Measuring Meaning Composition in the Human Brain with Composition Scores from Large Language Models

Anonymous ACL submission

Abstract

The process of meaning composition, wherein smaller units like morphemes or words combine to form the meaning of phrases and sentences, is essential for human sentence comprehension. Despite extensive neurolinguistic research into the brain regions involved in meaning composition, a computational metric to quantify the extent of composition is still lacking. Drawing on the key-value memory interpretation of transformer feed-forward network blocks, we introduce the Composition Score, a novel model-based metric designed to quantify the degree of meaning composition during sentence comprehension. Experimental findings show that this metric correlates with brain clusters associated with word frequency, structural processing, and general sensitivity to words, suggesting the multifaceted nature of meaning composition during human sentence comprehension. [1]

1 Introduction

When encountering words such as "milk" and "pudding", the human mind effortlessly combines them to form a complex concept, such as a milk-flavored pudding. This combinatory process is a fundamental aspect of human language comprehension and production, enabling us to generate an infinite array of meanings from a finite set of words. Despite extensive neurolinguistic research into the localization of meaning composition in the human brain (Bemis and Pylkkänen, 2011, 2013; Blanco-Elorrieta et al., 2018; Flick and Pylkkänen, 2020; Li and Pylkkänen, 2021; Zhang and Pylkkänen, 2015; Li et al., 2024), understanding the detailed mechanism of how a complex meaning is constructed from its components and how it is processed by the human brain has become a challenging problem. One of the primary difficulties lies in the absence of a suitable computational metric to quantify

[1] Our code and data will be released upon acceptance.

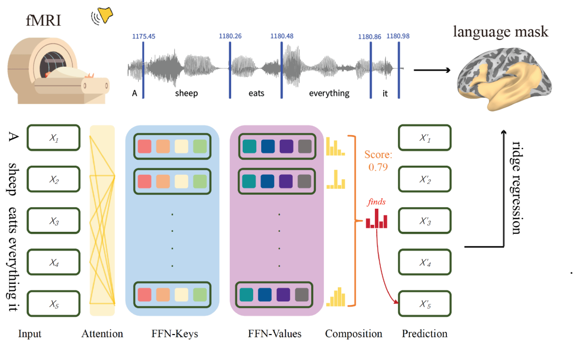

Figure 1: Comparing Composition Scores with fMRI data during naturalistic listening comprehension.

the extent of meaning composition. This absence significantly complicates quantitative analyses of meaning composition in the human brain.

Recent advancements in Large Language Models (LLMs) offer promising insights into this problem. By training on large-scale natural language corpora and aligning with human preferences, these computational models achieve unprecedented levels of proficiency in understanding and generating natural languages (OpenAI et al., 2023; Anil et al., 2023; Touvron et al., 2023). In addition to their high performance, studies have shown that their internal states correlate with human behavioral and neural data (Schrimpf et al., 2021; Caucheteux et al., 2022), suggesting shared principles between their algorithms and the human brain. Given this background, it is natural to inquire whether we can develop a computational metric to quantify meaning composition from the internal states of LLMs.

Motivated by this inquiry, our study introduces a novel model-based metric, the Composition Score, to evaluate meaning composition in the human brain. Leveraging the key-value memory interpretation of the Feed-Forward Network (FFN) modules in the transformer model (Geva et al., 2021, 2022), this metric computes the composition of memory-induced vocabulary distributions within the FFN blocks given an input prefix, thereby reflecting the degree of meaning composition of each word. To assess its validity, we examine the patterns of Composition Scores using the novel "The Little Prince" in English and compare them with other control variables such as word frequency and syntactic node count based on top-down, bottom-up, and left-corner parsing. Additionally, we correlate Composition Scores with an openly available fMRI dataset where participants listened to "The Little Prince" in the scanner (Li et al., 2022). Our findings reveal that:

- The Composition Score exhibits partial correlation with word frequency and syntactic node counts but reveals more intricate patterns;
- The Composition Score is associated with a broader brain cluster and exhibits a higher regression score with the fMRI data compared to the control variables;
- Brain regions associated with the Composition Score encompass those underlying word frequency, structural processing, and general sensitivity to words, indicating the multifaceted nature of meaning composition.

2 Related Work

2.1 Meaning composition in LLMs

Despite considerable efforts in interpreting transformer models and Large Language Models (LLMs), e.g. Hewitt and Manning, 2019; Clark et al., 2019; Voita et al., 2023, prior research has not extensively focused on meaning composition in LLMs. In their groundbreaking work interpreting the Feed-Forward Network (FFN) block as key-value memory, Geva et al. (2021) noted that the block engages in "memory composition" and quantified the degree of composition by examining the overlap between neuronal predictions and block predictions. Building on this, Geva et al. (2022) and Voita et al. (2023) proposed that the FFN block makes predictions by amplifying and suppressing concepts in the vocabulary space, akin to composing meaning. Inspired by this interpretation, we design the Composition Score to link the meaning composition in models and the human brain.

2.2 Meaning composition in the human brain

The process of meaning composition in the human brain has been localized to regions in the left temporal lobe. Studies have found that phrases like "red boat" trigger increased activity in the left anterior temporal lobe (LATL) compared to non-compositional word lists (Bemis and Pylkkänen, 2011, 2013), indicating LATL's involvement in conceptual combination. This effect is consistent across different word orders and languages (Westerlund et al., 2015), including American Sign Language (Blanco-Elorrieta et al., 2018).

