A vectorial code for semantics in human hippocampus

Melissa Franch[1], Elizabeth A. Mickiewicz[1], James L. Belanger[1,2], Assia Chericoni[1], Ana G. Chavez[1], Kalman A. Katlowitz[1], Raissa Mathura[1], Eleonora Bartoli[1], Suzanne Kemmer[2], Danika Paulo[1], Steven T. Piantadosi[3], 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, Houston, Texas, United States of America
2. Department of Linguistics, Rice University
3. Department of Psychology, University of California Berkeley

* Correspondence: Benjamin.Hayden@bcm.edu

Funding statement
This research was supported by the McNair Foundation and by NIH R01 MH129439.

Competing interests
The authors have no competing interests to declare.

Acknowledgements
We thank Jeff Johnston, Joshua Adkinson, Victoria Gates, and George Kokalas for invaluable assistance.

Data availability
The data that support the findings of this study are available from the corresponding authors upon reasonable request.

ABSTRACT

As we listen to speech, our brains actively compute the meaning of individual words. Inspired by the success of large language models (LLMs), we hypothesized that the brain employs vectorial coding principles, such that meaning is reflected in distributed activity of single neurons. We recorded responses of hundreds of neurons in the human hippocampus, which has a well-established role in semantic coding, while participants listened to narrative speech. We find encoding of contextual word meaning in the simultaneous activity of neurons whose individual selectivities span multiple unrelated semantic categories. Like embedding vectors in semantic models, distance between neural population responses correlates with semantic distance; however, this effect was only observed in contextual embedding models (like BERT) and was reversed in non-contextual embedding models (like Word2Vec), suggesting that the semantic distance effect depends critically on contextualization. Moreover, for the subset of highly semantically similar words, even contextual embedders showed an inverse correlation between semantic and neural distances; we attribute this pattern to the noise-mitigating benefits of contrastive coding. Finally, in further support for the critical role of context, we find that range of neural responses covaries with lexical polysemy. Ultimately, these results support the hypothesis that semantic coding in the hippocampus follows vectorial principles.

INTRODUCTION

As we listen to speech, we typically identify the meanings (semantics) of the words we hear. The principles by which neurons encode meaning remain poorly understood. One hypothesis is that semantic coding is fundamentally vectorial, so that word meanings are encoded in patterns of activation across a large number of distinct units (Frisby et al., 2023; Goldstein et al., 2022, 2024; Landauer & Dumais, 1997; Mitchell & Lapata, 2008; Piantadosi et al., 2024; Pulvermüller, 1999; Rogers & Mcclelland, 2004). Interest in vectorial encoding has grown in recent years because it is a core element of large language models (LLMs), which have shown great promise in a variety of applications (Devlin et al., 2018; Joulin et al., 2016; Mikolov, Chen, et al., 2013; Mikolov et al., 2013; Pennington et al., 2014; Radford et al., 2019; Vaswani et al., 2017). Inspired by the observation that such models can adroitly manipulate semantic information, we hypothesized that semantics in the human brain is also vectorial (Linzen & Baroni, 2025; Manning et al., 2020; Pavlick, 2025).

Typically, word vectors are composed of embedding values as units, but these units could be firing rates from neurons in the brain (Jamali et al., 2024). In a standard vectorial code, units do not have an identifiable semantic meaning; instead, they tend to have broad and complex response fields that encompass many apparently unrelated concepts (Frisby et al., 2023; Landauer & Dumais, 1997; McClelland & Rogers, 2003; Mitchell & Lapata, 2008; Pulvermüller, 1999b, 1999a; Rogers & McClelland, 2004). Likewise, a given word's meaning reflects the concurrent activation of a variety of units that are also selective for other words. In contrast, a good deal of evidence from the hippocampus supports the hypothesis that neurons are selective to single discrete concepts (Bausch et al., 2021,Quian Quiroga, 2023; Quian Quiroga et al., 2005 and 2009; Rey et al., 2025), consistent with ideas that hippocampus has highly sparse codes and narrow selectivities (Kolibius et al., 2023; O'Keefe & Dostrovsky, 1971; Poli et al., 2017; Zeng et al., 2019). In a similar vein, a recent study examining neuronal activity in posterior middle frontal gyrus found, likewise, that neurons there are selective for a single semantic category, such as words relating to actions (Jamali et al., 2024).

Nonetheless, we were interested in exploring the vectorial hypothesis because it can provide ready solutions to many important problems in cognitive science, such as similarity judgments, analogies, generalization, hierarchical reasoning, pattern completion, and efficient learning (Chen et al., 2017; Kommers et al., 2015; Mejri et al., 2024; Tversky, 1977; Piantadosi et al., 2024). For example, if embeddings recapitulate semantics, then similarity judgments can be as simple as calculating the neural embedding distance between two words (Devlin et al., 2018; Erk, 2012; Grand et al., 2022; Joulin et al., 2016; Pennington et al., 2014; Radford et al., 2019; Tversky, 1977; Vaswani et al., 2017). Indeed, it is the close association between embedding and semantic distance that endows vectorial codes with many of their most useful properties (Frisby et al., 2023; Mitchell & Lapata, 2008; Rogers & Mcclelland, 2004; Senel et al., 2017). One

difficulty in assessing the vectorial hypothesis is that contextual embedding models are relatively new (since 2017, Devlin et al., 2018; Peters et al., 2018; Radford et al., 2019; Vaswani et al., 2017) and neuroscience research has only recently begun relating neural and contextual word embeddings, albeit not at the single unit level (Deniz et al., 2023,Goldstein et al., 2022, 2023, 2024; Hong et al., 2024; Hosseini et al., 2024; Kumar et al., 2024; LeBel et al., 2021; Tang et al., 2023; Tuckute et al., 2024; Zada et al., 2024). Indeed, word meanings in normal language are highly contextually dependent. For example, the word "sharp" may mean something entirely different in the phrases "sharp knife" and "sharp mind". Contextual semantic embedders capture these nuances of language.

We were especially interested in the hippocampus because of its well-established role in semantic and conceptual coding (Blank et al., 2016; Davachi, 2006; Dijksterhuis et al., 2024; Fletcher et al., 1997; Quian Quiroga et al., 2005 and 2009; Rey et al., 2025; Squire & Zola-Morgan, 1991; Tulving, 2002). We recorded responses of neurons in the hippocampus of epilepsy patients undergoing intracranial monitoring for seizure activity while patients listened to continuous, narrative speech. We compared linguistic embeddings from both contextual and non-contextual embedding models to neural embedding vectors and neural responses to words and polysemy (Hartmann & James, 1998). We found that individual neurons have complex, multi-word tunings, and that lexical concepts are encoded by large and heteromorphic populations of hippocampal neurons. Moreover, we find that semantic embedding distance correlates with neural embedding distance, but only for models that involve a high degree of contextual embedding. Indeed, neural embedding variance correlated with polysemy, changing activity with different instances of the same word. Finally, we find evidence for a strong contrastive code for highly similar words. Together, these results endorse the hypothesis that semantic encoding in human hippocampus is fundamentally vectorial.

