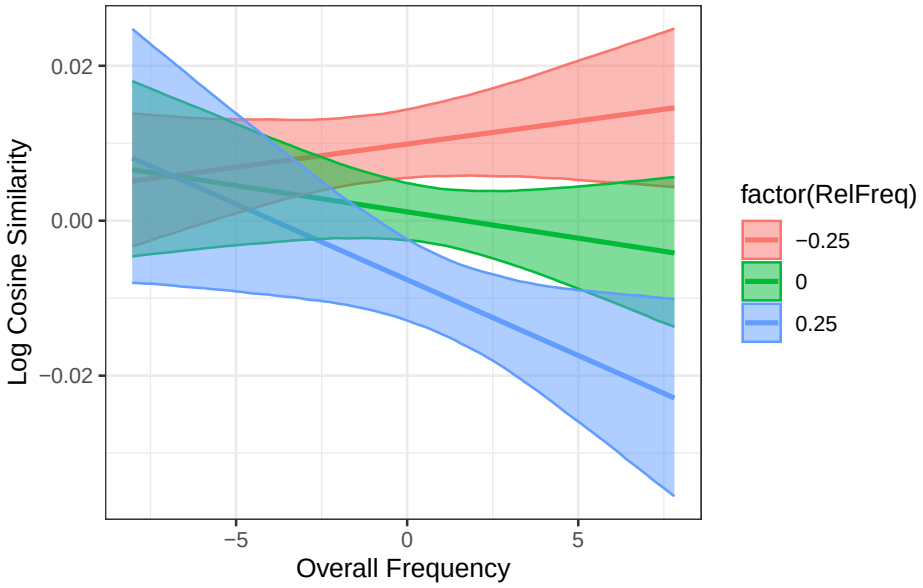

Figure 1: Visualization of our model predictions for GPT-2 at relative frequency values of -0.25, 0, and 0.25

Table 2: Model results for our Bayesian mixed-effects model for GPT-2 XL.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-0.002	0.002	-0.005	0.002	14.775
OverallFreq	0.002	0.001	0.000	0.003	99.675
RelFreq	0.003	0.006	-0.009	0.015	68.450
OverallFreq:RelFreq	0.003	0.001	0.000	0.006	97.825

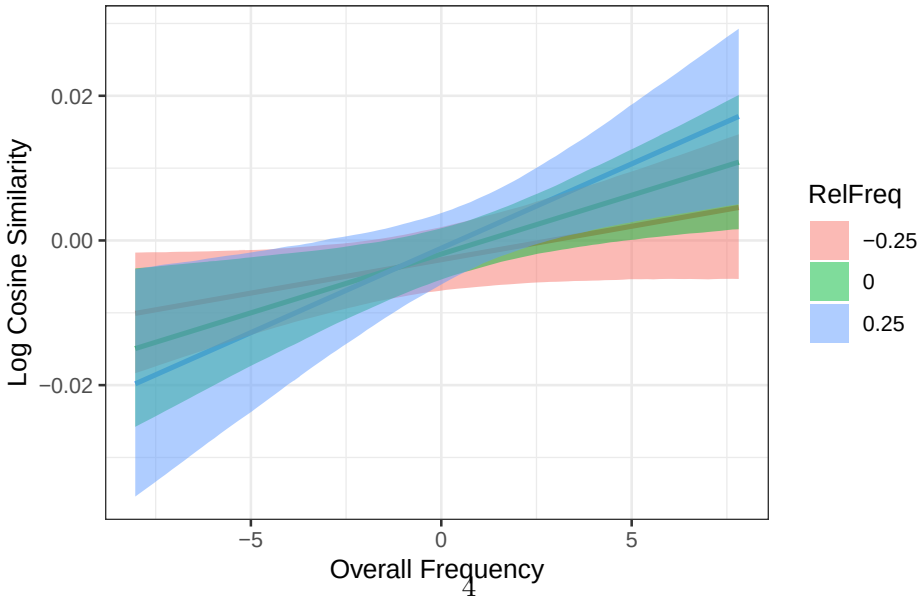

Figure 2: Visualization of our model predictions for GPT-2 XL at relative frequency values of -0.25, 0, and 0.25

Olmo-1B

Our results suggest that Olmo-1B, similar to GPT2, is learning separate representations for high-frequency binomials, but may not be learning separate representations for low-frequency binomials.

