Attention is all you need (in the brain): semantic contextualization in human hippocampus

Kalman A. Katlowitz[1], James L. Belanger[1*], Taha Ismail[1*], Ana G. Chavez[1], Assia Chericoni[1], Melissa Franch[1], Elizabeth A. Mickiewicz[1], Raissa K. Mathura[1], Danika L. Paulo[1], Eleonora Bartoli[1], Steven T. Piantadosi[2], Nicole R. Provenza[1], Andrew J. Watrous[1], Sameer A. Sheth[1**], and Benjamin Y. Hayden[1,2**]

1. Department of Neurosurgery, Baylor College of Medicine, and Neuroengineering Initiative, Rice University, Houston, Texas, United States of America
2. Department of Psychology, University of California Berkeley

* These authors contributed equally to this project
** These authors contributed equally to this project

Correspondence: Benjamin.Hayden@bcm.edu

Funding statement

This research was supported by the McNair Foundation and by NIH R01 MH129439, U01 NS121472, NINDS Research Education Grant Programs for Residents and Fellows in Neurology, Neurosurgery, Neuropathology, and Neuroradiology (UE5), the SNS Allan Friedman RUNN Research Grant

Competing interests

The authors have no competing interests to declare.

Acknowledgements

We thank Joshua Adkinson, Hanjie Chen, Justin Fine, Victoria Gates, Suzanne Kemmer, and George Kokalas for invaluable assistance.

ABSTRACT

In natural language, word meanings are contextualized, that is, modified by meanings of nearby words. Inspired by self-attention mechanisms in transformer-based large language models (LLMs), we hypothesized that contextualization in the brain results from a weighted summation of canonical neural population responses to words with those of the words that contextualize them. We examined single unit responses in the human hippocampus while participants listened to podcasts. We first find that neurons encode the position of words within a clause, that they do so at multiple scales, and that they make use of both ordinal and frequency-domain positional encoding (which are used in some transformer models). Critically, neural responses to specific words correspond to a weighted sum of that word's non-contextual embedding and the embedding of the words that contextualize it. Moreover, the relative weighting of the contextualizing words is correlated with the magnitude of the LLM-derived estimates of self-attention weighting. Finally, we show that contextualization is aligned with next-word prediction, which includes prediction of multiple possible words simultaneously. Together these results support the idea that the principles of self-attention used in LLMs overlap with the mechanisms of language processing within the human hippocampus, possibly due to similar prediction-oriented computational goals.

INTRODUCTION

If words are the building blocks of language, contextualization is its architecture (Clark 1996; Sperber and Wilson 1986; Q Liu, Kusner, and Blunsom 2020). That is, word meanings are strongly influenced by context; to quote J.R. Firth, "you shall know a word by the company it keeps" (Firth 1968). Contextualization includes, for example, adverbial modification ("***very*** *tall*"), pronoun reference ("the ***dog*** chased *its* tail"), adjectival composition ("***red*** *balloon*"), and prepositional marking ("***up*** the *mountain*"). Impairments in the ability to understand contextualization are diagnostic of schizophrenia, autism, and other psychiatric diseases (Happé 1997; Kuperberg 2010), and contextualization is an important element of not just speech, but many features of higher-level human cognition (Fedorenko, Ivanova, and Regev 2024; Kaiser, Inciuraite, and Cichy 2020). Classic models of the neural basis of contextualization highlight its functional neuroanatomy (Pylkkänen 2019; Hickok and Poeppel 2007) but surprisingly little is known about the neurocomputational basis of contextualization.

Large Language Models (LLMs) leverage large training sets and machine learning principles to replicate key features of human language use (Radford et al. 2019; Kumar et al. 2024; Cong et al. 2024). Modern LLMs implement contextualization through *self-attention*, by which they learn associations between specific word pairs, and perform a weighted summation of semantic embedding vectors (including both linear and non-linear processes, see **Discussion,** Vaswani et al. 2023; Raiaan et al. 2024). Notably, the magnitude of these weightings vary on a word-by-word basis, so the contextualization is graded, rather than all-or-none. To implement contextual weightings, LLMs depend on various forms of positional encoding that represent the locations of words within the larger clause (Zhao et al. 2024). Ultimately, these LLMs' training functions are optimized for next-word prediction, which is sufficient to drive the learning of contextual relationships and result in internal representations of language. Self-attention, therefore, instantiates a specific theory of contextualization, one with testable hypotheses for how the human brain may implement it.

Recent years have seen great advances in our understanding of the neural basis of contextualization (e.g., (Desbordes et al. 2023; Desbordes, King, and Dehaene 2024; Li and Pylkkänen 2021; Willems and Peelen 2021; Toneva, Mitchell, and Wehbe 2022; Q Liu, Kusner, and Blunsom 2020). Semantic contextualization has consistent effects on decodability that depend on specific roles (Anderson et al. 2021; Zuanazzi et al. 2024; Honari-Jahromi et al. 2021). Indeed, there is evidence for reactivation of specifical adjective semantic responses during listening to subsequent nouns they modify (Fyshe et al. 2019; Dijksterhuis et al. 2024; Zhao, Gao, and Brennan 2025). Inspired by transformer architecture models in LLMs, we hypothesized that contextualization makes use of weighted summation of semantic embedding vectors, which correspond to semantic encoding in the human brain (Turney and Pantel 2010; Mitchell and Lapata 2008; Piantadosi et al. 2024; Jamali et al. 2024; Franch et al. 2025).

We are especially interested in the hippocampus, an area that is responsible for encoding general semantic information (Duff et al. 2020; Manns, Hopkins, and Squire 2003), and whose single units show vectorial encoding of semantics as well (Franch et al., 2025; Katlowitz et al., 2025). Additionally, the hippocampus shows sensitivity to higher order features of speech, such

as grammatical categories (Dijksterhuis et al., 2024; Dragoi, 2024; Opitz and Friederici, 2003). Despite its exclusion from the canonical language network (Fedorenko, Ivanova, and Regev 2024), the hippocampus plays a critical role in contextualization across multiple functional domains (Smith and Bulkin 2014; Frankland and Greene, 2020), including language (Duff and Brown-Schmidt 2012). Indeed, patients with hippocampal damage often exhibit deficits in distinguishing similar experiences that occur in different contexts (Kirwan et al. 2012; Leutgeb et al. 2007; Yassa and Stark 2011; Marr, D 1971; Gold and Kesner 2005) including linguistic ambiguities (MacKay, Stewart, and Burke 1998).

To test these hypotheses, we examined hippocampal single unit responses during passive listening to podcasts. We found broad support for the idea that four key elements of self-attention transformer processes have parallels in the human hippocampus. ***First***, neurons in the human hippocampus show clear positional encoding by combining a multi-dimensional ordinal index scheme with a sinusoidal (Fourier) scheme. ***Second***, contextual responses to words reflect a vector sum of the features defining the word, i.e. neural responses to words are influenced by weighted combinations of prior words. ***Third***, the weighting of these contextual words reflects the weightings associated with those words used in three different LLMs. ***Finally***, we find evidence linking contextualization to next-word prediction, including finding that neurons encode meanings of multiple possible future words simultaneously. Together these results support the idea that the brain and LLMs share a common set of computational principles in the implementation of semantic contextualization.

RESULTS

We recorded responses of 356 isolated units in the hippocampus of ten native English-speaking patients who were undergoing neural monitoring for epilepsy. Each patient listened to the same 47 minutes of English speech from a popular podcast (six monologues taken from the Moth Radio Hour), comprising a total of 7346 words (**Figure 1A**; **Methods**). An analysis of the dataset used here was previously reported in a study of semantic encoding (Franch et al. 2025). Here, we examined how neuronal responses to words change as a function of context (**Figure 1B**).

Positional encoding by hippocampal neurons

Ramp Coding: We found a significant correlation between a neuron's response to each word and that word's position within a sentence in 30.6% of units (n=109/356, $p<0.05$, Kendall's rank correlation). The majority of significant correlations were positive (29.5% of all units, 96.4% of the 109 significant units) and correlations were low overall, average r=0.03 +/- 0.01. **Figure 1C**). This overall bias towards an increase in firing rate for words later in a sentence is reminiscent of previous studies showing an increase in spectral power across word index, which was attributed to working memory load (Howard et al. 2003; Nelson et al. 2017). Although we cannot fully disambiguate the effects of position and memory load, we reran the analysis, controlling for number of opening nodes or closing nodes (correlates of phrasal and thus language-specific working memory complexity (Gwilliams et al. 2024; Weissbart et al. 2024), **Figure 1D, bottom**) with a linear regression model incorporating all three parameters; we found a strong residual correlation between index and the firing rate in 18.0% of units (n=64/356), versus number of closing nodes (6.5%, 23/356) or opening nodes (12.9%, 46/356), consistent with a main effect of position.

Sentences are characterized by structure at multiple levels and a word may have different positions at different levels (e.g., the first word in a dependent clause can be in the middle of the sentence). We examined positional encoding separately at the first eight increasing syntactic depths (depths calculated using CoreNLP, see **Methods**; **Figure 1D**, (Manning et al. 2014). We found neurons with preferential encoding of position at each depth (**Figure 1E**). Firing was more likely to be modulated by the indices of deeper syntactic levels ($\chi 2$ test, $p<0.0001$, correlation between depth and number of significant units, r =0.74, $p<0.05$). To account for correlations in word index across syntactic depths, we ran a multivariate regression to predict firing rate as a function of indices at all depths. Consistent with the preference towards deeper syntactic levels found in the univariate analysis, the median absolute value of the coefficients increased as a function of syntactic depth (Pearson's correlation: r=0.89, p=0.003, **Figure 1F**).

Position Coding: A different (but not mutually exclusive) encoding scheme is selectivity for a specific point in time. For example, in the premotor nucleus of the zebra finch, a population of narrowly tuned projection neurons are each selective for a specific time point and together represent relative position within a song sequence (Kozhevnikov and Fee 2007, see also MacDonald et al. 2011 and Eichenbaum, 2014 for other animal models). We found index selectivity for a modestly sized population of units across multiple syntactic depths (13.2 +/-

4.5% of units, Kruskal-Wallis test on indices 1-10, $\alpha<0.05$). However, unlike in the zebra finch HVC, these neurons were typically selective for more than one word index (e.g. sentence level **Figure 1G)**; they had multiple positions significant on the post-hoc rank sum test for each index versus all others (1.8 +/- 0.2 significant positions per cell, proportion with one peak: 37.7 +/- 9.5%, chance at 38.7%). The existence of multiple peaks raises the possibility of phase coding.

Phase coding: A third (also non-exclusive) positional encoding scheme is used by some LLMs, including in Rotary Position Embedding (RoPe) (Vaswani et al. 2023; Su et al. 2023), and shares features with oscillatory coding mechanisms in the brain (Lisman and Jensen 2013; Moser, Kropff, and Moser 2008). Relative to the other encoding schemes, sinusoidal positional encoding has the advantage of scalability to long clauses. Sinusoidal positional encoding would show up as a consistently modified response to words occurring at specific periods, such as an increase in firing rate for every other word or every fourth word. To test for this code, we computed the power spectrum of the firing rate response to each word across the recording and detected peaks above the aperiodic component (Donoghue et al., 2020) for each neuron (**Figure 1H**). We found many instances of oscillatory tuning, with 50.8% of units (n=181/356) showing a preferred word rate ($\alpha=0.05$, peak greater than shuffled index power spectrum peak). Across the population, peak frequencies approximately tiled the frequency space, as would be predicted by an efficient code (**Figure 1I**).

This apparent sinusoidal encoding could be an artifact of neurons having an intrinsic (non-stimulus driven) firing rate oscillation at a fixed temporal frequency. To rule out this alternative explanation, we tested for any intrinsic periodicity for the neurons during the rest periods before and after the task (62.0 +/- 10.1 total minutes per patient) and compared the intrinsic periodicity to preferred word periods (range of 0.45 to 2.25 Hz, corresponding to a median word rate of 4.5 words per second). Of the units with a preferred word period, 34% (61/181) had no significant undriven oscillatory power (16.7% of total population). However, even among the inherently oscillatory neurons, we observed no systematic relationship between their preferred undriven period and the preferred word period ($r=0.04$, $p=0.7$), indicating that lexical sinusoidal positional encoding is not simply a by-product of intrinsic periodicity.

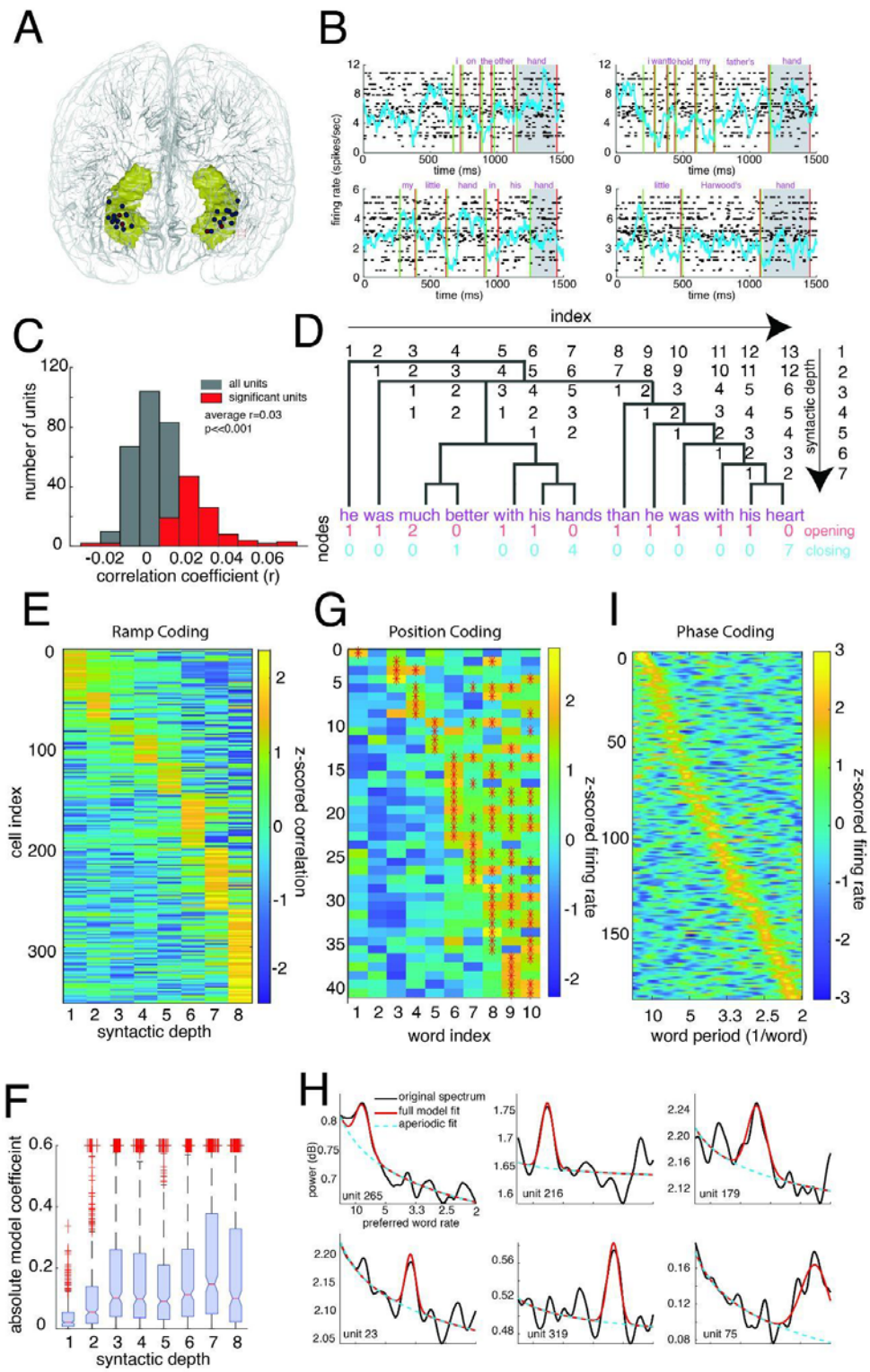


Figure 1. Human hippocampal neurons encode relative word positions during a passive listening task