Although the LATL remains the most consistently implicated locus for composition with the highest replication rates, recent evidence suggests a role for the surrounding temporal cortex as well. Investigations into the functional intricacies of the LATL have unveiled its conceptual, non-syntactic functions (Bemis and Pylkkänen, 2013; Li and Pylkkänen, 2021; Parrish and Pylkkänen, 2022; Zhang and Pylkkänen, 2015). For instance, the LATL can integrate concepts such as "boat red" even without explicit syntactic combination (Bemis and Pylkkänen, 2013; Parrish and Pylkkänen, 2022). Conversely, the posterior temporal cortex exhibits greater sensitivity to syntactic structures (Flick and Pylkkänen, 2020; Hagoort, 2005; Lyu et al., 2019; Matchin et al., 2019; Matchin and Hickok, 2020; Li and Pylkkänen, 2021). As outlined in Pylkkänen (2019), composition may entail syntactic, logico-semantic, and conceptual subroutines, engaging multiple areas across the temporal, parietal, and frontal cortex beyond the LATL (see Pylkkänen, 2019 for a review).

2.3 Correlating model predictions with the human brain

Previous studies comparing both symbolic models and LLMs to the human brain have revealed some shared principles between the two systems (e.g., Brennan et al., 2016; Caucheteux and King, 2022; Caucheteux et al., 2022; Goldstein et al., 2022; Nelson et al., 2017; Schrimpf et al., 2021; Toneva et al., 2022; Antonello et al., 2023; Gao et al., 2023). For example, Nelson et al. (2017) correlated syntactic complexity under different parsing strategies with the intracranial electrophysiological signals and found that the left-corner and bottom-up strategies fit the left temporal data better than the most eager top-down strategy; Goldstein et al. (2022) and Caucheteux et al. (2022) both showed that the human brain and the deep learning language models share the computational principles of predicting the next word as they process the same natural narrative. Toneva et al. (2022) constructed a computational representation for "supra-word meaning". They modeled composed meaning by regressing word embeddings from its context embeddings in ELMo (Peters et al., 2018), and found significant LATL and LPTL activity correlating with this metric. Antonello et al. (2023) and Gao et al. (2023) examined the scaling law in the correlation between model states (e.g. hidden states, attention matrices) and human neural and behavioral data.

3 Methods

3.1 Composition Scores from LLMs

The Composition Score proposed in this paper quantifies the compositionality of key-value memory stored in the FFN blocks of LLMs, building upon the key-value memory interpretation of the FFN blocks. We begin by formally describing the key-value memory hypothesis and subsequently introduce the definition of the Composition Score.

3.1.1 The key-value memory interpretation

Geva et al. (2021) first proposed the key-value memory interpretation of FFN blocks in transformer models. An FFN block (e.g., for transformer layer l) can be expressed as:

$$\mathrm{FF}^l(\mathbf{x}) = f(\mathbf{x} \cdot K^{l\top}) \cdot V^l$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input vector, $K, V \in \mathbb{R}^{d_m}$ are the two linear layers inside the FFN block, and f is the activation function. This formulation can be viewed as a generalized expression of a neural memory (Sukhbaatar et al., 2015):

$$\mathrm{MN}(\mathbf{x}) = \mathrm{softmax}(\mathbf{x} \cdot K^\top) \cdot V$$

Consequently, the first linear layer K^l corresponds to the "keys" matrix in the neural memory, each row of which (also referred to as a "neuron") is a key vector that triggers activation of a certain memory; and V^l corresponds to the "values" matrix, each row of which is a memory entry $\mathbf{v}_i^l$ that can affect the next-token prediction. The activation, $\mathbf{m}^l = f(\mathbf{x} \cdot K^{l\top})$, can then be viewed as a vector that contains the unnormalized coefficient of each memory entry in this FFN block. As a result, the output of the FFN block is a weighted mixture of memory values.

Geva et al. (2021, 2022) then translated the aforementioned vector-space analysis into human-readable representations, where $\mathbf{x}$, the vector representation of a word w_j in a sentence, corresponds to the input prefix $w_1, ..., w_j$. Additionally, the memory value of the i-th neuron $\mathbf{v}_i$ can be mapped to a vocabulary distribution $\mathbf{p}_i^l$ by the output embedding matrix E using:

$$\mathbf{p}_i^l = \mathrm{softmax}(\mathbf{v}_i^l \cdot E)$$

This same mapping can also be applied to the FFN output. In this context, the FFN block receives a sentence prefix, activates its stored memory accordingly, and then combines the predicted next-token distribution encoded by each neuron to produce the final prediction.