RESULTS

Neural encoding of meaning

Ten native English-speaking patients undergoing neural monitoring for epilepsy listened to 47 minutes of English speech (six monologues taken from the Moth podcast, **Methods, Figure 1A and B**). This stimulus set consisted of 7,346 words, of which 1,351 were unique. We collected responses of isolated single neurons in hippocampus (HPC, n=356 neurons) from 10 patients. Overall, neurons showed a clear response to the onset of individual words with a characteristic ramping to a peak, followed by a gradual decline in firing, although there was a great deal of diversity across neurons (**Figure 1C, D, and E**). We computed firing rate during an epoch starting 80 ms after the onset of each word (to account for the approximate response latency) and lasting the duration of the word (see **Methods)**. Rates were normalized by word duration. All analyses were repeated with a variety of response epochs which yielded similar results (data not shown).

A subset of patients (n=4) also listened to monologues containing nonsense words (Keuleers & Brysbaert, 2010, *jabberwocky stimuli*, 1,890 words, see **Methods**) following the podcast episodes. The nonsense narratives included words modified from real ones to create meaningless, novel terms, such as "toom" or "skook." These nonword stimuli preserved syntactic structure by retaining function words and morphemes but lacked coherent semantics. We found that 57% of neurons (75/125) exhibited a significant change in firing rate to non-word stimuli, with most cells showing markedly reduced activity for nonwords relative to words (**Figure 1D**). Additionally, hippocampal neural population activity strongly distinguishes words from nonwords, achieving an average decoding accuracy of 75% (chance is 50%; see **Methods**).

Next, we sought to better understand the relationship between neural activity and semantics, particularly how language unfolds over time as words are heard in sequence. To do this, we predicted each neuron's spike count for a given word from its dimensionality-reduced Word2Vec embedding (Mikolov, et al., 2013, **Figure 1F**). Word2Vec represents each word as a fixed (i.e., non-contextual) 300-dimensional vector, meaning, for example, the embedding for "camp" remains the same whether it is used as a noun or a verb. The Poisson linear model fit the responses of 354 neurons, showing a significant positive correlation between actual and predicted firing rates (**99%, Figure 1F;** see **Methods**). We found comparable results with contextual embeddings (e.g., BERT embeddings fit 355 neurons, not shown; Devlin et al., 2018).

Next, we used the same linear model to predict neural activity from up to ten words in the past and future. We found that neural activity could be better predicted from past words than future words. Neural responses to the current word could be predicted from embeddings up to six words in the past and up to three words in the future (actual model correlations were always significantly greater than shuffled, see **Methods**). However, prediction accuracy showed a characteristic decay in decodability as a function of distance forwards and backwards in time (**Figure 1G**). Thus, it appears that

hippocampal neurons keep track of recently heard words and also show some evidence of predicting upcoming words, just as other brain regions have been shown to do (Goldstein et al., 2024; Goldstein, Wang, et al., 2023; Jamali et al., 2024; Tang et al., 2023; Zada et al., 2023).

We next identified the broad semantic categories that exist in our dataset. Word2Vec embedding distances recapitulate semantic distance, such that the embedding vectors for synonyms like ‘dad’ and ‘father’ have a smaller semantic distance (cosine distance) than unrelated words like ‘dad’ and ‘emergency’ (Joulin et al., 2016; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). We leveraged this fact to build semantic categories (**Figure 1H**; Huth et al., 2016; Jamali et al., 2024). Silhouette criterion analysis (**Figure1I**; Rousseeuw, 1987) and k-means clustering of the word embeddings (see **Methods**) revealed 11 putative semantic categories in the story data, listed in increasing word frequency, with examples in brackets: (1) body parts [hand, lungs, scar], (2) places- [hospital, office, cabin], (3) emotional [cheer, disgust, feelings], (4) mental- [know, answer, imagine], (5) social [friends, queen, dad] (6) objects [book, flowers, pigeon], (7) visual [beautiful, green, sizzling], (8) numerical [another, eleven, moment], (9) actions [shimmied, writing, elected], (10) identity [she, y’all, their], and (11) function words [the, about, onto] (**Figure 1J**).