A. Electrode recording location of hippocampal neurons from ten patients, red sites refer to the data obtained in B. **B.** Population response of 54 units in a single patient to four different phrases, each of which had a different meaning of the word “hand.” Green

lines are onsets and red lines are offsets of each word. **C.** Distribution of correlations between firing rate and word index in a sentence for all (blue) and significant (red) units **D.** Diagram of hierarchical indexing as a function of syntactic depth. The number of nodes opening and closing are listed below in red and blue, respectively. **E.** Z scored r values for each unit sorted by the index of the highest z score. **F.** Box plot of weights for model coefficients predicting neural response from index. **G**. Z scored mean firing rate of units significant for any index as a function of index within a sentence for the first ten positions. Red asterisks indicate significance on rank sum test. **H.** Example power spectra (black) of the word level firing rates of significant units, with the dominant peak (red) highlighted above the aperiodic component (dotted blue line). **I**. Each row is a Z scored fitted PSDs with the aperiodic component subtracted for all significant units (e.g. panel **H**), sorted by the center frequency.

Hippocampal neurons encode attentionally weighted sum of semantic vectors

We next asked how neuronal responses are modified by the semantics of adjacent words. Inspired by self-attention models in LLMs, we hypothesized that neuronal responses might represent a weighted sum of semantic embedding vectors (Viswani et al., 2017; **Figure 2A**). Consider for example, a sentence from one of our podcasts: "*I should probably tell you that the people he's talking about are the football team.*" In this sentence, the neural response (a 356 dimensional vector corresponding to 356 neurons across 10 patients) evoked by the word "*team*" is shifted from its standard response (estimated from the average response evoked by the other four times it was presented). The contextualized response to "*team*" is directed towards the word "*football*", meaning that it can be represented as a weighted sum of the non contextual neural responses for "*team*" and "*football*" (**Figure 2B**). By this measure, the second most strongly contextualizing word was "*people*"; meanwhile, the word "*he's*" has a negative cosine similarity, indicating that the representation of "*team*" is shifted away from "*he's*". This may be because "he's" has a different, contrasting, referent. Averaged across all words in our stimulus set, we find that words are contextualized as a function of relative distance, i.e. they are more sensitive to the words that immediately precede them (**Figure 2C**).

To quantify these effects across our dataset, we investigated several well-known grammatical pairings (Manning et al. 2014). We found that responses to nouns are modified by antecedent adjectives (e.g. "**younger** *brother*", n=137 adjective-noun pairs, z=0.65+/- 1.5, p=0.004, right-tailed paired signed rank test versus nearby word pairs with a similar distribution of pairwise distances, **Figure 2D**). Note that in this analysis, the antecedent word does not necessarily immediately precede the word it modifies (e.g. "**coolest** and meanest *kids*"). Our task did not include a sufficient number of samples to test post-nominal adjectives (n=4). Verbs contextualized adverbs that followed them (e.g. "**know** very *well*" n=86 verb-adverb pairs, z=0.52+/-1.4, p=0.006 **Figure 2E**), just as adverbs contextualized the verbs that followed them (e.g. "**completely** *soaked*" n=158 adverb-verb pairs, z=0.24 +/- 1.1, p=0.02, **Figure 2F**). We also found that verbs contextualize the responses of nouns that follow them (typically, direct objects

of the verb, e.g., "**call** my two *brothers*", n=289 verb-noun pairs, z=0.56+/- 1.5, p=0.001, **Figure 2G**), but nouns (typically subjects) do not contextualize the responses of verbs that follow them (e.g., "my **father** *dies*", n=747 noun-verb pairs, z=0.02 +/- 0.76, p>0.5, **Figure 2H**). Note that this pattern is consistent with the idea that heads contextualize tails more strongly than the reverse (Honari-Jahromi et al. 2021). Indeed, not all pairing types demonstrated a significant contextualization, perhaps due to a weaker semantic relationship or an abstract relationship. In prepositional phrases ("**over** the *world*"), the prepositions did not contextualize following nouns, despite a relatively large sample size (n=346, z=-0.03 +/- 0.6, p>0.5, **Figure 2I**). Determiners also did not generally modify their respective nouns ("**that** *flight*", n=324, z=-0.13 +/- 0.51, p>0.5, **Figure 2J**); nor did possessive adjectives ("**his** *cup*") contextualize their nouns (n=206, z=-0.13 +/- 0.61, p>0.5, **Figure 2K**).

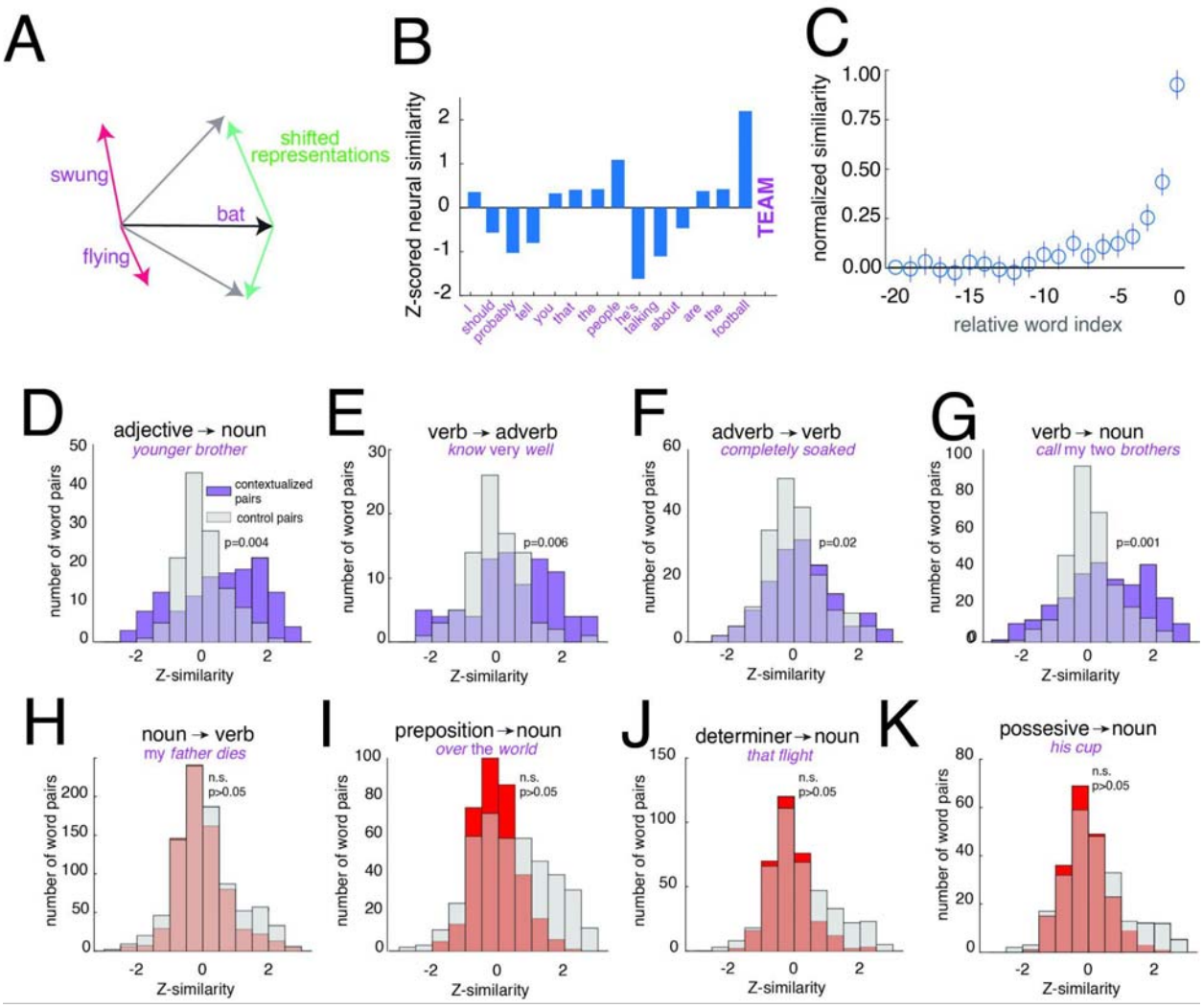


Figure 2. Semantic relationships are represented through contextually weighted vector summations

A. Vectorial diagram of how adjacent words can modify the neural vectorial representations of words (e.g."swung" appearing in the context of "bat" shifts the firing rate vector for "bat" to have a more similar cosine distance to the vector for "swung") **B.** Z-scored cosine similarity values between the shifted neural representation of the final word "team" and all preceding words in the sentence. **C.** Normalized (0 to 1) average z-scored cosine similarity values as a function of relative index. Error bars represent +/- 2*SEM, **D-K.** Distribution of the z-scored contextual pairing values for the word pairs designated above (purple/red) and the values for control word pairs (gray)

Correspondence between LLM contextual embedding and neural responses

The specific strength of mappings between the past and current words in transformer-based LLMs is given by the ***attention pattern grid*** (**APG**, Vig 2019) and is used by the LLM to define relationships. A high weight for a pair of words causes the embedding of the prior word to influence the representation of the current word. We reasoned that the projection of neural responses onto adjacent words (**Figure 2A,B**) allows us to infer what is in essence a **neural APG**. We next compared neural APGs to corresponding LLM-derived APGs. To generate the LLM-based APG, we treated GPT-2 dimension-wise responses as if they were unit responses and replicated the same process we used above to create an empirically derived estimator of the single APG for the LLM. (While GPT-2 has multiple attention heads, this empirically derived APG would necessarily be a single composite across all heads). For example, in our example sentence above, we find a close quantitative match between the neural data and the inferred APG from Layer 36 of GPT-2 ($r=0.78$ $p<0.001$, **Figure 3A**).

We next created 37 latent LLM APGs, one for the initial embedding and then each of the 36 layers of GPT-2, and compared each one to our neural APG. To restrict our analysis to words likely to interact with each other (per **Figure 2C**) we tested the correlations across a range of local contextual windows (2-20) for each layer. We find that the mean correlation between neural and LLM APGs across words is positive for all layers and contextual windows but can vary greatly as a function of both (**Figure 3B**). The correlation is maximized at Layer 20 and with a span of the five previous words ($r=0.23$ +/- 0.56 $p<0.001$, **Figure 3C**). This result is not unique to GPT-2, which is decoder-based; we obtain similar results using the encoder-based LLM DeBERTa V3 (maximized at layer 8 out of 12, $r= 0.16$ +/- 0.54 $p<0.001$, He et al., 2023) and a different state-of-the-art decoder based LLM Llama-3 (maximized at layer 18 out of 33, $r= 0.24$ +/-.07, Grattafiori et al. 2024).

A close quantitative match between neural and LLM-based APGs

We next examined the word-by-word temporal structures of the APGs (focusing on the five most recent words, see above). We identified ten distinct attention patterns using the Leiden algorithm (n=6579 words that all patients heard at least twice; **Figure 3D, left)**. While the immediately adjacent word was the most common strongest contextualizing influence, the LLM-based APGs demonstrated a striking variety of attention patterns. Remarkably, we find the same pattern of activity in the neural data when following the same sorting, even though the LLM and

neural data are completely independent of each other (**Figure 3D, right**). For example, a word with a high similarity index three words prior to it in LLM space tends to also have a higher similarity index three words prior to it in neural space. This effect is even clearer when only including words with a significant correlation between the neural and LLM attention vectors (n=1042 words, 16% of all words α=0.05, **Figure 3E**). A few examples demonstrating this phenomenon appear in **Figure 3F**.

What features predict stronger matches? First, there was a strong relationship between the standard deviation of each word's LLM attention vector and the subsequent match to the neural attention vector (r=0.21, p<0.0001, **Figure 3G**), suggesting that that if a word's meaning is not modified by prior words then the match will be poor simply because there are no features for our metric to capture (Sandoval-Segura et al. 2025). Interestingly, despite finding a decay in the match with window sizes larger than the 5 prior words (**Figure 3B**, right), this effect was not found for word pairs that are still within a clause (**Figure 3H**), demonstrating that the hippocampus can dynamically adjust its "pairing memory" as a function of the dynamic grouping of words. Of note, the mean clause size was 4.7 +/- 4.9 words, explaining our maximum at 5.

Next, we examined the semantic features of the words that showed contextualized effects. For each word, we defined a polysemy index as the variance in its contextual embeddings across occurrences (Franch et al. 2025). We found a positive correlation between polysemy index and neural contextualization effect (r=0.05, p<0.0001, Spearman's correlation), demonstrating that the more variable a word's embedding, the better the neural definitions of relationships correspond to the LLM defined relationships. This effect size was stronger when averaging across all representations of each unique word (r=0.12, p<0.005, **Figure 3I**). This difference could also be seen within different grammatical subcategories: direct objects (that are defined by the verbs describing them), had significantly better matches between the LLM and neural data than the less commonly linguistically modified subjects (p<0.001, rank sum test, **Figure 3J**).