3.1.2 Calculating Composition Score

The key idea of the Composition Score is to interpret the memory combination process described above as meaning composition, as manifested by the predicted vocabulary distributions. Given the predicted vocabulary distributions $\mathbf{p}_1^l, ..., \mathbf{p}_{d_m}^l$ of each neuron, and the final predicted distribution $\mathbf{p}^l$ of the FFN block, we first calculate the Jensen-Shannon distances (the square root of Jensen-Shannon divergence) between them:

$$\begin{aligned}\mathrm{dist}(\mathbf{p}_i^l, \mathbf{p}^l) &= D_{\mathrm{JS}}^{\frac{1}{2}}(\mathbf{p}_i^l \| \mathbf{p}^l) \\ &= \left[\frac{1}{2} D_{\mathrm{KL}}(\mathbf{p}_i^l \| \mathbf{p}_m^l) + \frac{1}{2} D_{\mathrm{KL}}(\mathbf{p}^l \| \mathbf{p}_m^l)\right]^{\frac{1}{2}}\end{aligned}$$

where $D_{\mathrm{KL}}(\cdot \| \cdot)$ is the Kullback–Leibler divergence between two distributions, and $\mathbf{p}_m^l = \frac{1}{2}(\mathbf{p}_i^l + \mathbf{p}^l)$. This quantifies the proximity of the final prediction to the individual memory values. If the distances are approximately equal across all the neurons in the block, we interpret the output as highly composed. Conversely, if the distance is close to zero for one or two neurons and significantly larger for others, we perceive the output as less composed. Hence, we define the Composition Score as:

$$S_{\mathrm{comp}}^l = \frac{\min_{1 \le i \le d_m} \mathrm{dist}(\mathbf{p}_i^l, \mathbf{p}^l)}{\max_{1 \le j \le d_m} \mathrm{dist}(\mathbf{p}_j^l, \mathbf{p}^l)}$$

The score ranges from 0 to 1, with a high score (close to 1) indicating that the largest distance is roughly equivalent to the smallest one, and vice versa. Conceptually, the Composition Score quantifies the degree of memory or meaning compositionality when predicting the next token, based on the input prefix. Since there is one score from each transformer layer, we incorporate the Composition Scores from all layers for analysis.

3.1.3 Activation-based approximation

Because computing the Composition Score is highly resource-intensive, we employed an approximation method to accelerate the computation: instead of considering all d_m neurons in layer l when calculating S^l_{comp}, we only include a fixed number d'_m of neurons. Specifically, we select neurons whose sum of absolute activation values comprises the majority of the total values. This approach is supported by the sparse activation phenomenon observed in the FFN neurons in LLMs (Voita et al., 2023), where most FFN neurons are either not activated or weakly activated during forward computation, with only a small fraction being strongly activated. It is primarily these latter neurons that contribute significantly to the meaning composition process in the FFN blocks.

To select an appropriate value for d'_m, we run the tested LLMs on the C4 validation corpus (Raffel et al., 2019) and gather their numbers of neurons (referred to as their majority k's) with the highest absolute activation values, which collectively contribute to over half of the total absolute activation. Subsequently, we set d'_m to a value significantly larger than the majority k's of all models under consideration. The approximated Composition Score is then calculated as:

$$S'^l_{\text{comp}} = \frac{\min_{1\leq i\leq d'_m} \text{dist}(\mathbf{p}^l_i, \mathbf{p}^l)}{\max_{1\leq j\leq d'_m} \text{dist}(\mathbf{p}^l_j, \mathbf{p}^l)}$$

We find the majority k is 1744.49 for LLaMA2-base, and 1754.14 for LLaMA2-chat. Therefore, we set d'_m to 3000 to cover the majority k's of both.

Figure 3 displays the averaged Composition Score of each layer of the LLaMA2 models alongside a randomly initialized LLaMA2-7B model. It can be seen that both the LLaMA2-base and LLaMA2-chat models exhibit a similar pattern, with the mean Composition Score increasing in the first 6 layers and plateauing thereafter. This result indicates that, as the layer number goes up, the degree of composition becomes higher. This is predictable as the input vector $\mathbf{x}$ in the higher layers is integrated with more contextual information, which makes it harder to find close matches in the neural memory. In contrast, the Composition Score for the randomly initialized model remains constant around 1.

As there is minimal difference between the results obtained from the two LLaMA2 models in all subsequent experiments, we present outcomes solely from the LLaMA2-chat model in the main text. For results pertaining to the LLaMA2-base model, please consult Appendix B.

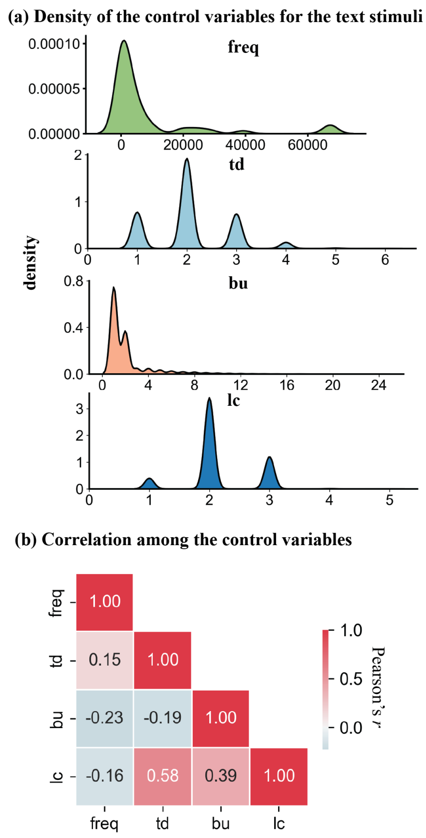

Figure 2: (a) Density plot of word frequency, node counts based on the top-down, bottom-up and left-corner node counts. (b) Correlation matrix among the 4 control variables.