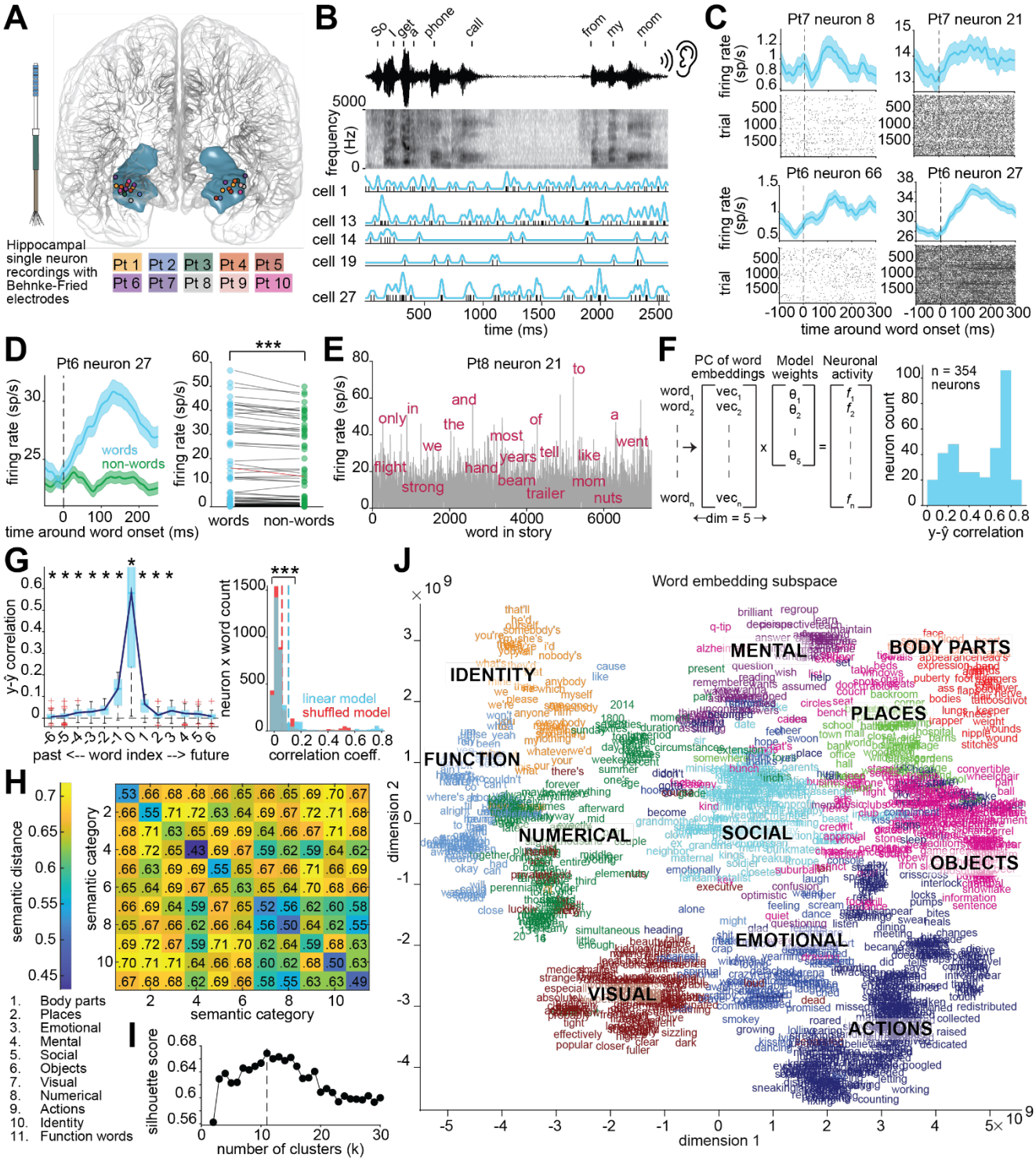


Figure 1. Hippocampal neurons track spoken words during language comprehension. | A, Electrode recording location of hippocampal neurons from ten patients. **B,** Continuous spiking activity from five example hippocampal neurons from patient 6 during words heard during the experiment, as shown in the spectrogram. **C**, Peri-stimulus time histograms and raster plots demonstrating responses to word onsets from four distinct hippocampal neurons in patient 6 (bottom row) and 7 (top row). Dashed lines

represent word onset. Plots display mean firing rate ±SEM shaded. **D,** Left - Peri-stimulus time histogram during words and jabberwocky (nonsense) non-words. Right - 71/125 neurons from four patients that exhibit significantly different firing rates between real and non-words. $P = 5.6x10^{-5}$, Wilcocxon signed-rank test. Red line represents the mean firing rate from words and non-words (14.78 and 13.98). **E**, The firing rate of one neuron occurring across the duration of all words from the six stories during language comprehension. **F**, Left - regression equation used to fit an individual neuron's spike count from word embeddings (word2vec). PC stands for principal components of the embeddings and *f* is a neuron's spike count for each word. Right – Correlation coefficients between the actual and model-predicted firing rates for significant neurons that were fitted by the model; $P < 0.05$, Spearman correlation. **G**, Left - boxplot of Spearman correlation coefficients between actual and model-predicted firing rates from the current word embeddings (index 0) and those of past and future words. Boxes reflect the interquartile range with median; navy curve is fit to the median of each box; red crosses are outliers. Index 0 box is clipped to improve visualization at other indices. Right - correlation coefficients as shown on the left across all neurons and word indexes (time). Dashed lines are the means of each distribution. Actual model performance is significantly greater than shuffled (one sided Wilcoxon signed-rank test, $P < 0.001$). **H**, Similarity matrix of median pairwise cosine distances between word embeddings within and between semantic categories that arise from all words. **I**, Silhouette score used to determine the number of semantic categories for K-means clustering of words. **J,** All words plotted in their dimensionality reduced embedding space and colored according to their semantic category.

A dense and distributed coding of semantics

One possible way to represent meaning in the brain is through neurons that selectively respond to single words, or to single categories of semantically related words, such as those associated with family, sports, or nature. However, this regime could be biologically inefficient, as it would require many neurons - enough needed for all the concepts learned across one's life. Alternatively, the same neuron could encode different concepts not simultaneously but at different moments based on context or cognitive demands. Recently, many studies have demonstrated clear and robust evidence for mixed selectivity coding principles in a range of contexts (Blanchard et al., 2018; Fusi et al., 2016; Johnston et al., 2020, 2024). Therefore, we hypothesized that word coding would be distributed and dense - meaning a single neuron should encode many, semantically unrelated words and one word should be encoded by many different neurons.

To quantify individual neuron encoding of words, we performed a Receiver Operating Characteristic (ROC) analysis (**Figure 2A-B**), a standard approach that has previously been used to characterize neural responses during behavior (Britten et al., 1992; Li et al., 2017). For statistical reasons (**Methods**), we used only words that occurred at least four times in the data (n = 283 words), and compared firing rates evoked

by each word to all others (see **Methods**). Every single one of these 283 words was encoded by at least 3 neurons (**Figure 2C-E**). The words with the least encoding (i.e., most sparsely coded) were ‘was’ and ‘and’, encoded by exactly 3 neurons; the most dense word, ‘hit”, was encoded by 67 neurons (0.8 and 18.8% of our sample, respectively, **Figure 2D-E**). On average, a single word is encoded by 37 neurons (10.4%, **Figure 2D**). Likewise, most neurons encoded many words. Only 6 out of 356 neurons did not encode any words, and the median number of words encoded by a neuron was 6 words (**Figure 2F**).

Are the words a neuron encodes semantically similar or different? Surprisingly, the same neuron does not reliably encode semantically similar words. For example, within the social category (category 5), we examined encoding of words related to the concept of family (mom, dad, father, brother, and son), and found minimal overlap between neurons that encode each family-related word (**Figure 2G**). Indeed, of the 28 neurons encoding "father" and 19 encoding "dad," only 7 encode both words (**Figure 2G**). Notably, the distinct words that a cell encoded spanned semantic categories and phonology. For example, neuron 167 encodes the words: ‘lot’, ‘remember’, ‘even’, ‘they’, ‘those, and ‘make’; neuron 252 encodes the words: ‘she’, ‘last’, ‘out’, ‘school’, and ‘team’. Each of those neurons encode words that span 5 unique semantic categories. Overall, a given neuron’s response field spanned 4 of our 11 semantic categories (**Figure 2I**). (And, recall that these proportions refer to all sampled neurons, not just a preselected subset of selective neurons).

We also wondered whether more specific concepts, like proper nouns, were encoded in this dense and distributed manner. Only three proper nouns appeared frequently enough in our data to perform statistics: ‘Mark,’ ‘God,’ and ‘Michelle’. These three are also densely encoded in the population, with ‘God’ showing the highest density at 60 neurons (**Figure 2H**). 89 neurons (25%) encoded these 3 proper nouns, with minimal neuronal overlap between them (14 neurons encoded all 3 names). Across semantic categories, fewer neurons encoded identity (pronouns) and function words, while more critical and less frequent categories, such as body parts and emotion words, were more densely encoded (**Figure 2J**). We found that such densely coded words had significantly higher classification performance (area under the curve) compared to sparsely coded words, suggesting neurons exhibit more distinct firing rates patterns for these less frequent words that are critical for understanding (**Figure 2K**).