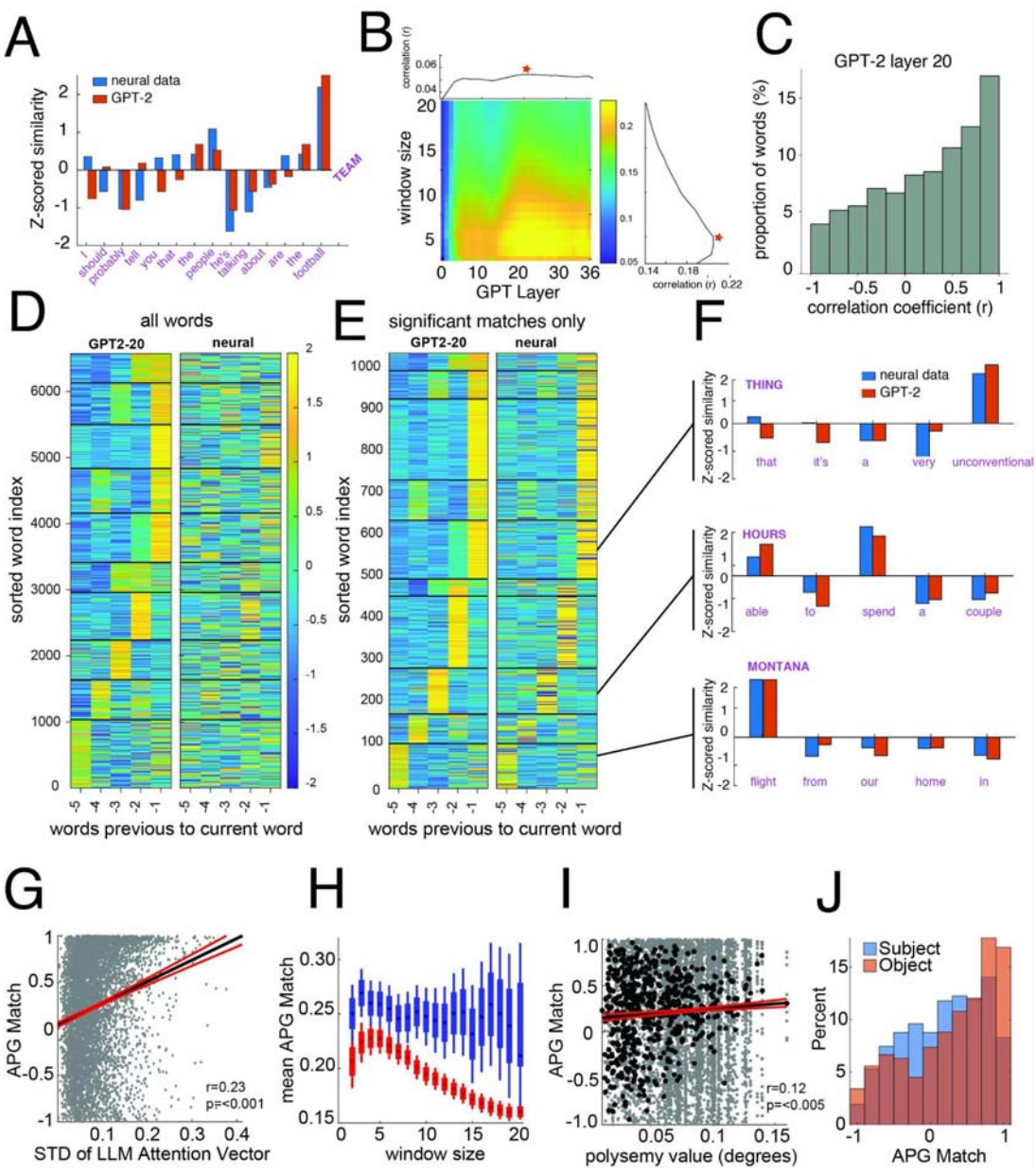


Figure 3. Neural attention patterns closely match LLM attention patterns.
A. Similar to Figure 2B but with the APG vector for GPT-2 Layer 20 overlaid. **B.** Average match across all words between neural APG and GPT-2 APG for all layers (x axis) and window lengths from 2 to 20 (y axis). Average values projected above and to the right. Red star reflects the maximum. **C.** Distribution of r-values for correlations between the neural APG and the LLM embedding (GPT-2 Layer 20) APG for the previous 5 words for all words. **D.** The z-scored APG vectors for all words from GPT-2 layer 20 sorted by cluster index and peak (left) with the neural data sorted on the same indices (right). **E.** same as **C** but only words that had significant matches. **F.** Example phrases from the clusters. **G.** Correlations between neural and LLM attention vectors as a function of index within a syntactic depth of 8. **H.** Mean and SEM (box) and 2*SEM (whisker) of r values between the APG vectors of the LLM and the neural data for each

word for all words (red) and only words whose antecedent window is still within the same clause (blue). **I.** correlation between the match of the neural and LLM vectors of each word and the polysemy score of that word for all words (grey) and all unique words (black, r values are average across all presentations). **J.** Box and whisker plot representing the distribution of correlations between LLM and neural APG vectors for words that are subjects, objects, and prepositional objects.

Contextual models capture neural variance better than non-contextual models

The fact that word-evoked neural responses depend on preceding words implies that the neural data should be better fit by contextual semantic models than by non-contextual ones (Jain and Huth 2018). To test this idea, we next compared model fits for contextual and non-contextual embedders, using principal component analysis (PCA) on model embeddings to ensure the same number (100) of predictive dimensions. For each source of semantic embeddings we created an explanatory Poisson model using ridge regression and compared the log likelihood of the model to that of a model that had its embedding vectors shuffled (ΔLLH). As predicted, the contextual models we tested (GPT-2, Llama-3 and DeBERTa V3, all at layers that maximized the fit to the neural APG) have better fits than non-contextual models (Glove and Word2Vec, signrank test, $p<0.001$ for all pairwise comparisons, **Figure 4A**).

Within a given LLM model, however, contextualization depends on the layer within the model. Typically, higher layers have greater contextualization, although the highest layers may be more oriented towards prediction of potential future words, making them less well correlated with the current semantically turned neural activity (Schrimpf et al. 2021; Caucheteux, Gramfort, and King 2021; Toneva and Wehbe 2019; Hong et al. 2024). Consistent with these past studies, we found the strongest model fits at middle layers - in our case, at Layer 25 out of the 37 total layers (**Figure 4B**), which is close to the Layer 20 that optimized the match between neural and LLM APG (**Figure 3C**).

These analyses, however, depend on interpreting the intricacies of LLMs, which include far more than local context. To isolate the effects of context, we next examined the explanatory power of recent words within the current neural representations using a non-contextual embedder (word2vec). Here, rather than quantifying the impact of representations of prior words on those of current words (**Figure 2,3**), we directly examined the relationship between firing patterns and the semantic information of prior words. We first found a significant relationship in 63.2% of units, though at a lower strength than current words (median ratio of ΔLLH = 40.8 +/- 21.6%). Notably, the same units that responded to the current word also responded to prior words (correlation of the ΔLLH per unit $r=0.86$, $p<0.0001$, Spearman's correlation, **Figure 4C**), demonstrating that the firing patterns of semantically responsive neurons are modulated by both the current and previous words. Further, if we use the GPT-2's empiric APG to select the most contextually relevant prior word for each word (as per **Figure 3C**) we find that the model using these contextual words to predict current firings rates outperforms a model that picks random words with a similar distribution of word distances (formed by randomizing the distances, $p<0.0001$, sign rank test on ΔLLH).

Yet perhaps these relationships are due to semantic coherence, in which adjacent words contain similar semantic features. To quantify the added information we calculated the Akaike information criterion (AIC) of the model using both the current and most contextualizing previous word and compared it to the AIC of the model using just the current word for each neuron. If the two words provide the same information then this should not improve the model fit. In contrast, we found that adding the immediately prior word improves the fit in 22.1% of the 272 significant units using just the current word (i.e. there is a decrease in the AIC). In stark contrast, adding the LLM APG defined contextual word improved the fit in nearly all, at 94.9% of units (**Figure 4D**) despite being farther away (an average of 2.14 +/- 1.3 words prior relative to 1 word prior). Additionally, the representations of the contextual words had a more significant role in the joint models: the norm of the coefficients for the contextualizing words was 70.1%+/- 31.5 of the norm for the current words' coefficients, significantly greater than the 58.8% +/-27.7 for the immediately preceding words (paired t-test, $p<0.001$, **Figure 4E**). Together, these two results demonstrate that the words used by an LLM to provide context are the best source of information to improve a model fitting ongoing words to neural data.

Role of context in ongoing next-word prediction

LLMs use contextualization solely to improve next-word prediction, which is their training goal. We therefore examined the relationship between neural contextualization and next-word prediction in our data. We first found that word-evoked neural responses are related to the semantics of upcoming words, just as they are by past words (**Figure 4C,F**). The degree of contextualization appears to be an inherent feature of units: units with a stronger relationship to past words (defined by the ΔLLH) had a stronger relation to future words as well ($r=0.78$; $p<0.001$, Spearman's correlation). LLMs however, do not just predict a single word, but rather a distribution of potential words (Vaswani et al., 2017). Similarly, we found that hippocampal activity is simultaneously associated with each of the top three most likely candidates for the next word given all the previous words (**Figure 4G**). Indeed, the strength of encoding (as inferred from LLH difference) is proportional to their relative probability of occurrence. We found a stronger match to the most likely next word than the actual next word spoken ($p<0.05$, rank sum test). These results suggest that the hippocampus is actively predicting the distribution of possible futures.

The model of next-word prediction has another testable implication: prediction errors should reduce the faithfulness of neural semantic encoding as they become decoupled from the predictive framework that normally supports neural representation (Goldstein et al. 2022). To test this hypothesis, we defined *prediction errors* as word surprisal, the negative logarithm of the probability of hearing a word given the previous context, and the *representational error* as the deviance of the models' prediction for that word. We find an average correlation of 0.07 +/- 0.09 between the prediction error and the representational error across units (**Figure 4H**). Overall, 52.0% of units (185/356) show a significant positive correlation versus just 8.2% that show a significant negative correlation (29/356) ($\alpha=0.05$). In other words, surprising words are represented less faithfully than unsurprising words.

Artifactual explanations for this result could include that surprising words are less common, can occur in moments of uncertainty (high entropy), or even just elicit a higher firing rate (any of which could decrease the accuracy of representation). We can control for these possibilities by creating a linear model that predicts the deviance of each word using the surprisal, frequency, entropy, spike count of the word, and all of their interactions. Doing so, we nonetheless still found a significant positive effect of surprisal in 32.9% of units (versus a negative effect in 2.5% of units). Interestingly, despite the significant relationship between the accuracy of the model in representing the current word and the impact of surprisal on its deviance (r=0.24, p<0.001 Pearson's correlation), it was not always the units with the highest semantic accuracy that had the strongest relationship between model error and surprisal (**Figure 4I**). This weak correlation at the level of units implies there is an unavoidable tradeoff between accurate representation of the current word and susceptibility to modulation by ongoing predictions.

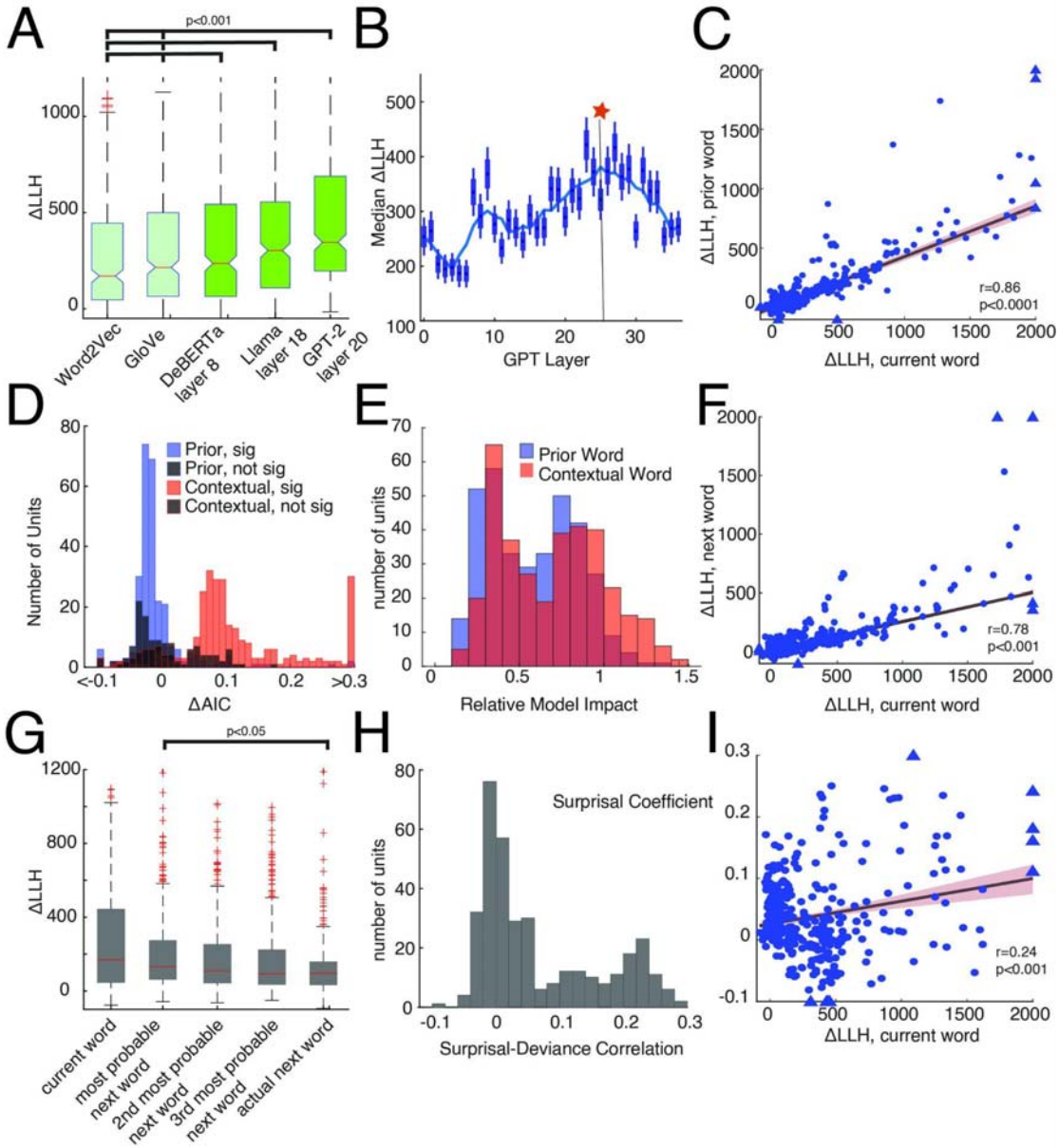


Figure 4. Neural activity reflects combinations of the semantic representations of the current, prior, and potential future words