Table 3: Model results for our Bayesian mixed-effects model for Olmo 1B.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	0.001	0.008	-0.015	0.018	56.375
OverallFreq	-0.001	0.003	-0.007	0.004	33.000
RelFreq	-0.152	0.028	-0.207	-0.099	0.000
OverallFreq:RelFreq	-0.017	0.006	-0.030	-0.005	0.425

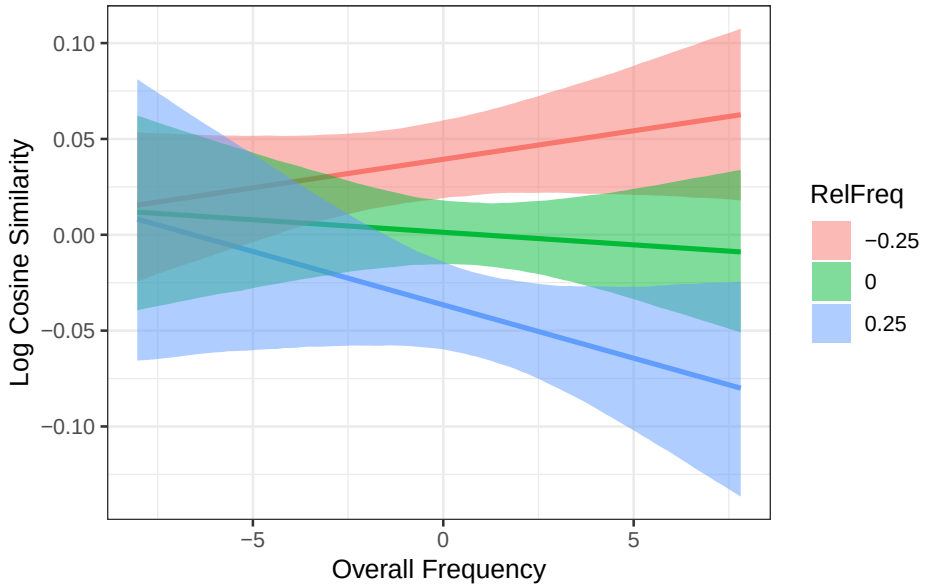

Figure 3: Visualization of our model predictions for Olmo 1B at relative frequency values of -0.25, 0, and 0.25

Olmo-7B

Our mixed-effects model is presented below in Table 4 and visualized in Figure 4. We found a meaningful main-effect of relative frequency ($\beta = -0.151$), suggesting that for binomials with a stronger preference for the alphabetical form, the holistic embeddings of the alphabetical form were less similar to the compositional form than the holistic embeddings of the nonalphabetical form were. We also found a meaningful interaction effect ($\beta = -0.017$), suggesting that for lower-frequency binomials there is not much of a difference between the alphabetical

and nonalphabetical forms in terms of their semantic embeddings, however for more-frequent binomials that occur more in the alphabetical form, the holistic embeddings for the alphabetical form are *less* similar to the compositional form than the holistic embeddings for the nonalphabetical form.

Our results suggest that Olmo-7B, similar to Olmo-1B and GPT2, is learning separate representations for high-frequency binomials, but may not be learning separate representations for low-frequency binomials.

Table 4: Model results for our Bayesian mixed-effects model for Olmo 7B.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	0.001	0.009	-0.015	0.018	55.775
OverallFreq	-0.001	0.003	-0.007	0.004	32.450
RelFreq	-0.151	0.029	-0.208	-0.094	0.000
OverallFreq:RelFreq	-0.017	0.007	-0.030	-0.004	0.575