3.2 Control variables

In addition to the Composition Score obtained from the LLMs, we incorporated five other control variables: Word rate, word frequency, and syntactic node counts derived from top-down, bottom-up, and left-corner parsing strategies. These variables have demonstrated correlations with notable brain clusters within the language network and provide a baseline for comparison with our Composition Score metric. Figure 2 shows the density and correlation matrix between word frequency and node count based on three parsing strategies.

Word rate. Word rate is a binary regressor that marks 1 at the offset of each word in the audiobook. It signifies an individual's overall responsiveness to words as opposed to other stimuli and has been associated with a widespread left temporal-frontal network within the language regions (Li et al., 2022).

Word frequency. We also included the log-transformed unigram frequency of each word, estimated using the Google ngrams Version 2012070129 [2] and the SUBTLEX corpora for Chinese (Cai and Brysbaert, 2010). Prior research on frequency effects has identified activity in the middle temporal lobe (e.g., Embick et al., 2001; Simon et al., 2012).

Node counts. Node count refers to the number of parsing steps between consecutive words according to a parsing strategy. This concept is associated with certain aspects of Yngve's (1960) Depth hypothesis (see also Frazier, 1985). Different parsing strategies yield varied predictions regarding the processing effort required for a given word. A top-down parser begins with a mother node and establishes phrase structures before validating them against the input string. Conversely, a bottom-up parser initiates with the first terminal word and verifies all evidence before applying a phrase structure rule. A left-corner parser combines elements of both top-down and bottom-up approaches, implementing a grammatical rule upon encountering the very first symbol on the right-hand side of the rule (Hale, 2014). We computed CFG-based node counts for the text stimuli using these three parsing strategies.

Prior research has shown significant left temporal and frontal activity for the left-corner and the bottom-up parsing strategies (Nelson et al., 2017), supporting bottom-up and/or left-corner parsing as tentative models of how human subjects process sentence structures.

3.3 Aligning Composition Scores and control variables with fMRI data

First-level regression. The Composition Score for each word, derived from each of the 32 hidden layers of the LLaMA2 models, was initially convolved with the canonical hemodynamic response function (HRF). Subsequently, two ridge regressions were conducted for each subject using the 32 Composition Scores from the two LLMs to predict the fMRI timecourses from each vertex within a left-lateralized language mask. The language mask (see the pink region in Figure 7) covered regions including the whole left temporal lobe, the left inferior frontal gyrus (LIFG; defined as the combination of BAs 44 and 45), the left ventromedial prefrontal cortex (LvmPFC; defined as BA11), the left angular gyrus (LAG; defined as BA39) and the left supramarginal gyrus (LSMA; defined as BA 40). The left AG and vmPFC have also been implicated in previous literature on conceptual combination (Bemis and Pylkkänen, 2011; Price et al., 2015) and the LIFG and the LMTG have been suggested to underlie syntactic combination (Flick and Pylkkänen, 2020; Hagoort, 2005; Lyu et al., 2019; Matchin et al., 2019; Matchin and Hickok, 2020). The optimal penalty term α of the ridge regressions was determined by automatic cross-validation.

Similarly, the five control variables, time-aligned to the offset of each word, were first convolved with the HRF and then regressed against each subject's fMRI timecourse of each vertex within the language mask using ordinary linear regression (OLS).

The regression scores R^2 for the Composition Scores and the control variables, obtained for each subject, were normalized by the noise ceiling, i.e., the Inter-Subject Correlation (ISC; Hasson et al., 2004) of the regression scores R^2_{ISC}. The R^2_{ISC} was computed as the mean regression score of all subjects, where the regressor is the mean fMRI signal of all subjects. The normalized regression scores were calculated as $\bar{R}^2 = R^2/R^2_{\text{ISC}}$. Figure 1 illustrates our model-brain comparison methods with an example sentence.

Statistical significance testing. At the group level, the β values for the control variables and the Composition Score at each layer of the two LLMs, averaged over subjects, underwent a one-sample one-tailed t-test with a cluster-based permutation test (Maris and Oostenveld, 2007) involving 10,000 permutations. Clusters were formed from statistics corresponding to a p-value less than 0.05, and only clusters spanning a minimum of 20 vertices were included in the analysis. These analyses were conducted using the Python packages MNE (v1.0.3) and Eelbrain (v0.39.8).

4 Experiment settings

4.1 Text stimuli

The text of the audiobook "The Little Prince" in English comprises 15,376 words and 1,499 sentences. The mean sentence length is 10.20, with a standard deviation of 6.94. Since the text is derived from an audiobook, the sentences lack punctuation. Consequently, we input the text data sentence by sentence into the LLMs to mitigate ambiguity.