A. A comparison of the LLH for each neuronal model with different semantic embeddings for non-contextual (Word2Vec and GloVe) versus contextual (DeBerta V3, Llama-3 and GPT-2). **B.** Box and whisker (median and SEM) plots of the distribution of LLH differences across all neurons using the hidden states of each layer of GPT-2 as the semantic predictors. Blue line represents the median smoothed with a gaussian of width 5 and star indicates maximum of fitted line. **C.** Plot of ΔLLH for current word versus next word for each neuron. Triangles are datapoints that are outside the limits of the plot. Black and red lines represent the best fit +\- 95% confidence interval. **D.** Histogram of the difference in the AIC per word between the basic current word model and the current + prior word model (blue) or the current + contextual word model (red). Positive values indicate that the unit's firing patterns are better explained with both words, even when penalizing for the increased number of parameters. **E.** Distribution of the norm of the coefficients for the prior word (blue) and the contextual word (red) normalized by the norm of the coefficients for the current word from the joint models, for each unit. **F.** Plot of ΔLLH for current word versus next word for each neuron **G.** Box and whisker plot for the dLLH distributions for current versus the top three candidates for next word ordered by their probabilities and was as actual next word. **H.** Distribution of correlations between the deviance of the model for each word and the relative surprise of that word for all neurons for all unis. **I.** The ΔLLH of the non-contextual model as a function of the coefficients of surprisal in the linear model that predicts deviance

DISCUSSION

Here we report that neural representations of word meaning within the hippocampus incorporate the activation patterns of the words that contextualize them and the strength of this incorporation is well approximated by the strength of contextualization used by transformer-based LLMs. These results therefore suggest a specific mechanism for semantic contextualization in the human brain, one that relies on the powerful and convenient properties of vectorial semantic representations, and that depends on positional encoding, and that is associated with next-word prediction. In other words, our results support the hypothesis that contextualization in the human brain works via a self-attention-like process similar to LLMs. Ultimately, our findings are grounded in a vectorial representational scheme for semantics that has been proposed in models of meaning in the neuronal responses (Franch et al. 2025; Jamali et al. 2024) and in cognitive science more broadly (Shepard 1980; McClelland and Rogers 2003; Minda and Smith 2001). Our results therefore extend earlier work supporting vectorial representation principles for meaning. Indeed, one of the major virtues of vectorial codes is that they allow for straightforward contextualization through vectorial operations (Piantadosi et al. 2024).

Neural codes for positions of words are essential for contextualization, as they represent a mapping principle (Ding 2025; Carpenter et al. 2018; Fló 2021). The finding of a sinusoidal positional encoding is particularly interesting. Sinusoidal codes have several benefits, including representation of relative distances (which enhances flexibility), smoothness, generalization to arbitrarily long sequences, and ease of learning (relative to learned positional representations) (Zheng, Ramasinghe, and Lucey 2021). These factors have also been proposed for why the entorhinal cortex has sinusoidal codes for navigational space, in the form of grid-cells (Monteiro, Pedro, and Silva 2022). Recent years have witnessed the extension of grid cell principles to non-navigational domains (Constantinescu, O'Reilly, and Behrens 2016; Park, Miller, and Boorman 2021). Our results, therefore, while not indicative of a grid-like pattern *per se*, suggest the extension of the principle of periodic codes to the domain of lexical position mapping.

Contextualization in the brain does not happen arbitrarily; instead, it is a learned process that attempts to achieve a specific goal. Understanding this "objective function" is a key question in linguistics. Computer scientists have independently found that next-word-prediction (or, more generally, next-token prediction, see below) is a tool powerful enough to introduce features not only remarkably similar to contextualization, but also other cognitive processes such as syntactic parsing (Oota, Gupta, and Toneva 2023), reasoning, and generalization (Goldstein et al. 2021; Liu et al. 2025; Cong et al. 2024), though see (Olatunji and Sheppard 2025). In other words, many of the core features of language are not built into LLMs; instead, they are an emergent feature of architecture and training principles. Our results here highlight the links between contextualized responses and next-word prediction in the human brain. The relationship between neural activity and upcoming words has been interpreted as preparatory activity (Goldstein et al. 2021), though others have raised concerns that this analysis is limited by inherent correlations

within natural language (Schönmann et al. 2025). We therefore also demonstrated that information about upcoming words must be present in some form, as the evoked responses were modulated by these predictions: expected words were better represented than words that were surprising (even when accounting for lexical frequency), indicating a latent encoding of next-word likelihood (Heilbron et al. 2022). Second, we show that these representations include not only the most-likely next word but also the second and third most-likely next word. Said differently, the brain seems to actively guess multiple possible upcoming words, presumably so that it can more efficiently represent upcoming sensory information (Bubic, Von Cramon, and Schubotz 2010; Comrie et al. 2024). Finally, we show that the units that are modulated by context are the same units that are involved in ongoing prediction, thereby directly connecting two separate features of speech processing: representations of the past and representations of the future (Merlin and Toneva 2024). Future research could investigate how prediction errors modulate learning signals in the hippocampus, and how these signals could contribute to the refinement of semantic representations.

The hippocampus is well-situated to perform the core elements of semantic contextualization: (1) it is closely associated with mapping functions that are related to temporal position encoding (Naya and Suzuki 2011; Whittington et al. 2020), (2) it integrates information across multiple modalities and uses that information to contextualize representations in a variety of domains (Maren, Phan, and Liberzon 2013), (3) and it is closely associated with the processes of prediction that drive contextualization (Aitken and Kok 2022; Mizumori 2013). Finally, (4) the hippocampus has a long-established role in memory, one that has many links with semantic knowledge and with semantic representations in particular (Duff et al. 2020; Binder and Desai 2011). Indeed, the role of the hippocampus in language is well established (Duff and Brown-Schmidt 2012; Piai et al. 2016; Binder et al. 2009), as is its role in semantic representations specifically (Franch et al. 2025). Recent studies using intracranial unit recordings have begun to show clear evidence for hippocampal role in language (Dijksterhuis et al. 2024; Franch et al. 2025; Katlowitz et al. 2025). We suspect some of the reason for the absence of the hippocampus in classical language models is negative evidence - the hippocampus can be difficult to image using fMRI and lesions isolated to the dominant hippocampus are rare. However, another factor is that the hippocampus is clearly not a site uniquely aligned to language coding processes, so it does not fit modern uniqueness-focused definitions of language areas (Fedorenko, Ivanova, and Regev 2024). In any case, we surmise that future work on the functional architecture of language will need to incorporate the hippocampus into its theorizing, specifically for applying rapidly changing contexts.

Our study has several notable limitations. First, we assume a single APG, whereas in LLMs the attention mechanism is multilayered and multiheaded. This simplification likely overlooks a dynamic and distributed attentional mechanism in the brain, wherein different networks may attend to different features (Kumar et al. 2024). Second, we restrict ourselves to the study of the attention mechanism of contextualization, and ignore the fascinating multi-layer perceptron (MLP) element of contextualization (Viswani et al., 2017). This nonlinear step is less well understood but accounts for more model parameters than the attention block and allows for

many higher order features, such as representation that distinguishes learned (rather than inferred) contextualization (Gerber 2025). Another limitation of the present work is that our study focuses only on words, whereas modern LLMs make use of sub-word tokens. There is no doubt that the human brain operates at the token level as well, so future work extending these ideas to tokens is likely to provide better fits to data. Finally, our work assumes a single network and infers the underlying “lower order” representations. In LLMs (and likely in the brain as well), context is incorporated in a staged matter across many layers. Future high density examinations of hippocampal and cortical networks will be better situated to assess the roles of individual networks and the transformations between them.

While we see some overlaps, we do not, ultimately, believe that the brain operates in the same way that LLMs do. There are many obvious differences between language computations in the brain and in LLMs. At a mechanistic level, the brain is made of recurrent networks of spiking neurons versus an LLM which is composed of feedforward matrix operations based in silicon. This dichotomy is responsible for many divergent features, such as learning speed, power efficiency, robustness, and flexibility. More broadly, LLMs were designed for a single task whereas the brain has evolved to solve a wide variety of problems, from motor control to intentionality. It is clear that these embodied processes have a large - perhaps definitional - role in cognition (Cisek and Pastor-Bernier 2014). As such, LLMs are not designed to be models of the brain. However, their remarkable efficacy at language generation means they can serve as a source of hypotheses for neuroscientific investigation.

METHODS

Experimental Paradigm

Data collection procedures have been described in detail previously (Franch et al. 2025). In brief, we recorded single-neuron activity from the hippocampus of ten adult patients with drug resistant epilepsy (six male, four female) undergoing intracranial monitoring at Baylor St. Luke's Hospital for seizure localization. None of the patients had hippocampal seizure foci. Electrode position was confirmed on postoperative CT and coregistered to a canonical brain (Wang et al. 2023; Magnotti, Wang, and Beauchamp 2020). Recordings were performed in the Epilepsy Monitoring Unit from Behnke-Fried stereoelectroencephalography (sEEG) probes with 8 microwires each, sampled at 30 kHz using a Blackrock Neuroport system. Spike sorting was performed with WaveClus and manually curated (REF). Participants listened to six autobiographical stories from *The Moth Radio Hour* (totaling 47 minutes and 7,346 words), presented via hospital TV speakers. The audio was synchronized with neural recordings and then transcribed by Assembly AI. Labels were then manually corrected using PRAAT (Boersma 2011).

Generation of Contextual Embeddings

We reconstructed each sentence by reading the transcript in order and appending tokens to a running string until we encountered sentence-ending punctuation (e.g., ".", "?", "!"). Residual words after the last delimiter were grouped into a final sentence. All punctuation tokens were retained to maintain contextual fidelity but removed from the final analysis.

We used three LLMs in this study: GPT-2 Large (Wolf et al. 2020), Llama-3 (Grattafiori et al. 2024), and DeBERTa V3 (He et al. 2020).

GPT-2 Large: We extracted 37 hidden states consisting of 36 transformer blocks + 1 embedding layer. Each word was a 1280-d vector, and the native context window was 1024 tokens.

Llama-3 7B: We extracted 33 hidden states consisting of 32 transformer blocks + 1 embedding layer. Each word was a 4096-d, vector and the native context window was 1024 tokens.

For both GPT2 and Llama-3, we built an expanding context window. Beginning with an empty string, we appended one new word, re-tokenized the running text with each model's tokenizer, ran the model in inference-only mode, and averaged the sub-word vectors belonging to the just-added word. When the running token sequence exceeded 1024 tokens, we truncated the leftmost words so that all three models would be compared under identical effective context sizes. This "progressive-context" procedure respects the decoder-only training objective of both models

while yielding word-level embeddings from every layer. The result was 37 layers×1,280-d vectors per word for GPT-2 Large and 33 layers×4,096-d vectors per word for Llama-3 7B.

DeBERTa-V3: Base (12 transformer blocks + 1 embedding layer, 768-d vectors, 512-token context) (He et al. 2020) was handled separately due to its bidirectional encoder architecture, which processes the entire input in parallel and has access to future context. To align its outputs more closely with the unidirectional processing of GPT-2 Large and Llama-3, we applied the same expanding-context procedure: beginning with an empty string, we appended one new word at a time, rebuilt the running text by concatenating words (preserving punctuation for contextual fidelity), and re-tokenized the updated segment with DeBERTa's tokenizer. Although the model sees both past and future tokens, we extracted only the hidden states corresponding to the just-added word which limits interpretation to left-contextual information. Sub-word vectors for that word were averaged to yield one word-level vector per layer. Hidden states from all 13 layers were retained, yielding one 768-d vector per layer per word (He et al., 2023).

All extraction scripts were written in Python 3.10 with PyTorch 2.3 (Paszke et al. 2019) and the Transformers library (Wolf et al. 2020, 20). NumPy, Python, and CUDA random seeds were fixed to ensure reproducibility, and the resulting layer-specific word embeddings were saved to Excel files for downstream analyses.

Generation of Static (non-contextual) embeddings

We used Word2Vec and GloVe to derive non-contextual ("static") semantic vectors for our speech stimulus. All words were converted to lowercase and stripped of punctuation before lookup. Surnames and proper nouns that were not present in the embedding vocabulary (e.g., "*Applebee's*") were excluded from downstream analyses.

For Word2Vec we employed the pre-trained fastText skip-gram model released by Facebook AI Research (Joulin et al. 2016), accessed through MATLAB's fastText Word Embeddings interface. The model supplies a 300-dimensional vector for each word, trained on hundreds of millions of tokens so that semantically similar items occupy nearby regions of the embedding space (Mikolov et al. 2013). For GloVe, we used the 840 billion-token, 300-dimensional Common Crawl model (glove.840B.300d.txt) (Pennington, Socher, and Manning 2014). A Python script loaded the plaintext file into memory, mapped each transcript word to its corresponding vector, and wrote the resulting word × 300 matrix to an Excel workbook for downstream analysis.

Although both methods learn from word co-occurrence statistics, their training objectives differ subtly: Word2Vec (and fastText) predict the surrounding context of a target word in a local sliding window, whereas GloVe factorizes a global word-co-occurrence matrix to preserve ratios of co-occurrence probabilities across the entire corpus. In practice, Word2Vec tends to capture fine-grained relational semantics e.g., king – man + woman ≈ queen, while GloVe emphasizes broad topical similarity.

Neural Analyses

Neural responses were defined as the spike count during the word presentation, shifted by an offset of 80 ms (to account for response latency).

Fourier analyses: Power spectra for the neural response to words were calculated through Welch's power spectral density (PSD) estimate. A frequency range of 100 frequencies between 0.05 and 0.5 Hz was analyzed, corresponding to periods of 2 through 20 words. We used windows of 512 words with a sample overlap of 256 words. Rare (30/7356 words in 1/10 patients) missing words from experimental interruptions were linearly interpolated between adjacent values for the purposes of this analysis. The resultant PSDs were then smoothed with a gaussian filter that had a full width half maximum (FWHM) at 0.023 Hz (a window of 15 samples). A peak in the spectrum above the aperiodic component was estimated through the FOOOF toolbox (Donoghue et al. 2020).The single largest peak was selected, as long as it had a minimum width of 0.01 Hz and its amplitude was at least 2.5 standard deviations above aperiodic fit.

We also calculated the intrinsic frequency of the units during the resting period, defined as the time before and after the experiment. Spike count was sampled at 1 kHz. Here we used windows of 4.4 minutes and 2.2 minutes of overlap to obtain similar frequency resolution. The frequency range was 0.045 to 4.5 Hz, corresponding to a median word rate of 4.5 words per second. For example, if a neuron fired once every other word, this would correspond to a preferred frequency of approximately 2.3 Hz.