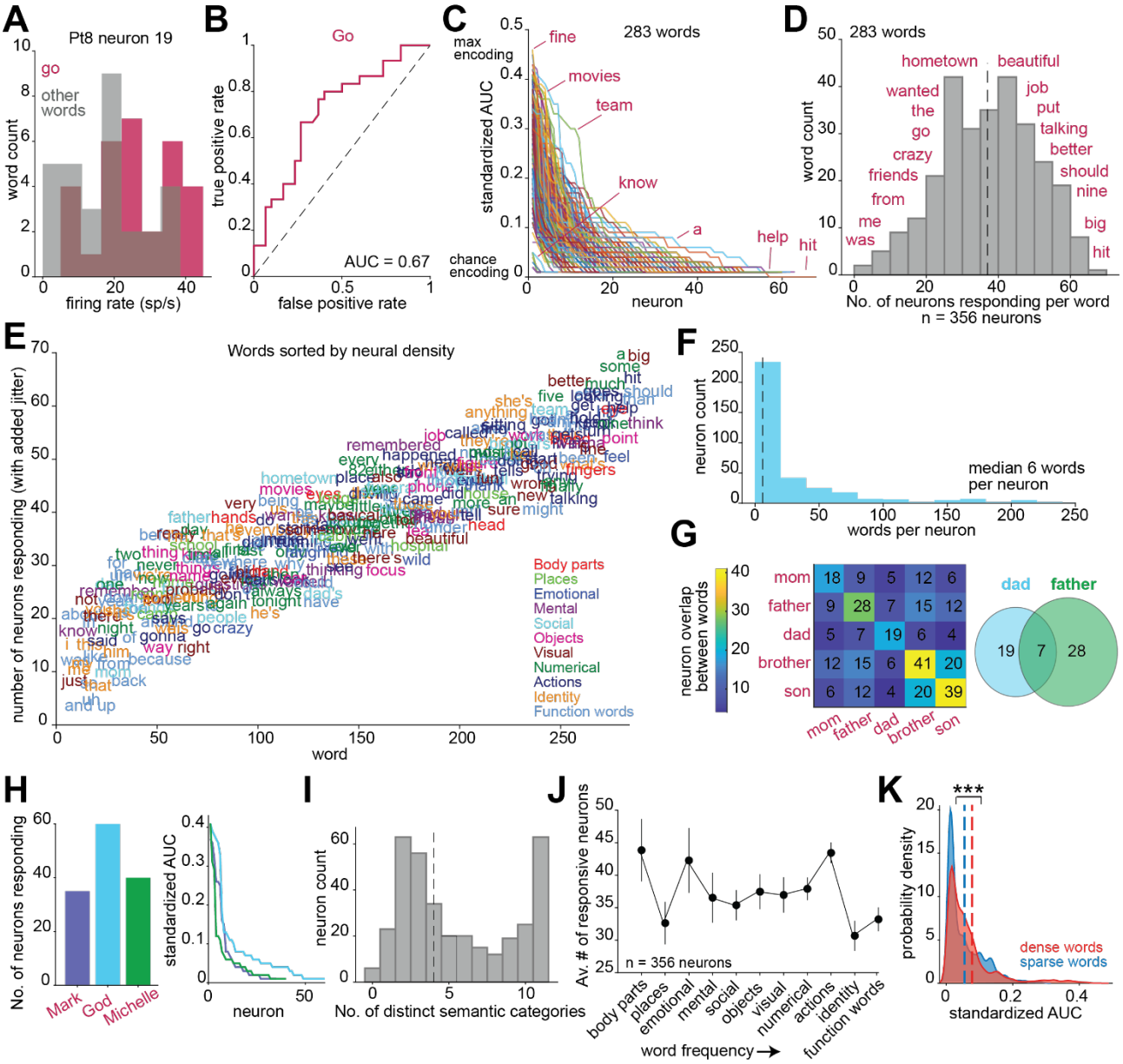


Figure 2. Hippocampal neurons represent words with a dense and distributed code. **I A**, Example firing rates from one neuron during all occurrences of the word, “go” and a random subset of other words. **B**, Receiver operator characteristic curve for the same neuronal activity and words in a. **C**, For each word (line), the standardized area under the curve (AUC) from each neuron that significantly classified that word from all others (see Methods). **D**, Histogram of the number of neurons encoding each word. Every word was encoded by at least 3 neurons. **E**, The same as in d but all 283 word text plotted based on the number of neurons that encode them, with added jitter for readability. **F,** Histogram of the number of words encoded by each neuron. **G,** Assessment of overlap between neurons encoding words related to the concept, family. Numbers in the matrix and circles represent the number of coding neurons. **H,** Left - Number of neurons encoding proper nouns from ROC analysis. Right - AUC values for proper noun encoding. **I,** Number of

semantic categories from the words that each neuron encodes. Dashed line represents the median. **J,** Average number of neurons encoding words from each semantic category. **K,** Probability density plots of AUC values from sparse words (n = 1,517 AUC values from words with < 28 cells encoding them) and dense words (n = 1,632 AUC values from words with more than 55 cells encoding them). P = 1.643×10^{-9}, Wilcoxon rank sum test.

Hippocampal semantic coding is vectorial

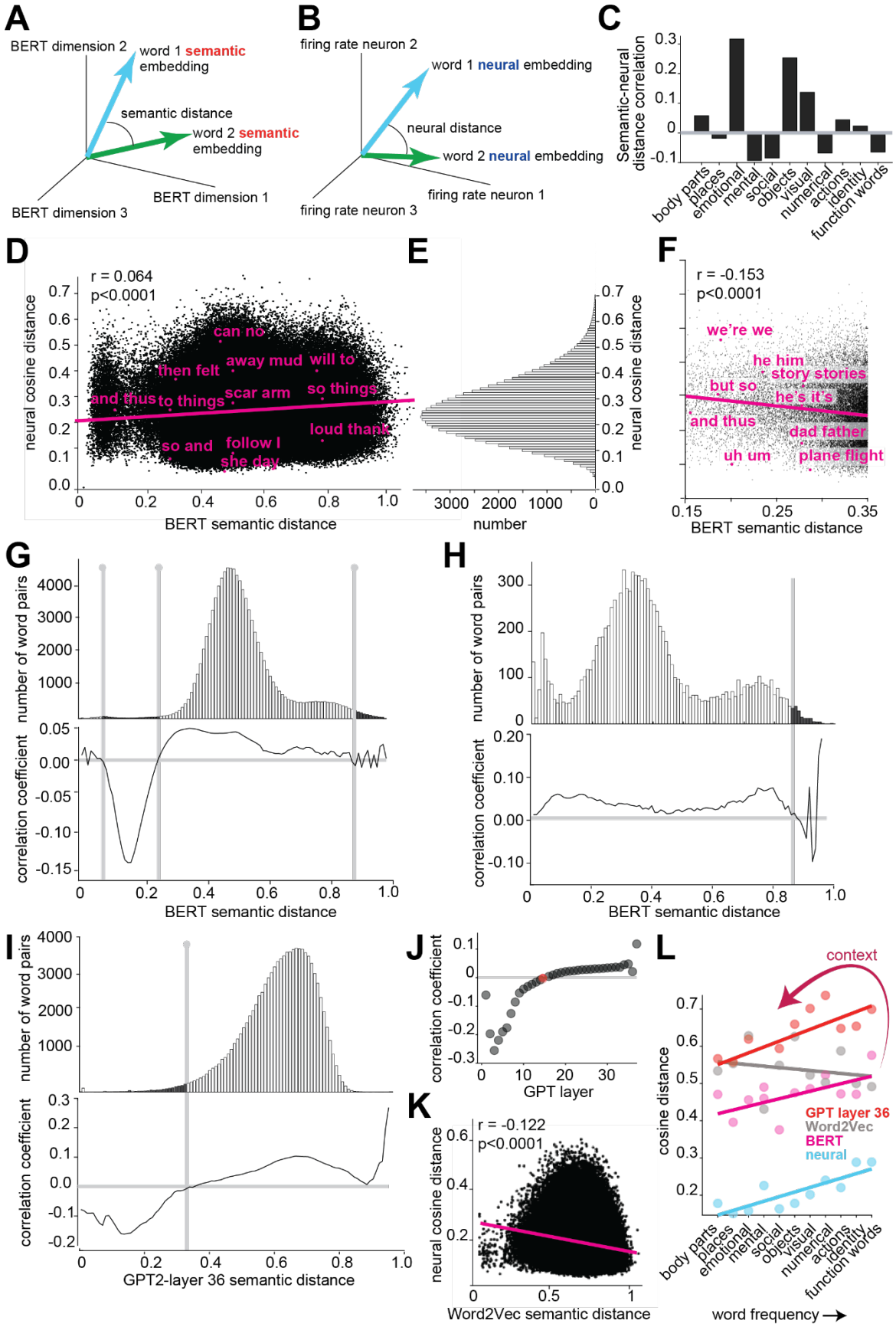
One convenient feature of semantic embeddings derived from large language models is that embedding distances correlate with semantic distances (Devlin et al., 2018; Mikolov, et al., 2013; Pennington et al., 2014; Radford et al., 2019; Vaswani et al., 2017) (**Figure 3A**). Thus, synonymous terms like "mom" and "mother" will be located close to each other in embedding space, while unrelated terms like "mom" and "tarmac" will be located far apart. Here, we will follow the standard convention of using angular (cosine) distance between embedding vectors (Mikolov et al., 2013; Pennington et al., 2014; Radford et al., 2019). We performed an analogous analysis on the responses of a population of neurons to ask whether *neural embedding distances* are correlated with semantic embedding distances (**Figure 3B**). Overall, we find that contextual embedding distance between randomly selected word pairs is correlated with their neural embedding distances (**Figure 3C-E**). Consider, for example, the contextual embedder BERT (Devlin et al., 2018)($r = 0.0650$, $p<0.0001$, **Figure 3D**). Note that while the size of this correlation is low, this is expected given the extremely high variance of single neuron firing rates and the fact that each datapoint reflects a single occurrence of a given word.