[2] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

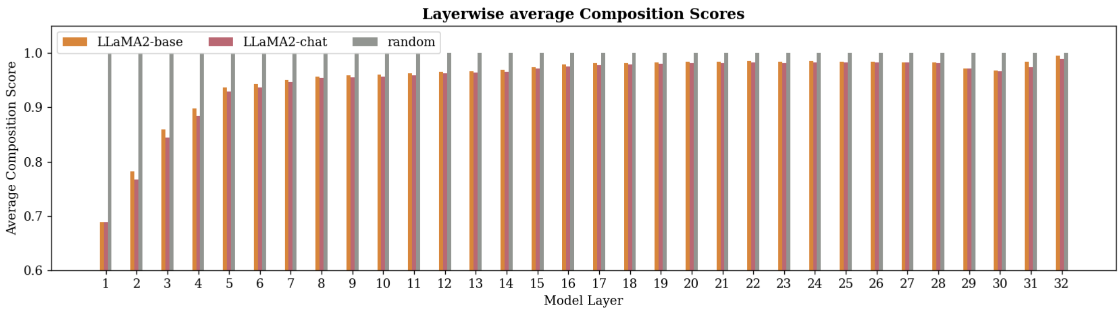

Figure 3: The average Composition Score of each layer of the LLaMA2 models and a randomly initialized model.

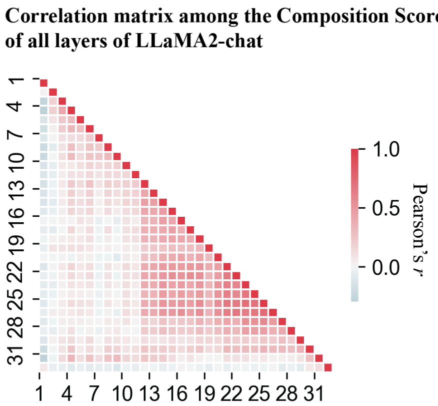

Figure 4: Correlation matrix among the 32 layers of LLaMA2-chat.

4.2 fMRI data

We use the fMRI recordings of the English subset of "The Little Prince" dataset (Li et al., 2022), a publicly available dataset containing the fMRI recordings of 49 English subjects (30 females, mean age=21.3 years, SD=3.6) listening to the audiobook "The Little Prince" in English for 94 minutes in total. The preprocessed volumetric data were projected onto a "fsaverage5" template surface (Fischl, 2012). The fMRI signals are z-scored across the time dimension for each participant, surface voxel and session independently.

4.3 Model

We use the widely-used open-source LLM, LLaMA2 (Touvron et al., 2023) in all our experiments. LLaMA2 comprises two versions: LLaMA2-base (pretrained on about 2.0T tokens in multiple languages) and LLaMA2-chat (the LLaMA2-base model fine-tuned with instructions in English), and we test both of the versions. To manage computational resources (see Appendix A), we employ the 7B-sized models.

4.4 Token-word alignment

To compare the LLM-based Composition Score of each subword token with the word frequency and syntactic node counts, we employ the following procedure for token-word alignment: Given a sentence with L words as $w_1, .., w_L$, when inputting the prefix $w_1, ..., w_k$ (up to the last subword token of w_k if it is split by the LLaMA2 tokenizer), the model state is aligned with the control variables of w_k, as well as the human fMRI recording corresponding to the offset of w_k (taking into account the delay and duration of BOLD signals). This alignment ensures that we compare the model state and the control variables given the same contextual input.

5 Results

5.1 Patterns of Composition Scores

Layerwise correlation. Given that the Composition Scores across different model layers exhibit different distributions, we hypothesize that they contain unique information regarding meaning composition. To validate this assumption, we compute the Pearson's r among the layerwise scores. The results are depicted in Figure 4 and Figure 8 (in Appendix B). It can be seen that in both the base and chat models, the layers form small correlated clusters, but the overall correlation among all layers is not high, with the highest absolute correlation coefficient reaching around 0.59.

Prefixes with high and low Composition Scores. To gain deeper insights into how the model assigns high and low Composition Scores under various input prefixes, we analyze prefixes with the highest and lowest Composition Scores in each layer. Table 1 presents examples of such prefixes with high and low Composition Scores across lower, middle, and higher layers.

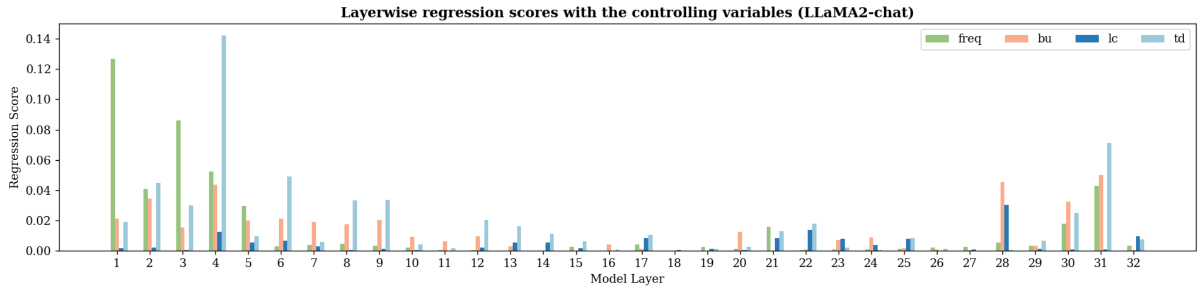

Figure 5: The regression scores R^2 between the Composition Scores from LLaMA2-chat and the control variables.