Attention Pattern Grid

The ***empiric attention pattern grid*** is an estimate of the degree to which all past words influence each current word in terms of the proportion of influence they have on the semantics of that word. It therefore recapitulates the function performed by the APG used in transformer architecture LLMs. We first estimated the neural APG from neural data alone and, for comparison, performed an analogous process on the LLM embeddings. Note that the LLM-derived empiric APGs were not derived directly from the LLM APGs themselves; they were inferred so as to keep the estimation procedure the same for brain and LLM. For reference, a true APG from an LLM is the result of the dot product of the Query and Key matrices, and there are many heads of attention per layer.

Our empiric APGs were calculated in the same manner for all embedding matrices. Embedding matrices were 7,346 rows, corresponding to the number of words in our stimulus set, by M columns, corresponding to the number of embedding dimensions per word. For the neural embedding matrix M was equal to 356, corresponding to the number of neurons collected across patients. For the semantic empiric APGs, M ranged from 300 (word2vec) to 4,096 (Llama-3).

We then created two new embedding vectors for each word: the average vector across all instances of that word $\underline{\square_{\square}}$ and the shifted representation of the word j is $\Delta\square_{\square} = \square_{\square} - \underline{\square_{\square}}$. Then each value of the APG was created by measuring the cosine similarity between the shifted vector and the mean vector of the previous words

$$\square\square\square(\square,\square) = (\underline{\square_{\square}} \cdot \square\square_{\square}) / (||\underline{\square_{\square}}|| * ||\square\square_{\square}||)$$

for j=1:7,346 and i=1:j

Words with only one instance in the entire task were excluded from analysis as it was impossible to calculate a shifted representation. Only causal influences were calculated.

Clustering Attention Vectors

To explore the structure of our empiric APGs, we constructed a k-nearest neighbors (k-NN) graph on truncated APGs going back 5 words. This graph served as the basis for unsupervised clustering. We applied the Leiden algorithm—a community detection method known for identifying well-connected, non-overlapping clusters—to segment the graph into distinct groups of words with similar recent attention dynamics. These clusters were hypothesized to reflect shared structure patterns in word relationships. For the sake of graphical presentation we then sorted these clusters by the index of the highest average value in the cluster.

Poisson Regression

<u>Data Preprocessing and Feature Construction:</u>The spike count of a neuron in response to a word was defined as the number of spikes recording for the duration of that word, with the onset and offsets shifted by 80ms to account for synaptic transmission to the hippocampus. Neural response vectors were then defined as the response for each of *N* neurons over *W* words. These were stored in a matrix **Y** with shape ($W \times N$). Each word was also represented by a feature vector (e.g., GPT-derived semantic embeddings), forming a matrix of shape ($W \times D$). For each feature, PCA was used to generate the 100 orthogonal dimensions that explained the most variance, creating **X_raw** of shape (Wx100). If two features were used (e.g. current and prior word) then they were concatenated into a matrix of shape *W x 200*. A scalar **duration** vector (length *W*) represented the duration of each word which needed to be included as words of different lengths would have different spike counts assuming a fixed firing rate. Rows with missing data (NaNs) in any of **X_raw** (e.g. names not represented in Word2Vec), or **Y** (e.g. not heard by every patient) were removed. To model duration-dependent effects, each feature in **X_raw** was multiplied elementwise by **duration** to form an interaction matrix. The full design matrix X was defined as

$$\square = [\square\square\square\square\square\square\square\square\square \mid \square\_\square\square\square \mid \square\_\square\square\square * \square\square\square\square\square\square\square\square\square]$$

yielding a total of 201 dimensions (401 for two word models). Each column of **X** was standardized to have zero mean and unit variance. Additionally, 100 shuffled versions of **X** were generated by permuting the rows of **X_raw**, recomputing the interaction terms, and standardizing again. These permutations were used to generate null distributions for hypothesis testing. Notably, the durations were not shuffled.

$$\square_{\square\square\square\square\square} = [\square\square\square\square\square\square\square\square\square \mid \square\_\square\square\square_{\square\square\square\square\square} \mid \square\_\square\square\square_{\square\square\square\square\square} * \square\square\square\square\square\square\square\square\square]$$

<u>Poisson Ridge Regression Model</u>

We modeled spike counts with a Poisson generalized linear model (GLM) using an exponential link function. For each word *w*:

$$\hat{\square}_{\square} = \square^{\square_{\square} * \square + \square}, \square_{\square} \sim \square\square\square\square\square\square\square(\square_{\square})$$

where: $\square_{\square}$ is the feature vector for word w, β is the weight vector, *b* is a scalar bias, and $\hat{\square}_{\square}$ is the predicted spike count, and $\square_{\square}$ is the observed spike count.

We used ridge regression to prevent overfitting, yielding the following loss function:

$$L(\beta, \alpha) = \sum_{w} [\, \hat{y}_w - y_w * log(\lambda_w) \,] + \alpha * ||\beta||^2$$

wherein α is the regularization parameter.

The model was implemented in PyTorch on a NVIDIA GeForce RTX 4070 Ti SUPER to decrease training time. Training was performed using the Adam optimizer (Kingma and Ba 2017) with a default learning rate of 0.01. Training proceeded for up to 300 epochs, with early stopping if the decrease in penalized loss fell below 10^{-4} for ten consecutive epochs. Loss was computed at each training epoch. Gradients were back propagated and the weights were updated until convergence or early stopping.

Hyperparameter Selection:

To choose the regularization strength α, we used a two-stage grid search:

1. **Coarse Search** over: $\square \in \{ 10^{-5}, 10^{-4}, \ldots, 10^{3} \}$
2. **Fine Search** around the best coarse value, spanning two orders of magnitude.

For each value of α, we performed 5-fold cross-validation. For each fold, the model was trained on the other four folds and evaluated on the held□out fold by computing the Poisson log□likelihood:

$$LL = \sum_{w} [y_w * log(\hat{y}_w) - \hat{y}_w - log(y_w!) \,]$$

These fold□wise log□likelihoods were averaged to yield an overall score for each α. The α yielding the highest average log-likelihood was selected.

Final Model Fitting and Evaluation

Once the best α was determined using cross-validation, we created an explanatory model by training it on all available data for each neuron using the selected α.

We computed the following evaluation metrics:

- Log-likelihood of the model (as above)

$$LL = \sum_{w} [y_w * log(\hat{y}_w) - \hat{y}_w - log(y_w!) \,]$$

- Deviance of the residuals for each word:

$$\square_{\square} = 2 * [\, \square_{\square} * \square\square\square(\square_{\square} / \square_{\square}) - (\square_{\square} - \square_{\square}) \,]$$

- Akaike Information Criterion (*AIC*):

$$\square\square\square = 2 * \square\square\square - 2 * \square\square$$

  where EDF is the effective degrees of freedom

$$\square\square\square\square\square((\square\square\ \square\square + \square\square)^{-1} * \square\square\ \square\square).$$

  For this equation W is a diagonal matrix of the models predicted firing rates $\square\square\square\square(\square\square, \square\square, \ldots)$ and X is the design matrix above.

Permutation Testing

To determine whether the observed model performance exceeded what would be expected by chance, we computed a null distribution of log-likelihoods using 100 permuted feature matrices. For each permutation:

1. A Poisson ridge model was fitted using optimal α determined by the true data
2. Log-likelihood was computed.

The empirical p-value was defined as

$$\square = (\#\ \square\square\ \square\square\_\square\square\square\square\square\square\square\square\square \geq \square\square\_\square\square\square\square) \,/\, 100$$

Software and Environment:

- **Language**: Python 3.12.9
- **Libraries**: PyTorch, NumPy, SciPy, scikit-learn, pandas
- **Platform**: Jupyter Notebook with GPU acceleration

Surprisal and Entropy

We quantified how predictable each word in the story was to a large language model according to previously described methods (Gwilliams et al. 2024; Weissbart and Martin 2024). For each word, $\square_{\square}$, we slid a progressive 1,024-token window through the transcript ending at $\square_{\square-1}$, fed this context into *GPT-2-large*, and extracted the probability distribution over the vocabulary for the current word, $\square_{\square}$. From this distribution, we computed two standard information-theoretic measures:

$$\square\square\square\square\square\square\square\square\square\square(\square_{\square}) = -\square\square\square\ \square(\square_{\square} \mid \square_{\square-1}, \square_{\square-2}, \ldots, \square_{\square-\square})$$

the unexpectedness of the actual word $\square_{\square}$, and

$$\square\square\square\square\square\square\square\square(\square_{\square}) = -\sum_{w \in V} P(w \mid \mathrm{w}_{i-k}, .., w_{i-1}) log\ P(w \mid \mathrm{w}_{i-k}, \ldots, w_{i-1})$$

which quantifies the model's uncertainty over all possible next words in the vocabulary (50,257 tokens in English for GPT-2). Tokens consisting only of punctuation were excluded from the calculation. Each word was thus assigned a single surprisal and entropy value based on its prior context.

Top Three Lexical Predictions

For every word $\square_{\square}$, we fed the preceding progressive 1024-token context into *GPT-2-Large* and obtained the next-token distribution:

$$\square(\cdot \mid \square_{1}, \ldots, \square_{\square})$$

At the immediate next position, we discarded punctuation-only tokens, ranked the remaining vocabulary by probability, and extracted the top three most likely lexical candidates:

$$\{\square_{\square+1}^{(1)}, \square_{\square+1}^{(2)}, \square_{\square+1}^{(3)}\}$$

along with their associated probabilities:

$$\square_{\square+1}^{(\square)} = \square(\square_{\square+1}^{(\square)} \mid \square_{1}, \ldots, \square_{\square}), \square = 1,2,3$$

The resulting six values per word—three predicted tokens and their probabilities—capture the model's uncertainty and lexical expectations about the immediate next word.

Dependency Grammar Relations

We derived governor–dependent pairs from each utterance using the Stanford CoreNLP dependency parser (Manning et al. 2014). After reconstructing sentences from the transcript, every sentence was sent to a locally hosted CoreNLP server where it was parsed on a locally hosted CoreNLP server; the parser outputs the directed, labeled relations of the Stanford Typed Dependencies scheme (de Marneffe and Manning 2008), closely aligned with Universal Dependencies (Nivre et al. 2020). Example links include **nsubj** (nominal subject)—e.g., ⟨*dog*, *barks*, *nsubj*⟩—and **dobj** (direct object)—e.g., ⟨*eat*, *apple*, *dobj*⟩—which respectively mark the clausal subject and the verb's direct object.

Because CoreNLP tokenizes contractions (e.g., "did", "n't") and multi-word expressions (e.g., "t","-", "shirt") differently from the manual transcript, we implemented a post-hoc merging routine: special three- and two-token patterns ("t","-", "shirt", "y'", "all") and common clitic suffixes (n't, 're, 've, etc.) were collapsed so that each governor or dependent indexed exactly one transcript word. Tokens that were purely punctuation or part of an internal merge were excluded, yielding a clean mapping from parser indices to global transcript indices. The resulting edge list,(governor, dependent, relation), for both "basic" and, when available, enhanced++ graphs (Schuster & Manning, 2016) were saved to Excel for downstream analyses.

Part of Speech Pairing & Subject and Object Pairing

We adopted the Stanford CoreNLP POS (part of speech) Tagger, which assigns fine-grained part-of-speech labels as per the standards of NLTK (Bird, Klein, and Loper 2009) (Bird et al., 2009) such as ***NN*** (common noun singular), ***NNS*** (plural noun), ***NNP*** (proper noun singular), ***NNPS*** (proper noun plural), ***VB*** (verb base form), ***VBD*** (verb past tense), ***VBZ*** (verb third-person singular), ***JJ*** (adjective), ***RB*** (adverb), and so forth. These granular labels were also grouped into broader categories (e.g., "Noun," "Verb," "Adjective," "Adverb") for ease of reference. The tagging process allowed us to identify each word's grammatical function without performing additional text pre-processing steps beyond basic sentence segmentation.

Next, for each sentence we queried a local CoreNLP server for enhanced++ dependency graphs (see **Dependency Grammar Relations**, de Marneffe & Manning, 2008). We built six part of speech-defined governor–dependent relations and one subject–object relations:

- Verb–Noun pairs, formed whenever a verb governed a noun via *nsubj, nsubj:pass, obj, dobj,* or *iobj*. The record notes whether the noun precedes or follows the verb, distinguishing subject-first ("noun ← verb") from object-after ("verb ← noun") configurations.

- Adjective–Noun links, identified through the *amod* label in either direction. We tagged each instance as *pre-nominal* (adjective before noun) or *post-nominal* (adjective after noun).

- Preposition–Noun relations, captured by *case*, which bind a preposition to the head noun of its prepositional phrase.

- Determiner–Noun relations, drawn from *det* edges, tracking classic determiner usage (e.g., *the* → *book*).

- Adverb–Verb modifiers, obtained from *advmod*, indicating how an adverb modifies its governing verb.

- Possessive Pronoun–Noun links, collected from *nmod:poss* and *poss* edges, recording bidirectional pronoun–noun possessives (e.g., *her* → *idea*).

- Subject–Object triplets, created only when the same verb simultaneously governed at least one *nsubj* (or *nsubj:pass*) and one *obj/dobj/iobj*.

For each such verb we paired every subject noun with every object noun and stored the relative order (subject-before-object or object-before-subject). For every pairing we logged the lexical forms, their row indices in the original transcript, the dependency label, sentence number, and the pre- vs. post-head position flag.

Opening Nodes, Closing Nodes, Syntactic depth

Following prior work (Gwilliams et al. 2024; Weissbart and Martin 2024), we treat syntactic depth as a simple count of how many constituent phrases remain open above a word in a Penn-Treebank parse:

□□□□□□□□□ □□□□□(□) = |□□□□ □□□□□ □□□□ (□)|

where □□□□ □□□□□ □□□□(□) is the sequence of non-terminal nodes (e.g., *S*, *VP*, *NP*) on the path from the leaf node *w* up to the root. In practical terms, this equals the number of open brackets at the position of *w* when the tree is read left-to-right. For example, in:

[□ [□□ □□□ □□□] [□□□□ □□□ [□□ □□□ [□□ □□□ □□□□□]]] [□□ □□□□□]]

boy has two open nodes (*S*, *NP*), so Depth = 2, whereas *apple* sits four levels deep (*S*, *SBAR*, *VP*, *NP*), giving Depth = 4. Alongside depth we extracted two complementary phrase-boundary features from the same parses: Opening nodes defined as the number of brackets that open at a word (i.e., new constituents that begin with that word), and closing nodes defined as the number of brackets that close at a word (constituents that terminate with that word).