Indeed, this problem of high noise could cause confusion for words with the most similar embeddings. Such word pairs include "but and so", "we and we're", and "he and him" (although the embeddings also include local context not given here, **Figure 3F**). While the brain needs to recognize their similarity, it also needs to maintain their semantic separation as well. For example, we would not want to mix up holiday gifts for our two parents because their semantic embeddings are nearly identical. Thus, we hypothesized that, at the finest scale, neuronal embeddings would have an additional feature of contrastive coding (sometimes called repulsive coding; Bromley et al., 1994; Chanales et al., 2017; Prince et al., 2025; Radford et al., 2021; Tian et al., 2020), which would allow for separation between similar, but still distinct, words. We find precisely this: pairs of closely related words (regardless of semantic category) tend to show a *negative* correlation between semantic and neural distance (**Figure 3F-G**). In other words, the relation between semantic and neural embedding is scale-dependent: negative for the finest scale and positive for all others (**Figure 3G**). Specifically, the negative pattern is observed when considering the word pairs whose semantic distances span the range of $r = 0.15$ to $r = 0.35$ (these comprise 2.16% of the words in our sample, **Figure 3F-G**). (Note that words with a range of $r < 0.15$ comprise less than one tenth of one percent of pairs, and data in this range are too noisy to meaningfully interpret). These

highly similar pairs include many pairs of identical words; however, this negative correlation in semantically close pairs is driven entirely by word pairs consisting of different words (near-synonyms like “sort/kind” and “are/were”), rather than identical word pairs (such as “and/and” or “the/the”). Specifically, in the range of 0.15-0.35, we found $r = -0.151$ ($p < 0.0001$) for different word pairs (**Figure 3F**) and $r = +0.032$ ($p = 0.45$) for identical pairs (**Figure 3H**).

Notably, this coding regime was not unique to BERT embeddings. The same contrastive coding for semantically similar word pairs and pattern separation for semantically different words is also observed in GPT-2 layer 36 embeddings (notice the similarity between **Figure 3G and 3I**; Radford et al., 2019). Interestingly, we found a negative correlation between semantic and neural distance for early GPT-2 layers (1-13) and for Word2Vec embeddings (**Figure 3J-K**). We think this is due to the early layers of GPT-2 being less contextual than the late layers, and thus more similar to static, non-contextual embeddings from models like Word2Vec. Therefore, the neural coding of semantics is both vectorial and highly contextual.

Additionally, the distance between neural embeddings of words within a semantic category is a function of increasing word frequency (**Figure 3L, blue**; this is the case for inter-category distances as well). Meaning, low-frequency body part words like "hand" and "finger" are closer in both neural space and the semantic space of contextual embeddings, while high-frequency identity words like "everyone" and "they" show greater distance, as seen in **Figure 3L** (red, pink, and blue). In other words, like the cluster of stars near the center of the Milky Way, frequent words are more centrally positioned in the high-dimensional embedding space, a pattern that is replicated in the hippocampal coding. Indeed, applying context to semantic embeddings causes a reduction in semantic distance for low-frequency words and increases it for high-frequency words, compared to static embeddings like Word2Vec (**Figure 3L, gray**). This shift, influenced by context, aligns with the pattern of category and frequency-dependent neural distance, explaining why Word2Vec embedding distances are negatively correlated with neural data, whereas contextual models show a positive correlation, more accurately reflecting the vectorial and contextual encoding of semantics in the hippocampus.


A
BERT dimension 2
word 1 semantic embedding
semantic distance
word 2 semantic embedding
BERT dimension 1
BERT dimension 3
B
firing rate neuron 2
word 1 neural embedding
neural distance
word 2 neural embedding
firing rate neuron 1
firing rate neuron 3
C
Semantic-neural distance correlation
body parts
places
emotional
mental
social
objects
visual
numerical
actions
identity
function words
D
r = 0.064
p<0.0001
neural cosine distance
BERT semantic distance
can no
away mud
will to
then felt
scar arm
so things
and thus
to things
loud thank
so and
follow I
she day
E
number
neural cosine distance
F
r = -0.153
p<0.0001
we're we
he him
story stories
but so
he's it's
and thus
dad father
uh um
plane flight
BERT semantic distance
G
number of word pairs
correlation coefficient
BERT semantic distance
H
number of word pairs
correlation coefficient
BERT semantic distance
I
number of word pairs
correlation coefficient
GPT2-layer 36 semantic distance
J
correlation coefficient
GPT layer
K
r = -0.122
p<0.0001
neural cosine distance
Word2Vec semantic distance
L
context
cosine distance
GPT layer 36
Word2Vec
BERT
neural
word frequency