The lower layers exhibit clearer patterns. For example, in Layer 1, prefixes ending with common function words such as prepositions and conjunctions (e.g., "of", "by" etc.) tend to receive low Composition Scores, while those ending with the determiner "the" receive high Composition Scores. However, in Layer 3, these patterns appear to reverse, with some less common words like "boa constrictor" receiving high scores. In the higher layers, the patterns become less clear. One potential trend is that prefixes ending with specific words such as "able" tend to receive low scores.

We hypothesize that the varying patterns of Composition Scores across different layers may be attributed to the residual connection structure and the nature of model training. Due to the presence of residual connections, neural memories across different layers are somewhat parallel (Voita et al., 2023). As a result, a prefix may match the key-value memory in some layers but not in others, leading to distinct scores across layers. Moreover, in the language modeling task, the model must optimize its neural memory storage to better fit the training corpus. Consequently, both frequent and infrequent prefixes may be memorized, resulting in intricate memory composition patterns.

Composition Score vs. control variables. To investigate whether the Composition Scores contain information regarding word frequency or syntactic structure, we conduct regressions of the Composition Score for each word against their word frequency and the node counts based on the three parsing strategies. Figure 5 illustrates the regression scores R^2.

The R^2 scores reveal that the bottom and top layers exhibit higher R^2 scores with the control variables, particularly the log frequency and the node count from top-down parsing. However, the overall R^2 scores across layers are not notably high, suggesting the presence of additional information in the Composition Scores beyond word frequency and syntactic information.

5.2 fMRI results for the control variables

5.2.1 Regression scores

The normalized regression scores of the control variables on the fMRI data are shown in Table 2. Among the control variables, wordrate shows the highest maximum and mean R^2 scores over the significant brain clusters. Log-transformed word frequency and the node count based on left-corner parsing also show relatively higher regression scores.

5.2.2 Significant brain clusters

Word rate. Consistent with prior research (e.g., Li et al., 2022), we find a widespread left temporal-frontal network in the LIFG, the left anterior superior temporal gyrus (LaSTG) and the left posterior middle temporal gyrus (LpMTG) for wordrate (N vertices=948, t=2.99, p<0.0001), indicating a general sensitivity to words.

Word frequency. The log word frequency is associated with a cluster in the LSTG (N vertices=73, t=-2.33, p=0.02), suggesting that lower word frequency induces higher LSTG activity.

Node counts. We find a significant cluster in the LaSTG (N vertices = 217, t = -2.54, p = 0.0001) associated with the node counts based on the left-corner parsing strategy. No significant clusters are identified for the node counts based on top-down or bottom-up parsing. These results further corroborate prior findings (Nelson et al., 2017) suggesting that left-corner parsing may align more closely with human processing of hierarchical sentence structures. See Figure 6 for the significant brain clusters for wordrate, log-transformed word frequency and node counts based on left-corner parsing.

Layer	Prefixes with low Composition Scores	Prefixes with high Composition Scores
1	I was discouraged by the failure of → my the second time was eleven years ago by → an	thus I abandoned at the → age after grooming oneself in the → morning
3	then he added so you also come from the → sky little drinking water left that I had to fear the → worst	I then drew the inside of the boa → con(strictor) I am beginning to → understand
16	it would suffice to be able → to he should be able → for	I have seen them from close → up who are you asked → the
32	it would suffice to be able → to on what planet have I come down on asked → the	I would like to see → a I was very worried because → my

Table 1: Example prefixes with low and high Composition Scores in different layers of the LLaMA2-base model. The token after the right arrow (→) is the next token to predict in the text corpus.

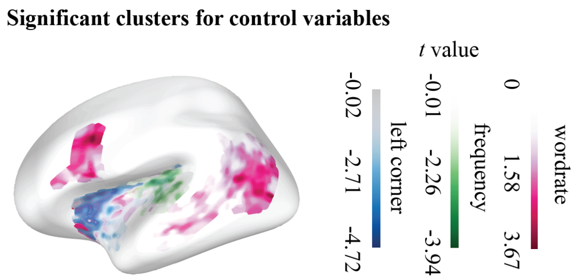

Figure 6: Significant brain clusters for the word rate, word frequency, and left corner parsing steps.

Regressor	Max	Mean
score-base	.1774	.0603
score-chat	.1361	.0462
word rate	.0697	.0229
bottom-up	.0005	.0002
top-down	.0037	.0011
left-corner	.0064	.0018
log freq	.0067	.0020

Table 2: Normalized regression scores R^2 on the fMRI data by the Composition Score and the control variables.

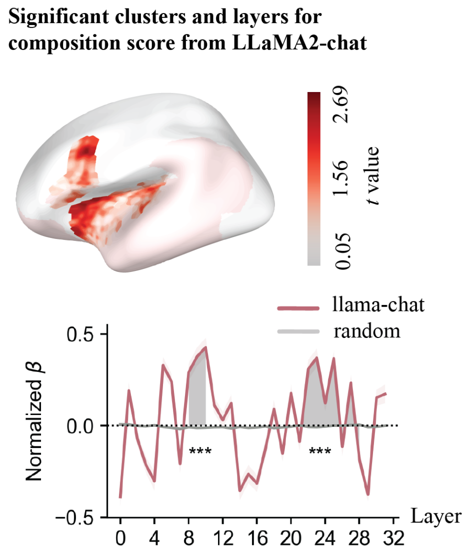

Figure 7: Significant brain clusters for Composition Scores and the significant layers from LLaMA2-chat. The light pink regions in the brain indicate the language mask. The orange and red lines depict the normalized β value for each layer of the two models. The grey lines depict the normalized β value for each layer of the random models. The shaded region indicates the significant layers. *** indicates p <0.001.