Hierarchical Phrase Indexing

We developed a novel measure to capture the nested phrase structure that surrounds every word. We converted each Penn-Treebank parse of a sentence into a depth-by-word index matrix (see **Opening Nodes, Closing Nodes, Syntactic depth**). After collapsing unary chains and inserting a virtual *ROOT* node, every word w_i is located on a unique path:

$$(-1, \square_1, \square_2, \ldots, \square_{\square})$$

where -1 marks *ROOT* and each $\square_{\square}$ the index of the constituent that dominates w_i at depth $\square$. We store these local indices in a matrix $\square$ whose rows correspond to successive depths and whose columns correspond to transcript words. Formally:

$$\square_{\square,\square} = \begin{cases} i+1, & k=1 \ (surface\ word\ order) \\ position\ of\ w_i\ inside\ its\ depth{-}k\ phrase, & k \geq 2 \end{cases}$$

Thus, the first row lists the running word indices, while each deeper row re-indexes words relative to their surrounding phrase at that depth: identical numbers within a row indicate membership in the same NP, VP, CP, etc. Because the *ROOT* placeholder occupies Depth-1, top-level phrases (e.g., the sentence's main NP and VP) appear at Depth-2, ensuring that deeper levels faithfully reflect increasing syntactic embedding. For every sentence we exported its matrix to a dedicated worksheet and subsequently concatenated all worksheets into a sheet with a common row index (*Depth-1* … *Depth-N*). The resulting spreadsheet provides a compact, numeric representation of hierarchical constituencies.

Bibliography

Aitken, Fraser, and Peter Kok. 2022. “Hippocampal Representations Switch from Errors to Predictions during Acquisition of Predictive Associations.” *Nature Communications* 13 (1): 3294. https://doi.org/10.1038/s41467-022-31040-w.

Anderson, A. J., D. Kiela, J. R. Binder, L. Fernandino, C. J. Humphries, L. L. Conant, and E. C. Lalor. 2021. “Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning.” *Journal of Neuroscience* 41 (18): 4100–4119.

Binder, Jeffrey R., and Rutvik H. Desai. 2011. “The Neurobiology of Semantic Memory.” *Trends in Cognitive Sciences* 15 (11): 527–36. https://doi.org/10.1016/j.tics.2011.10.001.

Binder, Jeffrey R., Rutvik H. Desai, William W. Graves, and Lisa L. Conant. 2009. “Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies.” *Cerebral Cortex (New York, N.Y.: 1991)* 19 (12): 2767–96. https://doi.org/10.1093/cercor/bhp055.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc.

Boersma, P. 2011. “Praat: Doing Phonetics by Computer [Computer Program].” *Http://Www. Praat. Org/*.

Bubic, Andreja, D. Yves Von Cramon, and Ricarda I. Schubotz. 2010. “Prediction, Cognition and the Brain.” *Frontiers in Human Neuroscience* 4 (March). https://doi.org/10.3389/fnhum.2010.00025.

Carpenter, Adam F., Gabriel Baud-Bovy, Apostolos P. Georgopoulos, and Giuseppe Pellizzer. 2018. “Encoding of Serial Order in Working Memory: Neuronal Activity in Motor, Premotor, and Prefrontal Cortex during a Memory Scanning Task.” *The Journal of Neuroscience* 38 (21): 4912–33. https://doi.org/10.1523/JNEUROSCI.3294-17.2018.

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Remi King. 2021. “Disentangling Syntax and Semantics in the Brain with Deep Networks.” arXiv. https://doi.org/10.48550/arXiv.2103.01620.

Cisek, Paul, and Alexandre Pastor-Bernier. 2014. “On the Challenges and Mechanisms of Embodied Decisions.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369 (1655): 20130479. https://doi.org/10.1098/rstb.2013.0479.

Clark, Herbert H. 1996. *Using Language*. “Using” Linguistic Books. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620539.

Comrie, Alison E., Emily J. Monroe, Ari E. Kahn, Eric L. Denovellis, Abhilasha Joshi, Jennifer A. Guidera, Timothy A. Krausz, Joshua D. Berke, Nathaniel D. Daw, and Loren M. Frank. 2024. “Hippocampal Representations of Alternative Possibilities Are Flexibly Generated to Meet Cognitive Demands.” *bioRxiv: The Preprint Server for Biology*, September, 2024.09.23.613567. https://doi.org/10.1101/2024.09.23.613567.

Cong, Shan, Hang Wang, Yang Zhou, Zheng Wang, Xiaohui Yao, and Chunsheng Yang. 2024. “Comprehensive Review of Transformer-Based Models in Neuroscience, Neurology, and Psychiatry.” *Brain-X* 2 (2): e57. https://doi.org/10.1002/brx2.57.

Constantinescu, Alexandra O., Jill X. O’Reilly, and Timothy E. J. Behrens. 2016. “Organizing Conceptual Knowledge in Humans with a Gridlike Code.” *Science (New York, N.Y.)* 352 (6292): 1464–68. https://doi.org/10.1126/science.aaf0941.

Desbordes, Thibault, Jean-Rémi King, and Stanislas Dehaene. 2024. “Tracking the Neural Codes for Words and Phrases during Semantic Composition, Working-Memory Storage, and Retrieval.” *Cell Reports* 43 (3).

Desbordes, Thibault, Yair Lakretz, Valentin Chanoine, Maxime Oquab, Jean-Marc Badier, Antoine Trébuchon, and Jean-Rémi King. 2023. “Dimensionality and Ramping: Signatures of Sentence Integration in the Dynamics of Brains and Deep Language

Models." *Journal of Neuroscience* 43 (29): 5350–64.
Dijksterhuis, D. E., M. W. Self, J. K. Possel, J. C. Peters, E. C. W. van Straaten, S. Idema, J. C. Baaijen, et al. 2024. "Pronouns Reactivate Conceptual Representations in Human Hippocampal Neurons." *Science* 385 (6716): 1478–84. https://doi.org/10.1126/science.adr2813.
Ding, Nai. 2025. "Sequence Chunking through Neural Encoding of Ordinal Positions." *Trends in Cognitive Sciences*, February. https://doi.org/10.1016/j.tics.2025.01.014.
Donoghue, Thomas, Matar Haller, Erik J. Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, et al. 2020. "Parameterizing Neural Power Spectra into Periodic and Aperiodic Components." *Nature Neuroscience* 23 (12): 1655–65. https://doi.org/10.1038/s41593-020-00744-x.
Dragoi, George. 2024. "The Generative Grammar of the Brain: A Critique of Internally Generated Representations." *Nature Reviews. Neuroscience* 25 (1): 60–75. https://doi.org/10.1038/s41583-023-00763-0.
Duff, Melissa C., and Sarah Brown-Schmidt. 2012. "The Hippocampus and the Flexible Use and Processing of Language." *Frontiers in Human Neuroscience* 6 (April):69. https://doi.org/10.3389/fnhum.2012.00069.
Duff, Melissa C., Natalie V. Covington, Caitlin Hilverman, and Neal J. Cohen. 2020. "Semantic Memory and the Hippocampus: Revisiting, Reaffirming, and Extending the Reach of Their Critical Relationship." *Frontiers in Human Neuroscience* 13 (January). https://doi.org/10.3389/fnhum.2019.00471.
Fedorenko, Evelina, Anna A. Ivanova, and Tamar I. Regev. 2024. "The Language Network as a Natural Kind within the Broader Landscape of the Human Brain." *Nature Reviews Neuroscience* 25 (5): 289–312. https://doi.org/10.1038/s41583-024-00802-4.
Firth, J. R. 1968. *Selected Papers of J.R. Firth, 1952-59.* Indiana University Studies in the History and Theory of Linguistics. Bloomington: Indiana University Press.
Fló, Ana. 2021. "Evidence of Ordinal Position Encoding of Sequences Extracted from Continuous Speech." *Cognition*, Special Issue in Honour of Jacques Mehler, Cognition's founding editor, 213 (August):104646. https://doi.org/10.1016/j.cognition.2021.104646.
Franch, Melissa, Elizabeth A. Mickiewicz, James Belanger, Assia Chericoni, Ana G. Chavez, Kalman Katlowitz, Raissa Mathura, Eleonora Bartoli, Suzanne Kemmer, Steven T. Piantadosi. et. al. 2025. "A Vectorial Code for Semantics in Human Hippocampus." *bioRxiv*, 2025–02.
Fyshe, A., G. Sudre, L. Wehbe, N. Rafidi, and T. M. Mitchell. 2019. "The Lexical Semantics of Adjective–Noun Phrases in the Human Brain." *Human Brain Mapping* 40 (15): 4457–69.
Gerber, Isaac. 2025. "Attention Is Not All You Need: The Importance of Feedforward Networks in Transformer Models." arXiv. https://doi.org/10.48550/arXiv.2505.06633.
Gold, April E., and Raymond P. Kesner. 2005. "The Role of the CA3 Subregion of the Dorsal Hippocampus in Spatial Pattern Completion in the Rat." *Hippocampus* 15 (6): 808–14. https://doi.org/10.1002/hipo.20103.
Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, et al. 2021. "Thinking Ahead: Spontaneous next Word Predictions in Context as a Keystone of Language in Humans and Machines." bioRxiv. https://doi.org/10.1101/2020.12.02.403477.
Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K.A., Devinsky, O., Hasson, U., 2022. Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. 25, 369–380. https://doi.org/10.1038/s41593-022-01026-4
Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian,

Ahmad Al-Dahle, Aiesha Letman, et al. 2024. “The Llama 3 Herd of Models.” arXiv. https://doi.org/10.48550/arXiv.2407.21783.
Gwilliams, Laura, Alec Marantz, David Poeppel, and Jean-Remi King. 2024. “Hierarchical Dynamic Coding Coordinates Speech Comprehension in the Brain.” bioRxiv. https://doi.org/10.1101/2024.04.19.590280.
Happé, Francesca G. E. 1997. “Central Coherence and Theory of Mind in Autism: Reading Homographs in Context.” *British Journal of Developmental Psychology* 15 (1): 1–12. https://doi.org/10.1111/j.2044-835X.1997.tb00721.x.
He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. “Deberta: Decoding-Enhanced BERT with Disentangled Attention.” *arXiv Preprint* arXiv:2006.03654.
Heilbron, Micha, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. 2022. “A Hierarchy of Linguistic Predictions during Natural Language Comprehension.” *Proceedings of the National Academy of Sciences* 119 (32): e2201968119. https://doi.org/10.1073/pnas.2201968119.
Hickok, Gregory, and David Poeppel. 2007. “The Cortical Organization of Speech Processing.” *Nature Reviews Neuroscience* 8 (5): 393–402. https://doi.org/10.1038/nrn2113.
Honari-Jahromi, M., B. Chouinard, E. Blanco-Elorrieta, L. Pylkkänen, and A. Fyshe. 2021. “Neural Representation of Words within Phrases: Temporal Evolution of Color-Adjectives and Object-Nouns during Simple Composition.” *PloS One* 16 (3): e0242754.
Hong, Zhuoqiao, Haocheng Wang, Zaid Zada, Harshvardhan Gazula, David Turner, Bobbi Aubrey, Leonard Niekerken, et al. 2024. “Scale Matters: Large Language Models with Billions (Rather than Millions) of Parameters Better Match Neural Representations of Natural Language.” *eLife* 13 (October). https://doi.org/10.7554/eLife.101204.1.
Howard, Marc W., Daniel S. Rizzuto, Jeremy B. Caplan, Joseph R. Madsen, John Lisman, Richard Aschenbrenner-Scheibe, Andreas Schulze-Bonhage, and Michael J. Kahana. 2003. “Gamma Oscillations Correlate with Working Memory Load in Humans.” *Cerebral Cortex (New York, N.Y.: 1991)* 13 (12): 1369–74. https://doi.org/10.1093/cercor/bhg084.
Jain, Shailee, and Alexander Huth. 2018. “Incorporating Context into Language Encoding Models for fMRI.” In *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html.
Jamali, Mohsen, Benjamin L Grannan, Jing Cai, Arjun R Khanna, William Muñoz, Irene Caprara, A Paulk, Sydney S Cash, Evelina Fedorenko, and Ziv M Williams. 2024. “Semantic Encoding during Language Comprehension at Single-Cell Resolution.” *Nature* 631 (July):610–16.
Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. “Bag of Tricks for Efficient Text Classification.” arXiv. https://doi.org/10.48550/arXiv.1607.01759.
Kaiser, Daniel, Gabriele Inciuraite, and Radoslaw M. Cichy. 2020. “Rapid Contextualization of Fragmented Scene Information in the Human Visual System.” bioRxiv. https://doi.org/10.1101/2020.01.06.895870.
Katlowitz, Kalman A., Shraddha Shah, Melissa C. Franch, Joshua Adkinson, James L. Belanger, Raissa K. Mathura, Domokos Meszéna, et al. 2025. “Learning and Language in the Unconscious Human Hippocampus.” *bioRxiv*, April, 2025.04.09.648012. https://doi.org/10.1101/2025.04.09.648012.
Kingma, Diederik P., and Jimmy Ba. 2017. “Adam: A Method for Stochastic Optimization.” arXiv. https://doi.org/10.48550/arXiv.1412.6980.
Kirwan, C. Brock, Andrew Hartshorn, Shauna M. Stark, Naomi J. Goodrich-Hunsaker, Ramona O. Hopkins, and Craig E. L. Stark. 2012. “Pattern Separation Deficits Following Damage to the Hippocampus.” *Neuropsychologia* 50 (10): 2408–14. https://doi.org/10.1016/j.neuropsychologia.2012.06.011.
Kozhevnikov, Alexay A., and Michale S. Fee. 2007. “Singing-Related Activity of Identified HVC