338

Figure 3. Neural distance recapitulates semantic distance. | A, We define semantic distance as the angular distance in a high dimensional space between two semantic vectors (embedding values for each word). **B,** We define neural distance in the analogous manner, where each word's neural embedding is the vector of firing rates from all neurons for that word. **C,** Semantic BERT embedding and neural embedding distance correlations among all word-pairs within each of the eleven semantic categories. While most are positive, some categories have negative correlations, likely due to contrastive coding. **D,** Scatter plot illustrating the correlation between neural and semantic embedding distances between all 26,978,185 word pairs. Pink line: regression line. Each dot represents one randomly selected word pair; magenta text indicates example pairs. **E,** Histogram of neural embedding distances; this plot is the marginal of D across the x-axis. **F,** Expanded view of the data from D in the neural range of 0.15-0.25; this range represents 2.16% of the data. In this range, the correlation between neural and semantic distances is negative. **G,** Histogram of BERT semantic embedding distances; this plot is the marginal of D across the y-axis. Below it is shown the semantic/neural correlation coefficient for the sub-range of BERT values centered on the x-axis point. While this line is generally positive in value, at especially short distances (low x-values), it turns negative. That negative correlation may reflect contrastive coding. **H,** Same data as in G, but limited to pairs in which the same word is compared in two locations. Because BERT is a contextual embedder, these will not have precisely the same value. **I.** Same data as in G, but using GPT-2 layer 36 rather than BERT. Same effects are observed, indicating that effects are not due to BERT. **J,** Correlation between semantic and neural distance across all layers of GPT-2. Early layers that are non-contextual embeddings have negative correlations with neural data, while higher, contextual layers are positive. Red indicates a layer with nonsignificant correlation. **K**, The same as in D but with Word2Vec semantic embeddings. These embeddings are static and non-contextual (similar to early layers in GPT-2), and negatively correlate with neural distance. **L,** Within-category neural distance and semantic distances from different language models. The arrow reflects how context changes semantic embeddings, improving their association to neural patterns of semantic coding. Word2vec $r = -0.22$, BERT $r = 0.62^*$, GPT-2 $r = 0.72^*$, Neural $r = 0.83^*$; Pearson correlation coefficient, $^*P < 0.05$, regression lines plotted.

Semantic coding is high dimensional and contextual

Next, we examined neural population coding and dimensionality of semantics. Using Principal Component Analysis (PCA) of each neuron's firing rate for every word (7,346 words x 356 neurons), we reduced the dimensionality of our neurons' responses to components. Only 35% of variance is explained in the first 10 components, telling us that this code is relatively high dimensional (95% of variance explained in first 190 components, **Figure 4A**; Jazayeri & Ostojic, 2021; Manley et al., 2024)**.** The fact that many dimensions are needed to explain most of the variance is consistent with the hypothesis that meaning is distributed across many neurons. Neural interactions underlying meaning may be nonlinear and cannot be fully captured by a low-dimensional

projection. Additionally, each semantic category can be decoded from the others using population activity (see **Methods**), but high frequency words like numbers, actions, and function words had reduced decoding performance compared to low frequency words (**Figure 4B**). This finding suggests, as do others above, that neural coding principles differ markedly for high frequency and low frequency words.

Indeed, for each neuron, lifetime sparseness increased as a function of word frequency across semantic categories, meaning that neurons respond to fewer common words than they do rare ones (**Figure 4C, left**). Likewise, population sparseness increased with word frequency, meaning a rare word activates more neurons than a common word does. This is consistent with, among other explanations, an energy efficient coding scheme, wherein words that are heard more often elicit less neural activity (**Figure 4C, right**).

Why do common words have reduced decoding and more sparseness compared to other, less spoken words? We think this is due, in large part, to the fact that common words are more polysemous (**Figure 4E, top;** Zipf, 1945), meaning the same word can have different meanings depending on the context in which it is spoken. To explore this idea further, we computed polysemy for each word by taking the cosine distance between BERT embedding vectors of repeated occurrences of the same word (n = 594 distinct and repeated words, see **Methods**). Polysemy is likely driving the larger neural distance between words occurring *within* high-frequency categories compared to low-frequency ones (**Figure 4D and 3L**). The words within these categories have more nuanced and flexible meanings, resulting in greater semantic (in contextual models) and neural distance within categories (**Figure 4D and 3L**).

Polysemy also explains why high frequency categories are harder to decode from all others - because these common words activate broad, overlapping neural patterns rather than category-specific ones. For example, ‘the’ may produce similar neural patterns to the noun that it modifies, such as ‘the school’, making ‘the’ and ‘school’ harder to neurally distinguish, despite being members of two different semantic categories (function words and places, respectively). Likewise, if a neuron is selective for a particular meaning of a highly polysemous word, it will only fire in contexts where that meaning is relevant, leading to a sparser response across all instances of the word. Since polysemy increases with word frequency, this can explain the similar trend we found across semantic categories in sparseness (**Figure 4C**). Given that neural distance positively correlates with semantic distance in contextual models (**Figure 3**), it follows that neural activity patterns also strongly correlate with polysemy (**Figure 4E bottom and 4F**). In other words, the meaning of a given high-polysemy word is more contextually driven than the meaning of a low-polysemy word, and hippocampal encoding tracks with that contextually driven meaning.