5.3 fMRI results for the Composition Scores

5.3.1 Regression scores

The normalized regression scores with the Composition Score exceed those with the control variables in both maximum and mean values. This indicates that the Composition Score provides a better fit to the human neural data compared to the control variables (refer to Figure 2).

5.3.2 Significant brain clusters

The Composition Scores derived from LLaMA2-chat exhibit a significant association with a cluster in the LIFG and the LaSTG (N vertices = 517, t = 3.52, p < 0.0001). These regions overlap with significant clusters for word rate, word frequency, and left-corner node count (refer to Figure 6), indicating the multifaceted nature of meaning composition during human sentence comprehension. Notably, the significant model layers include the middle layers 8-13 and the higher layers 21-25, suggesting that meaning composition in the human brain cannot solely be attributed to word frequency or memorization of specific words (for patterns of Composition Scores across layers, see Section 5.1).

6 Conclusion

In this paper, we introduce a novel model-based metric, the Composition Score, designed to quantify sentence-level meaning composition, and examine its correlation with human neural activity.

We identify several brain clusters significantly correlated with the Composition Score, including those associated with word frequency, syntactic structure, and general sensitivity to words. This suggests a multifaceted nature of meaning composition during human sentence comprehension.

Limitations

One key limitation of this study is that we have yet to fully comprehend the patterns of high and low Composition Scores for different sentences across different layers. We hypothesize that these patterns are related to the optimized memory efficiency of the LLMs, which may resemble memory mechanisms in the human brain.

Another limitation is that we solely employ the LLaMA2-7B models for the analysis, which may not guarantee the generalizability of our findings to other LLMs. However, given that the architecture of the FFN block remains largely consistent across LLMs, our method can be adapted to other models with minor modifications to the code. Additionally, our study solely focuses on English text stimuli, leaving the potential for further exploration in multilingual experiments.

Ethics Statement

The authors declare no competing interests. The fMRI dataset used in the analysis is publicly available and does not contain sensitive content, such as personal information. The adaptation and use of the fMRI dataset are conducted in accordance with its license. The model states of LLaMA2 are utilized solely for research purposes, aligning with its intended use.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Richard Antonello, Aditya Vaidya, and Alexander Huth. 2023. Scaling laws for language encoding models in fMRI. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Douglas K Bemis and Liina Pylkkänen. 2011. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*, 31(8):2801–2814.

Douglas K Bemis and Liina Pylkkänen. 2013. Flexible composition: Meg evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PloS one*, 8(9):e73949.

Esti Blanco-Elorrieta, Karen Emmorey, and Liina Pylkkänen. 2018. Language switching decomposed through meg and evidence from bimodal bilinguals. *Proceedings of the National Academy of Sciences*, 115(39):9708–9713.

Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94.

Qing Cai and Marc Brysbaert. 2010. Subtlex-ch: Chinese word and character frequencies based on film subtitles. *PloS one*, 5(6):e10729.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022. Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, 12(1):16327.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

David Embick, Martin Hackl, Jeannette Schaeffer, Melltem Kelepir, and Alec Marantz. 2001. A magnetoencephalographic component whose latency reflects lexical frequency. *Cognitive brain research*, 10(3):345–348.

Bruce Fischl. 2012. Freesurfer. *Neuroimage*, 62(2):774–781.

Graham Flick and Liina Pylkkänen. 2020. Isolating syntax in natural language: Meg evidence for an early contribution of left posterior temporal cortex. *Cortex*, 127:42–57.

L. Frazier. 1985. Syntactic complexity. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural language parsing: Psychological, computational, and theoretical perspectives*, pages 129–189. Cambridge, UK: Cambridge University Press.

Changjiang Gao, Shujian Huang, Jixing Li, and Jiajun Chen. 2023. Roles of scaling and instruction tuning in language perception: Model vs. human attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13042–13055, Singapore. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Peter Hagoort. 2005. On broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–423.

John T. Hale. 2014. *Automaton Theories of Human Sentence Comprehension*. CSLI Publications, Center for the Study of Language and Information, Stanford, California.

Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. 2004. Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022. Le petit prince multilingual naturalistic fmri corpus. *Scientific data*, 9(1):530.

Jixing Li, Marco Lai, and Liina Pylkkänen. 2024. Semantic composition in experimental and naturalistic paradigms. *Imaging Neuroscience*, 2:1–17.

Jixing Li and Liina Pylkkänen. 2021. Disentangling semantic composition and semantic association in the left temporal lobe. *Journal of Neuroscience*, 41(30):6526–6538.

Bingjiang Lyu, Hun S Choi, William D Marslen-Wilson, Alex Clarke, Billi Randall, and Lorraine K Tyler. 2019. Neural dynamics of semantic composition. *Proceedings of the National Academy of Sciences*, 116(42):21318–21327.

Eric Maris and Robert Oostenveld. 2007. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190.

William Matchin and Gregory Hickok. 2020. The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498.