Neurons in the Zebra Finch." *Journal of Neurophysiology* 97 (6): 4271–83. https://doi.org/10.1152/jn.00952.2006.
Kumar, Sreejan, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. 2024. "Shared Functional Specialization in Transformer-Based Language Models and the Human Brain." *Nature Communications* 15 (1): 5523. https://doi.org/10.1038/s41467-024-49173-5.
Kuperberg, Gina R. 2010. "Language in Schizophrenia Part 1: An Introduction." *Language and Linguistics Compass* 4 (8): 576–89. https://doi.org/10.1111/j.1749-818X.2010.00216.x.
Leutgeb, Jill K., Stefan Leutgeb, May-Britt Moser, and Edvard I. Moser. 2007. "Pattern Separation in the Dentate Gyrus and CA3 of the Hippocampus." *Science* 315 (5814): 961–66. https://doi.org/10.1126/science.1135801.
Li, Jixing, and Liina Pylkkänen. 2021. "Disentangling Semantic Composition and Semantic Association in the Left Temporal Lobe." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 41 (30): 6526–38. https://doi.org/10.1523/JNEUROSCI.2317-20.2021.
Lisman, John E., and Ole Jensen. 2013. "The Theta-Gamma Neural Code." *Neuron* 77 (6): 1002–16. https://doi.org/10.1016/j.neuron.2013.03.007.
Liu, Q, Matt J. Kusner, and Phil Blunsom. 2020. "A Survey on Contextual Embeddings." arXiv. https://doi.org/10.48550/arXiv.2003.07278.
Liu, Yuhang, Dong Gong, Yichao Cai, Erdun Gao, Zhen Zhang, Biwei Huang, Mingming Gong, Anton van den Hengel, and Javen Qinfeng Shi. 2025. "I Predict Therefore I Am: Is Next Token Prediction Enough to Learn Human-Interpretable Concepts from Data?" arXiv. https://doi.org/10.48550/arXiv.2503.08980.
MacDonald, Christopher J., Kyle Q. Lepage, Uri T. Eden, and Howard Eichenbaum. 2011. "Hippocampal 'Time Cells' Bridge the Gap in Memory for Discontiguous Events." *Neuron* 71 (4): 737–49. https://doi.org/10.1016/j.neuron.2011.07.012.
MacKay, Donald G., Rachel Stewart, and Deborah M. Burke. 1998. "H.M. Revisited: Relations between Language Comprehension, Memory, and the Hippocampal System." *Journal of Cognitive Neuroscience* 10 (3): 377–94. https://doi.org/10.1162/089892998562807.
Magnotti, John F., Zhengjia Wang, and Michael S. Beauchamp. 2020. "RAVE: Comprehensive Open-Source Software for Reproducible Analysis and Visualization of Intracranial EEG Data." *NeuroImage* 223 (December):117341. https://doi.org/10.1016/j.neuroimage.2020.117341.
Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit." In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.
Manns, Joseph R., Ramona O. Hopkins, and Larry R. Squire. 2003. "Semantic Memory and the Human Hippocampus." *Neuron* 38 (1): 127–33. https://doi.org/10.1016/S0896-6273(03)00146-6.
Maren, Stephen, K. Luan Phan, and Israel Liberzon. 2013. "The Contextual Brain: Implications for Fear Conditioning, Extinction and Psychopathology." *Nature Reviews Neuroscience* 14 (6): 417–28. https://doi.org/10.1038/nrn3492.
Marneffe, Marie-Catherine de, and Christopher D. Manning. 2008. "The Stanford Typed Dependencies Representation." In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8.
Marr, D. 1971. "Simple Memory: A Theory for Archicortex." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 262 (841). https://doi.org/10.1098/rstb.1971.0078.
McClelland, James L., and Timothy T. Rogers. 2003. "The Parallel Distributed Processing Approach to Semantic Cognition." *Nature Reviews Neuroscience* 4 (4): 310–22.

https://doi.org/10.1038/nrn1076.
Merlin, Gabriele, and Mariya Toneva. 2024. “Language Models and Brains Align Due to More than Next-Word Prediction and Word-Level Information.” arXiv. https://doi.org/10.48550/arXiv.2212.00596.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” arXiv. https://doi.org/10.48550/arXiv.1301.3781.
Minda, John Paul, and J. David Smith. 2001. “Prototypes in Category Learning: The Effects of Category Size, Category Structure, and Stimulus Complexity.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27 (3): 775–99. https://doi.org/10.1037/0278-7393.27.3.775.
Mitchell, Jeff, and Mirella Lapata. 2008. “Vector-Based Models of Semantic Composition.” In *Proceedings of ACL-08: HLT*, edited by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, 236–44. Columbus, Ohio: Association for Computational Linguistics. https://aclanthology.org/P08-1028/.
Mizumori, Sheri J. Y. 2013. “Context Prediction Analysis and Episodic Memory.” *Frontiers in Behavioral Neuroscience* 7 (October). https://doi.org/10.3389/fnbeh.2013.00132.
Monteiro, Jânio, André Pedro, and António João Silva. 2022. “A Gray Code Model for the Encoding of Grid Cells in the Entorhinal Cortex.” *Neural Computing and Applications* 34 (3): 2287–2306. https://doi.org/10.1007/s00521-021-06482-w.
Moser, Edvard I., Emilio Kropff, and May-Britt Moser. 2008. “Place Cells, Grid Cells, and the Brain’s Spatial Representation System.” *Annual Review of Neuroscience* 31:69–89. https://doi.org/10.1146/annurev.neuro.31.061307.090723.
Naya, Yuji, and Wendy A. Suzuki. 2011. “Integrating What and When across the Primate Medial Temporal Lobe.” *Science (New York, N.Y.)* 333 (6043): 773–76. https://doi.org/10.1126/science.1206773.
Nelson, Matthew J., Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S. Cash, et al. 2017. “Neurophysiological Dynamics of Phrase-Structure Building during Sentence Processing.” *Proceedings of the National Academy of Sciences* 114 (18): E3669–78. https://doi.org/10.1073/pnas.1701590114.
Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, and Daniel Zeman. 2020. “Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection.” *arXiv Preprint* arXiv:2004.10643.
Olatunji, Ibukun, and Mark Sheppard. 2025. “Next Token Prediction Is a Dead End for Creativity.” arXiv. https://doi.org/10.48550/arXiv.2505.19277.
Oota, Subba Reddy, Manish Gupta, and Mariya Toneva. 2023. “Joint Processing of Linguistic Properties in Brains and Language Models.” arXiv. https://doi.org/10.48550/arXiv.2212.08094.
Opitz, Bertram, and Angela D. Friederici. 2003. “Interactions of the Hippocampal System and the Prefrontal Cortex in Learning Language-like Rules.” *NeuroImage* 19 (4): 1730–37. https://doi.org/10.1016/s1053-8119(03)00170-8.
Park, Seongmin A., Douglas S. Miller, and Erie D. Boorman. 2021. “Inferences on a Multidimensional Social Hierarchy Use a Grid-like Code.” *Nature Neuroscience* 24 (9): 1292–1301. https://doi.org/10.1038/s41593-021-00916-3.
Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.
Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans, 1532–43. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Piai, Vitória, Kristopher L. Anderson, Jack J. Lin, Callum Dewar, Josef Parvizi, Nina F. Dronkers, and Robert T. Knight. 2016. “Direct Brain Recordings Reveal Hippocampal Rhythm Underpinnings of Language Processing.” *Proceedings of the National Academy of Sciences* 113 (40): 11366–71. https://doi.org/10.1073/pnas.1603312113.

Piantadosi, Steven T., Dyana C.Y. Muller, Joshua S. Rule, Karthikeya Kaushik, Mark Gorenstein, Elena R. Leib, and Emily Sanford. 2024. “Why Concepts Are (Probably) Vectors.” *Trends in Cognitive Sciences* 28 (9): 844–56. https://doi.org/10.1016/j.tics.2024.06.011.

Pylkkänen, Liina. 2019. “Neural Basis of Basic Composition: What We Have Learned from the Red–Boat Studies and Their Extensions.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1791): 20190299. https://doi.org/10.1098/rstb.2019.0299.

Radford, Alec, Jeff Wu, R. Child, D. Luan, Dario Amodei, and I. Sutskever. 2019. “Language Models Are Unsupervised Multitask Learners.” In . https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Raiaan, Mohaimenul Azam Khan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges.” *IEEE Access* 12:26839–74. https://doi.org/10.1109/ACCESS.2024.3365742.

Sandoval-Segura, Pedro, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. 2025. “Using Attention Sinks to Identify and Evaluate Dormant Heads in Pretrained LLMs.” arXiv. https://doi.org/10.48550/arXiv.2504.03889.

Schönmann, Inés, Jakub Szewczyk, Floris P. de Lange, and Micha Heilbron. 2025. “Stimulus Dependencies—Rather than next-Word Prediction—Can Explain Pre-Onset Brain Encoding during Natural Listening.” bioRxiv. https://doi.org/10.1101/2025.03.08.642140.

Schrimpf, Martin, 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. https://doi.org/10.1073/pnas.2105646118.

Shepard, Roger N. 1980. “Multidimensional Scaling, Tree-Fitting, and Clustering.” *Science* 210 (4468): 390–98. https://doi.org/10.1126/science.210.4468.390.

Smith, David M., and David A. Bulkin. 2014. “The Form and Function of Hippocampal Context Representations.” *Neuroscience and Biobehavioral Reviews* 0 (March):52–61. https://doi.org/10.1016/j.neubiorev.2014.01.005.

Sperber, Dan, and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.

Su, Jianlin, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. “RoFormer: Enhanced Transformer with Rotary Position Embedding.” arXiv. https://doi.org/10.48550/arXiv.2104.09864.

Toneva, Mariya, Tom M. Mitchell, and Leila Wehbe. 2022. “Combining Computational Controls with Natural Text Reveals Aspects of Meaning Composition.” *Nature Computational Science* 2 (11): 745–57. https://doi.org/10.1038/s43588-022-00354-6.

Toneva, Mariya, and Leila Wehbe. 2019. “Interpreting and Improving Natural-Language Processing (in Machines) with Natural Language-Processing (in the Brain).” In *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html.

Turney, Peter D., and Patrick Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research* 37:141–88.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. “Attention Is All You Need.” arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Vig, Jesse. 2019. “A Multiscale Visualization of Attention in the Transformer Model.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, edited by Marta R. Costa-jussà and Enrique Alfonseca, 37–42. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3007.

Wang, Zhengjia, John F. Magnotti, Xiang Zhang, and Michael S. Beauchamp. 2023. “YAEL: Your Advanced Electrode Localizer.” *eNeuro* 10 (10). https://doi.org/10.1523/ENEURO.0328-23.2023.

Weissbart, Hugo, and Andrea E. Martin. 2024. “The Structure and Statistics of Language Jointly Shape Cross-Frequency Neural Dynamics during Spoken Language Comprehension.” *Nature Communications* 15 (1): 8850. https://doi.org/10.1038/s41467-024-53128-1.

Whittington, James C. R., Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. 2020. “The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation.” *Cell* 183 (5): 1249-1263.e23. https://doi.org/10.1016/j.cell.2020.10.024.

Willems, Roel M., and Marius V. Peelen. 2021. “How Context Changes the Neural Basis of Perception and Language.” *iScience* 24 (5): 102392. https://doi.org/10.1016/j.isci.2021.102392.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. “Transformers: State-of-the-Art Natural Language Processing.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by Qun Liu and David Schlangen, 38–45. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

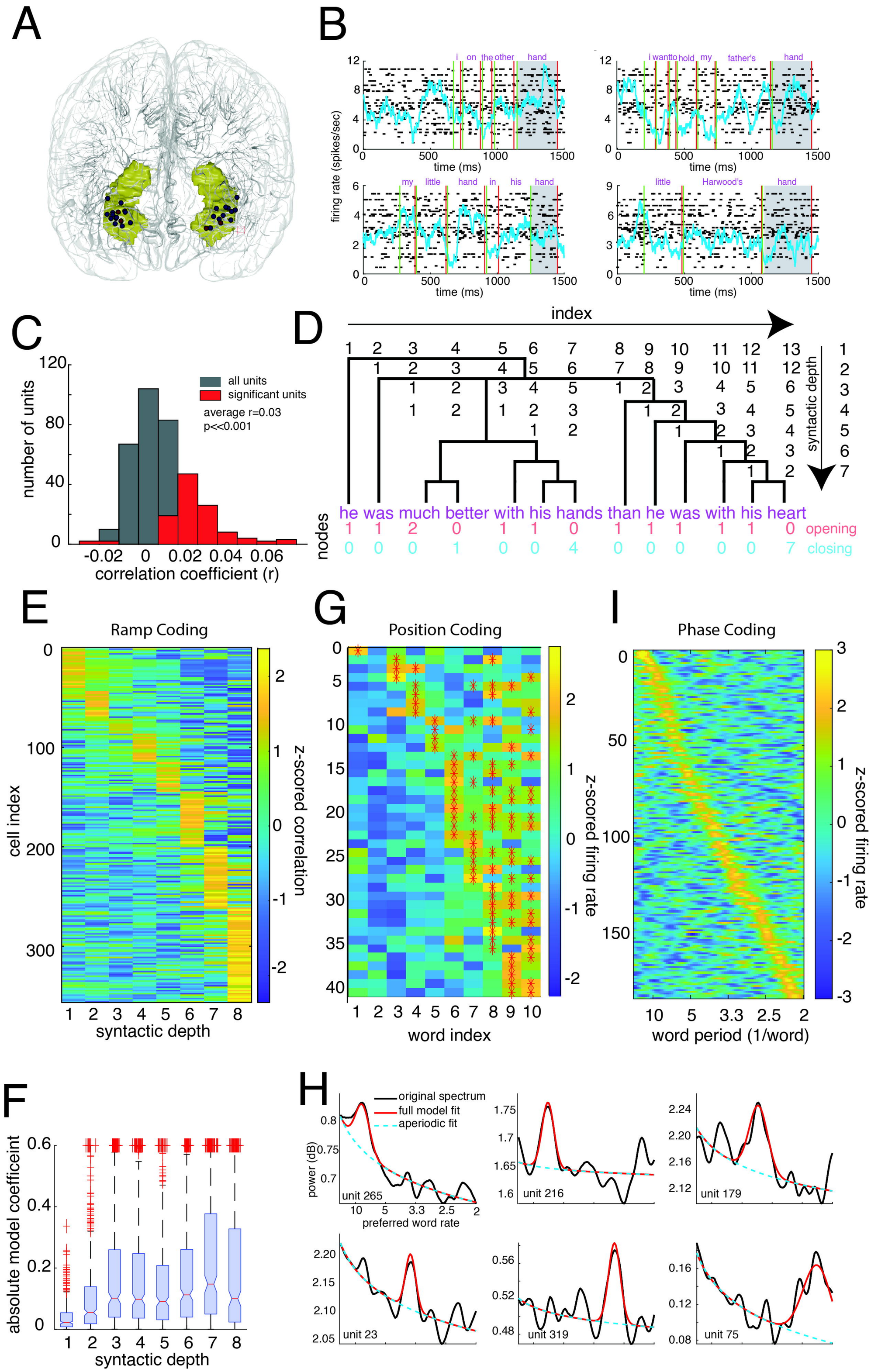
Yassa, Michael A., and Craig E. L. Stark. 2011. “Pattern Separation in the Hippocampus.” *Trends in Neurosciences* 34 (10): 515–25. https://doi.org/10.1016/j.tins.2011.06.006.