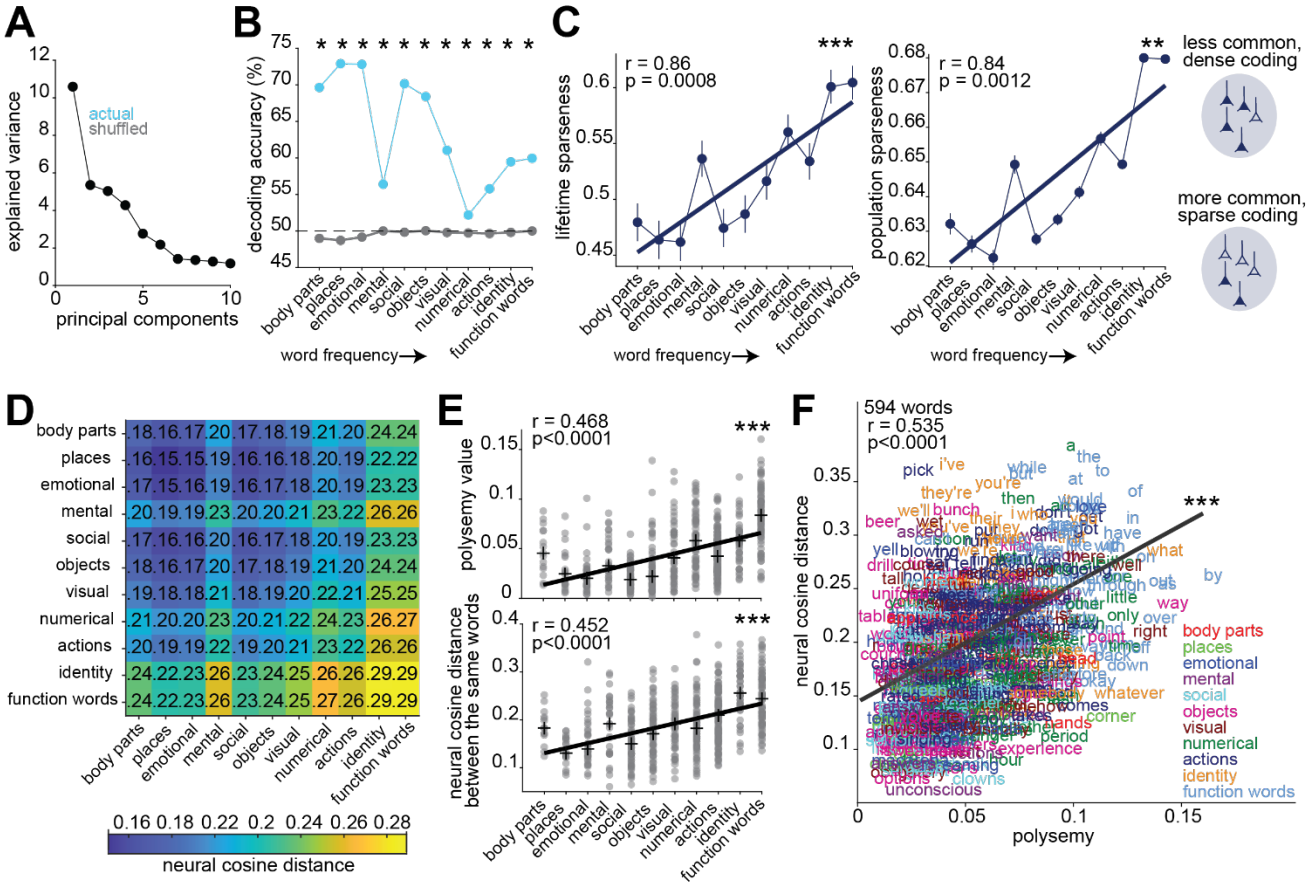


Figure 4. Population coding of words across frequency and context. | A, The amount of variance explained for each component from PCA of neural firing rates across all words. 35% of variance is explained in the first 10 components. **B**, Decoding accuracy for each semantic category and all other categories from neural population activity, for models using actual (blue) and shuffled (gray) labeled data. Chance is 50% (dashed line). Plots display mean prediction accuracy on test observations (±SEM). P < 0.0001 for each category, Wilcoxon signed-rank test comparing actual and shuffled accuracy. **C**, Left - lifetime and population sparseness measures across semantic categories, sorted by increasing word frequency. Plots display mean sparseness over neurons (lifetime) or words (population) ±SEM. Right - schematic showing population sparseness coding regime of words based on frequency. P and r-values are from Pearson correlation. **D**, Similarity matrix of neural distance between all word pairs within and across each semantic category. Diagonal of this matrix is plotted in Figure 3L. **E**, Top - polysemy values for each unique and repeated word within each semantic category. Bottom - neural distance between different occurrences of the same word, within each semantic category. P and r-values are from Pearson correlation. **F**, Each word is plotted according to its polysemy value and neural distance (median distance between all occurrences of that word). Words are colored based on their semantic category. P and r-values are from Pearson correlation. Lines in c, e, and f are regression lines.

DISCUSSION

Here, we show that the hippocampus exhibits a dense, vectorial, and contextual coding scheme for semantics. That is, individual words are encoded in the simultaneous activity of multiple neurons whose selectivities span multiple unrelated semantic categories (**Figure 2**). While a single neuron can distinguish individual words, the semantic *relationship* between words (even multiple occurrences of the same word) is reflected in vectors of the neural population activity (**Figure 3-4**). We find that *neural embedding* vectors, derived from the collective responses of neural populations, closely track word meanings as they shift across different contexts. This pattern is reflected in the association between neural activity and embeddings derived from high layers of contextual embedders (BERT and GPT-2), which incorporate meaning from adjacent words and sentences. Moreover, regardless of the embedder, at short semantic distances, this pattern is reversed. We propose that this reversal reflects a contrastive code that allows the brain to distinguish semantically very similar words (**Figure 3**).

Our finding of dense and vectorial encoding of meaning contrasts with the idea that there is a one-to-one correspondence between neurons and concepts (Bausch et al., 2021, Dijksterhuis et al., 2024; Quian Quiroga et al., 2005 and 2009; Rey et al., 2025). Of course, words are not exactly the same as concepts; our results, therefore, raise the possibility that word coding in the hippocampus may be dense while concept coding is sparse. Our attempt to isolate the most canonically concept-like words speech resulted in three cases where the single concept was repeated enough times for statistical analysis (‘God’, ‘Mark’, and ‘Michelle’); in these cases, we still find that those identities are encoded by many neurons (**Figure 2H**). These results, if found to be generally true, would suggest that concept-related words are densely coded when situated in the context of natural speech. That situation may imbue words with meanings of adjacent ones (Vaswani et al., 2017) (on the other hand, see Rey et al., 2025). Nonetheless, a direct comparison of results is difficult because of the many differences between our methods and these earlier ones.

Why are semantic embedding distances anti-correlated with neural embedding distances for word2Vec? Our data do not provide a clear answer to this question, but, in aggregate, they suggest the outlines of an answer. Overall, we find that contextual embedding distances and neural embeddings both align with word frequency and polysemy levels (which themselves are highly correlated, **Figure 3L**). Thus, the meaning of a high-polysemy word is more influenced by context than that of a low-polysemy word - this makes sense because high polysemy words, almost by definition, have meanings that are highly influenced by context. We suspect that it is this context that is the primary driver of the correlation between neural and semantic embeddings. The non-contextual embedders lack all of this context and give words a single context-free embedding. Absent the context, the underlying overall negative correlation is uncovered. This underlying negative correlation does not have a ready explanation, but may reflect a

special property of the hippocampus, which is associated with contrastive coding (Bromley et al., 1994; Chanales et al., 2017; Prince et al., 2025; Radford et al., 2021; Tian et al., 2020). One study shows a positive correlation between single neuron and semantic non-contextual embeddings in a region outside of the hippocampus (Jamali et al., 2024).

If hippocampal semantic coding is vectorial, then it shares important features with semantics in large language models (LLMs). This does not mean, however, that the brain is simply a LLM itself - indeed, we find clear differences between hippocampus and LLMs (for example, anti-correlations between neural and semantic embedding for non-contextual embeddings and for very close words). However, if the brain is not just one big LLM, LLMs more generally do offer a good comparison for thinking about neural data, and as a source of interpretational guidance for human neuroscientific data. In other words, LLMs may be more akin to a distant neuroscience model organism, like a fruit fly, that may share some key principles in common with human brains.

METHODS

Human intracranial neurophysiology

All experiments were performed in the Epilepsy Monitoring Unit (EMU) at Baylor St. Luke's Hospital using standardized approaches (Xiao et al., 2024A and B).

Experimental data were recorded from 10 adult patients (6 males and 4 females) undergoing intracranial monitoring for epilepsy. The hippocampus was not a seizure focus area of any patients included in the study. Single neuron data were recorded from stereotactic (sEEG) probes, specifically AdTech Medical probes in a Behnke-Fried configuration. Each patient had an average of 3 probes terminating in the left and right hippocampus. Electrode locations are verified by co-registered pre-operative MRI and post-operative CT scans. Each probe includes 8 microwires, each with 8 contacts, specifically designed for recording single-neuron activity. Single neuron data were recorded using a 512-channel *Blackrock Microsystems Neuroport* system sampled at 30 kHz. To identify single neuron action potentials, the raw traces were spike sorted using the *WaveClus* sorting algorithm (Chaure et al., 2018)and then manually evaluated. Noise was removed and each signal was classified as multi or single unit using several criteria: consistent spike waveforms, waveform shape (slope, amplitude, trough-to-peak), and exponentially decaying ISI histogram with no ISI shorter than the refractory period (1 ms). The analyses here used all single and multiunit activity.