William Matchin, Chia-Hsuan Liao, Phoebe Gaston, and Ellen Lau. 2019. Same words, different structures: An fmri investigation of argument relations and the angular gyrus. *Neuropsychologia*, 125:116–128.

Matthew J Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S Cash, Lionel Naccache, John T Hale, Christophe Pallier, et al. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave

Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Alicia Parrish and Liina Pylkkänen. 2022. Conceptual combination in the latl with and without syntactic composition. *Neurobiology of Language*, 3(1):46–66.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Amy R Price, Michael F Bonner, Jonathan E Peelle, and Murray Grossman. 2015. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, 35(7):3276–3284.

Liina Pylkkänen. 2019. The neural basis of combinatory syntax and semantics. *Science*, 366(6461):62–66.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Dylan Alexander Simon, Gwyneth Lewis, and Alec Marantz. 2012. Disambiguating form and lexical frequency effects in meg responses using homonyms. *Language and Cognitive Processes*, 27(2):275–287.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. 2022. Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science*, 2(11):745–757.

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.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional.

Masha Westerlund, Itamar Kastner, Meera Al Kaabi, and Liina Pylkkänen. 2015. The latl as locus of composition: Meg evidence from english and arabic. *Brain and Language*, 141:124–134.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Linmin Zhang and Liina Pylkkänen. 2015. The interplay of composition and concept specificity in the left anterior temporal lobe: An meg study. *NeuroImage*, 111:228–240.

A Computational Resource

All experiments are performed on platforms with 20 Intel Xeon Gold 6248 CPUs, 236 GB ROM, and 4 Nvidia Tesla v100 32 GB GPUs. Calculating the Computation Scores requires around 1 GPU hour for each model, and each regression requires around 2 hours on the platform for each human subject.

B Results of LLaMA2-Base

Figure 8 in Appendix B displays Pearson's r among the layerwise Composition Score from LLaMA2-base. Similar to LLaMA2-chat, the layers form small correlated clusters and do not exhibit high overall correlation. Figure 10 illustrates the regression scores between the layerwise Composition Score from LLaMA2-base and the control variables. The results mirror those of LLaMA2-chat. Figure 9 in Appendix B depicts the significant brain clusters correlated with the layerwise Composition Scores from LLaMA2-base. Similar to LLaMA2-chat, there are two separated layer clusters in the first and second half of the model layers respectively, and the brain clusters closely resemble those of LLaMA2-chat.

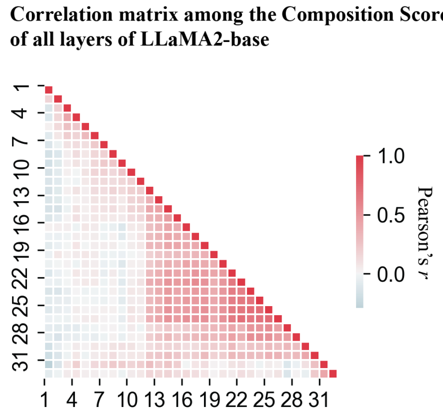

Figure 8: Correlation matrix among the 32 layers of LLaMA2-base.

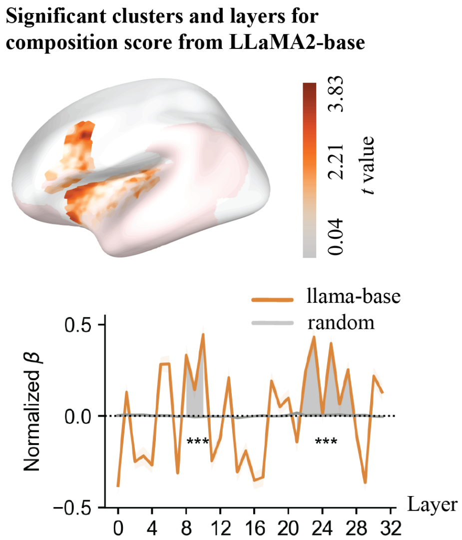

Figure 9: Significant brain clusters for Composition Scores and the significant layers from LLaMA2-base. The orange and red lines depict the normalized β value for each layer of the two models. The grey lines depict the normalized β value for each layer of the random models. The shaded region indicates the significant layers. *** indicates p <0.001.

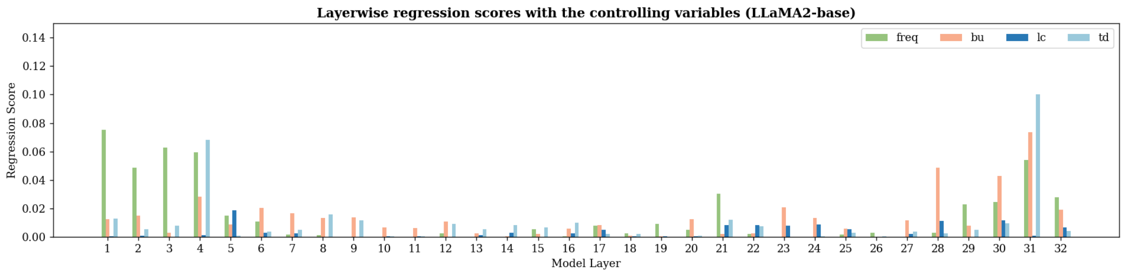

Figure 10: The regression scores R^2 between the Composition Score from LLaMA2-base and the control variables.