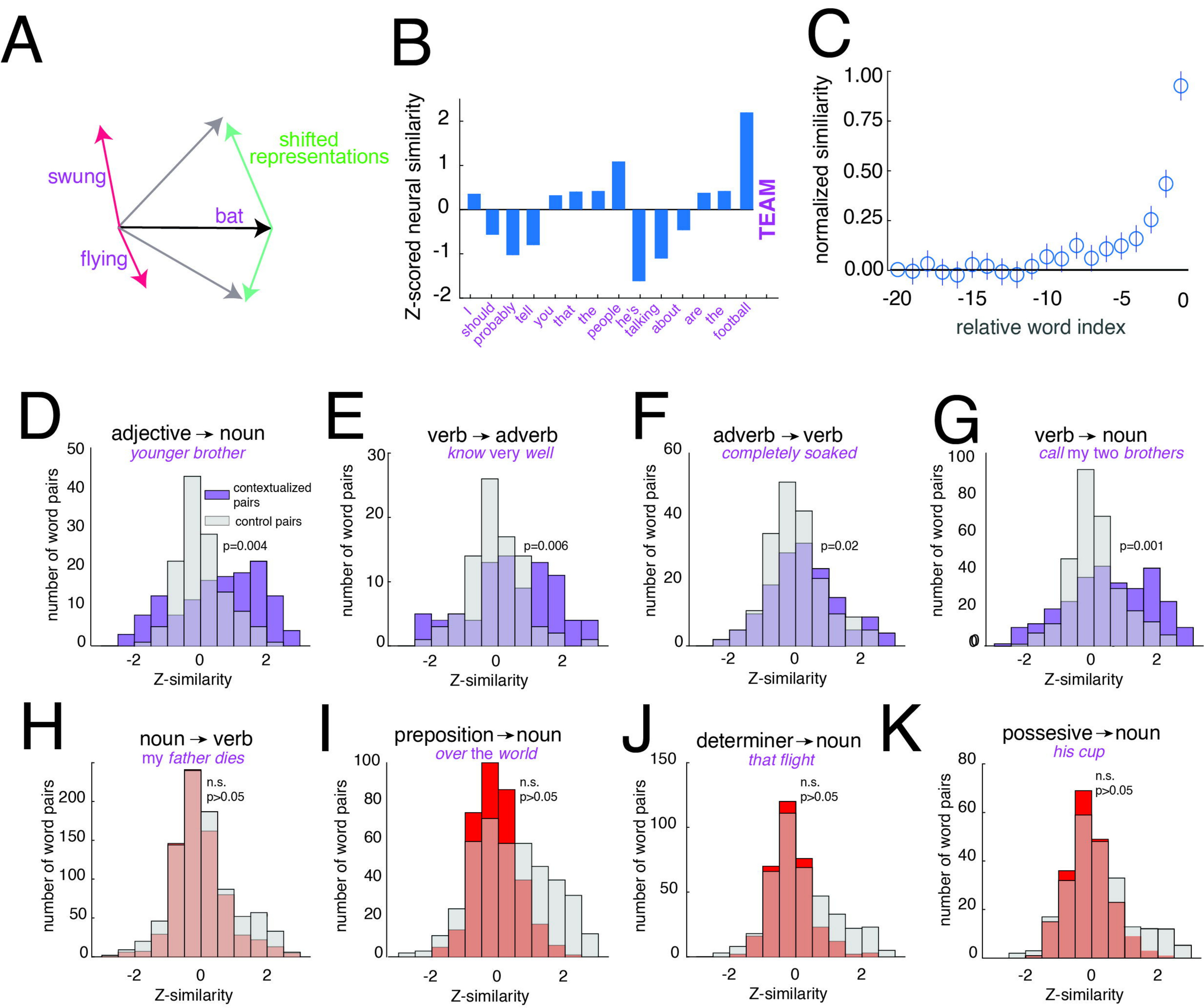
Zhao, Jingya, Ruiqi Gao, and Jonathan R. Brennan. 2025. “Decoding the Neural Dynamics of Headed Syntactic Structure Building.” *Journal of Neuroscience* 45 (17).

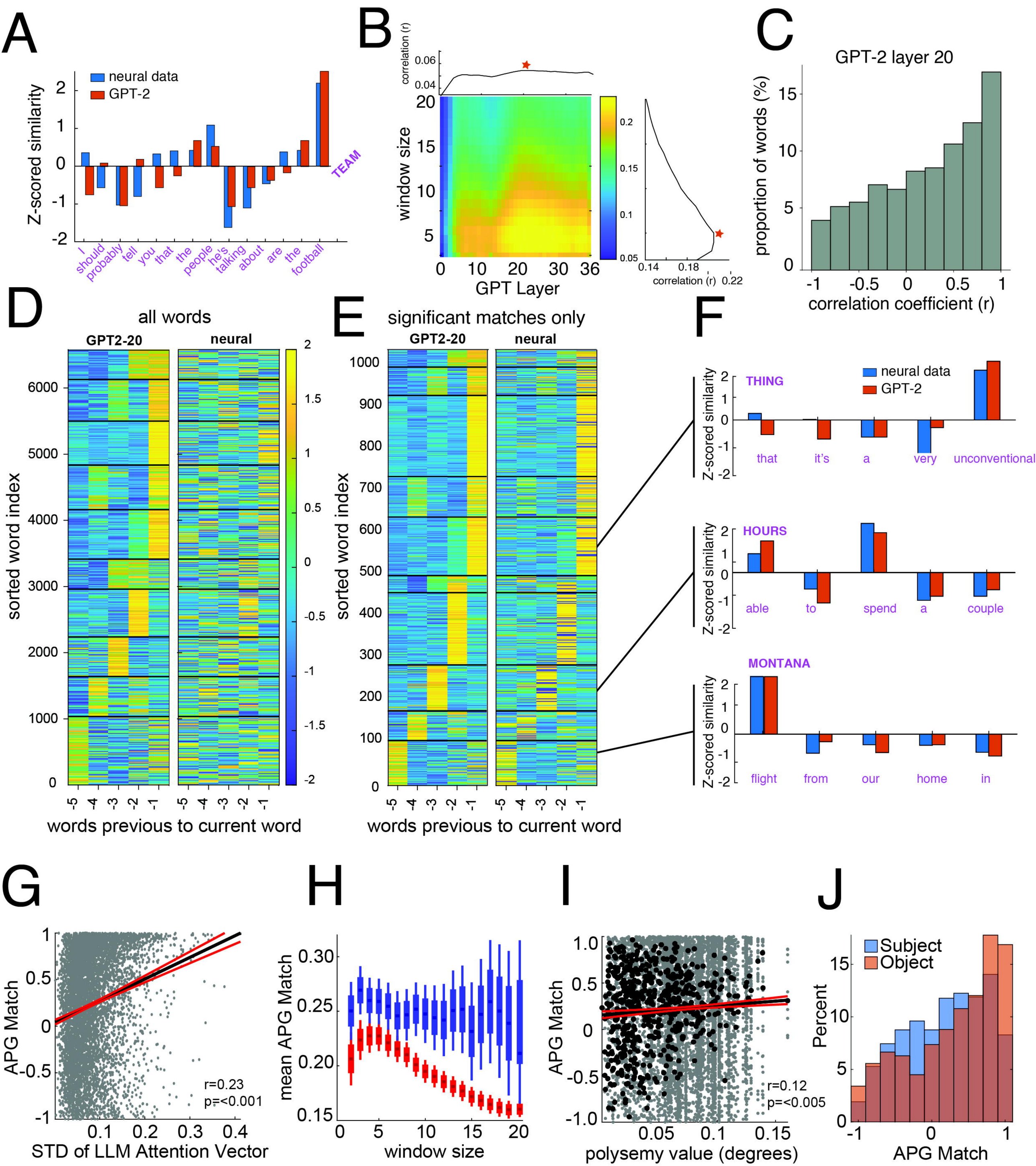
Zhao, Liang, Xiachong Feng, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. 2024. “Length Extrapolation of Transformers: A Survey from the Perspective of Positional Encoding.” In *Findings of the Association for Computational Linguistics: EMNLP 2024*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 9959–77. Miami, Florida, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.582.

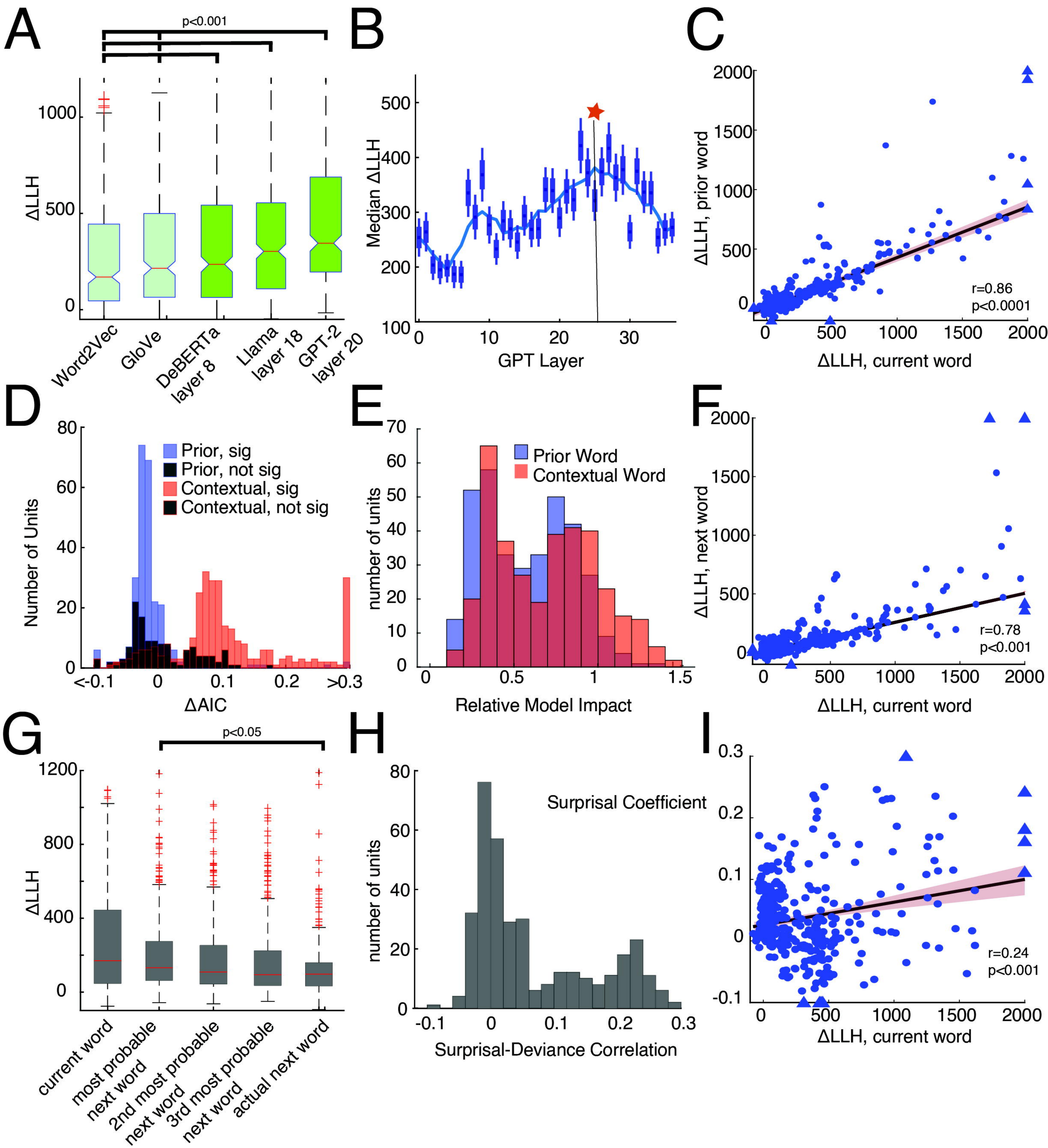
Zheng, Jianqiao, Sameera Ramasinghe, and Simon Lucey. 2021. “Rethinking Positional Encoding.” arXiv. https://doi.org/10.48550/arXiv.2107.02561.

Zuanazzi, Andrea, Pablo Ripollés, Wei Mei Lin, Laura Gwilliams, Jean-Rémi King, and David Poeppel. 2024. “Negation Mitigates Rather than Inverts the Neural Representations of Adjectives.” *PLoS Biology* 22 (5): e3002622.


A
B
firing rate (spikes/sec)
time (ms)
i on the other hand
i wantto hold my father's hand
my little hand in his hand
little Harwood's hand
C
number of units
correlation coefficient (r)
all units
significant units
average r=0.03
p<<0.001
D
index
syntactic depth
he was much better with his hands than he was with his heart
nodes
opening
closing
E
Ramp Coding
cell index
syntactic depth
z-scored correlation
G
Position Coding
word index
z-scored firing rate
I
Phase Coding
word period (1/word)
z-scored firing rate
F
absolute model coefficeint
syntactic depth
H
original spectrum
full model fit
aperiodic fit
power (dB)
preferred word rate
unit 265
unit 216
unit 179
unit 23
unit 319
unit 75


A
swung
bat
flying
shifted representations
B
Z-scored neural similarity
2
1
0
-1
-2
I
should
probably
tell
you
that
the
people
he's
talking
about
are
the
football
TEAM
C
normalized similarity
1.00
0.75
0.50
0.25
0.00
-20
-15
-10
-5
0
relative word index
D
adjective → noun
younger brother
number of word pairs
contextualized pairs
control pairs
p=0.004
Z-similarity
E
verb → adverb
know very well
p=0.006
F
adverb → verb
completely soaked
p=0.02
G
verb → noun
call my two brothers
p=0.001
H
noun → verb
my father dies
n.s.
p>0.05
I
preposition → noun
over the world
J
determiner → noun
that flight
K
possesive → noun
his cup


A
neural data
GPT-2
Z-scored similarity
I
should
probably
tell
you
that
the
people
he's
talking
about
are
the
football
TEAM
B
correlation (r)
window size
GPT Layer
C
GPT-2 layer 20
proportion of words (%)
correlation coefficient (r)
D
all words
GPT2-20
neural
sorted word index
words previous to current word
E
significant matches only
F
THING
that
it's
a
very
unconventional
HOURS
able
to
spend
a
couple
MONTANA
flight
from
our
home
in
G
APG Match
STD of LLM Attention Vector
r=0.23
p=<0.001
H
mean APG Match
window size
I
polysemy value (degrees)
r=0.12
p=<0.005
J
Subject
Object
Percent
APG Match


A
p<0.001
ΔLLH
Word2Vec
GloVe
DeBERTa layer 8
Llama layer 18
GPT-2 layer 20
B
Median ΔLLH
GPT Layer
C
ΔLLH, prior word
ΔLLH, current word
r=0.86
p<0.0001
D
Prior, sig
Prior, not sig
Contextual, sig
Contextual, not sig
Number of Units
ΔAIC
E
Prior Word
Contextual Word
number of units
Relative Model Impact
F
ΔLLH, next word
ΔLLH, current word
r=0.78
p<0.001
G
p<0.05
ΔLLH
current word
most probable next word
2nd most probable next word
3rd most probable next word
actual next word
H
Surprisal Coefficient
number of units
Surprisal-Deviance Correlation
I
ΔLLH, current word
r=0.24
p<0.001