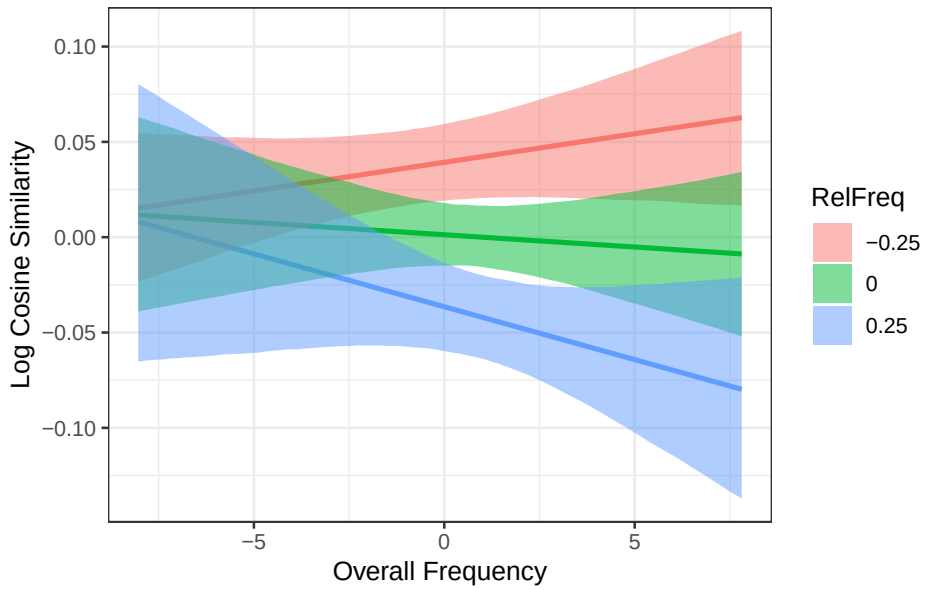

Figure 4: Visualization of our model predictions for Olmo 7B at relative frequency values of -0.25, 0, and 0.25

Llama2-7B

Our mixed-effects model is presented below in Table 5 and visualized in Figure 5. While the credible interval for the interaction effect crosses zero, over 96% of the posterior samples were

less than zero, suggesting that there is a meaningful interaction effect. The results suggest that for lower-frequency binomials there is not much of a difference between the alphabetical and nonalphabetical forms in terms of their semantic embeddings, however for more-frequent binomials that occur more in the alphabetical form, the holistic embeddings for the alphabetical form are *less* similar to the compositional form than the holistic embeddings for the nonalphabetical form.

Table 5: Model results for our Bayesian mixed-effects model for Llama2 7B.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-0.004	0.003	-0.009	0.002	9.2500
OverallFreq	-0.002	0.001	-0.003	0.000	3.4000
RelFreq	-0.013	0.009	-0.031	0.006	8.0625
OverallFreq:RelFreq	-0.004	0.002	-0.008	0.000	3.1750

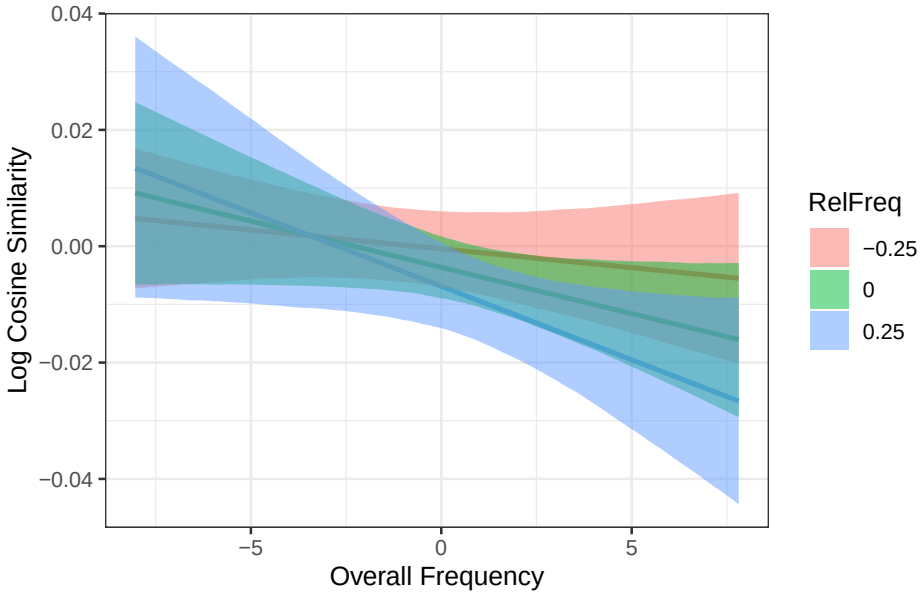

Figure 5: Visualization of our model predictions for Llama2 7B at relative frequency values of -0.25, 0, and 0.25

Discussion

Groeneveld, Dirk, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, et al. n.d. "OLMo: Accelerating the Science of Language Models."

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.

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.

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."