Electrode visualization

Electrodes were localized using the software pipeline intracranial Electrode Visualization (iELVis; Groppe et al., 2017) and plotted across patients on an average brain using Reproducible Analysis & Visualization of iEEG (RAVE; Magnotti et al., 2020). For each patient, DICOM images of the preoperative T1 anatomical MRI and the postoperative Stealth CT scans were acquired and converted to NIfTI format (Li et al., 2016). The CT was aligned to MRI space using FSL (Jenkinson et al., 2002; Jenkinson & Smith, 2001). The resulting coregistered CT was loaded into BioImage Suite (version 3.5β1; Joshi et al., 2011) and the electrode contacts were manually localized. Electrodes coordinates were converted to patient native space using iELVis MATLAB functions (Yang et al., 2012) and plotted on the Freesurfer (version 7.4.1) reconstructed brain surface (Dale et al., 1999). Microelectrode coordinates are taken from the first (deepest) macro contact on the Ad-Tech Behnke Fried depth electrodes. RAVE (Magnotti et al., 2020) was used to transform each patient's brain and electrode coordinates into MNI152 average space. The average coordinates were plotted together on a glass brain with the hippocampus segmentation and colored by patient.

Natural language stimuli

Patients with epilepsy and healthy language function listened to six episodes (5- to 13-minutes in duration) taken from *The Moth Radio Hour,* and totaling 47:25 minutes of listening time (7,346 words). The six stories were, "Life Flight", "The Tiniest Bouquet", "The One Club", "Wild Women and Dancing Queens", "My Father's Hands" and "Juggling and Jesus". In each story, a single speaker tells an autobiographical narrative in front of a live audience. The six selected stories were chosen to be both interesting and linguistically rich. Stories were played continuously through the built-in audio speakers of the patient's hospital television. The audio signal was synchronized to the neural recording system via analog input going directly from the computer playing the audio into the Neural Signal Processor at 30 kHz.

Jabberwocky stimuli

After story listening, a subset of four patients listened to 17 minutes of Jabberwocky stories (1,890 words) to assess neural specificity for semantics (**Figure 1D**). Jabberwocky stimuli are words that are altered from the real words in the stories used in the experiment to become words that have no meaning, for example, "toom" or "skook". An example sentence of jabberwocky stimuli is: Before I smow it, the foodbace toom is at my frose woor, and a coakle of them have bapes of bour under their alk. We used the software, *Wuggy*, to transform story transcripts into jabberwocky (Keuleers & Brysbaert, 2010). This program matched the subsyllabic segment length of each word and replaced words with letters that match phonetic probabilities of orthographic English. Jabberwocky preserves syntactic structure by retaining function words and morphemes, even though it lacks coherent semantics.

Audio Transcription

After experiments, the audio .wav file was automatically transcribed using Python and Assembly AI, a state-of-the-art AI model to transcribe and understand speech. The transcribed words and corresponding timestamp output from Assembly AI was converted to a TextGrid and then loaded into Praat, a software for speech analysis. The original .wav file was also loaded into Praat and we manually checked the spectrograms and timestamps, correcting each word to ensure the word onset and offset times are accurate. The TextGrid output of corrected words and timestamps from Praat was converted to a .xls and loaded into Matlab and Python for further analysis.

Clustering of semantic vectors

To identify the natural semantic categories present in our word data (**Figure 1H-J**), all unique words were clustered into groups based on related meanings using a word embedding approach (Joulin et al., 2016; Mikolov, et al., 2013). We used the pre-trained *fastText Word2Vec* model in MATLAB to extract word embeddings for all words in our dataset (Joulin et al., 2016; Mikolov et al., 2013). This pre-trained model provides 300-dimensional word embedding vectors, trained on millions of words text, to capture semantic relationships between words. Any surname words, such as "Harwood" or proper nouns like "Applebee's" that did not have word embeddings was discarded from the analysis. To compute and visualize semantic clusters, we used a t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm on word embedding values to reduce the dimensionality of each unique word based on their cosine distance to all other words, thus reflecting their semantic similarity. Words with similar meanings now have similar 2D coordinates. We then applied the K-means clustering algorithm to these 2D word representations and visualized clustered words on a 2D word map ($k = 11$ clusters, determined from the silhouette score; Rousseeuw, 1987). We then manually inspected and assigned a distinct semantic label to each cluster and adjusted categories for accuracy. For example, words bordering the edges of categories might get mis-grouped and were manually corrected. This method is comparable to those used in other studies of semantics (Huth et al., 2016; Jamali et al., 2024).

Firing rate responses to words

We computed the duration for each word in the story and for each non-word in Jabberwocky stimuli. The firing rate of a neuron for each word was the number of spikes

occurring during the word, beginning with an 80 ms delay (and adding a further 80 ms delay afterwards) to account for the approximate delay of information to hippocampus. This value was divided by word duration and then multiplied by 1000 to compute spikes per second.

Regressing spike counts on word embeddings

We regressed the activity of each neuron onto five principal components derived from Word2Vec or BERT semantic embedding vectors (**Figure 1F-G**). We first used Principal Component Analysis (`pca`, Matlab 2023b) to reduce the dimensionality of word embeddings, selecting the first five principal components (PCs) to represent each word. We chose five PCs based on cross-validation to minimize model deviance (error), effectively balancing under- and over-fitting. These five PCs collectively explained approximately 30% of the variance in the original embedding data. Each of the five embedding-derived PCs served as predictors in a Poisson generalized linear model with a log-link function (`fitglm`, Matlab 2023b). Since spike counts for each word follow a Poisson distribution, we modeled the spike count of neuron *n* for word *w* using a Poisson regression with a log-duration offset, represented mathematically as:

$$log(\lambda_{w,n}) = log(duration_w) + \beta_0 + \beta_1 V_{w1} + \beta_2 V_{w2} + \beta_3 V_{w3} + \beta_4 V_{w4} + \beta_5 V_{w5}$$

Where $\lambda_{w,n}$, the expected spike count for word (*w)* and neuron n (*n*), is modeled by the linear regression coefficients (β), one for each of the five principal components of that word (V_w), plus the y-intercept (β_0) and the duration offset, log(duration_w). The duration offset adjusts for the varying lengths of words, ensuring that the model predicts a spike rate (spikes per unit time) rather than just an absolute spike count. We used 10-fold cross validation to train and test the model on held-out data, over 200 iterations. A neuron was fit by the model if the Spearman correlation between predicted and actual firing rates (concatenated across all folds) was significant ($p < 0.05$). The deviance measures the goodness-of-fit of the Poisson model by quantifying the difference between the observed data and the model's predictions, with lower values reflecting a better fit. Model performance was also evaluated by comparing whether the deviance of the real model was significantly smaller than the deviance of a model fit on shuffled spike counts for that neuron ($p < 0.05$, one-sided Wilcoxon signed-rank test). All neurons had significantly lower deviance in their actual model compared to the shuffled.

We used the same Poisson linear model to predict neural activity from past and future word embeddings. Here, a separate model was fit for each word index, by which we shifted the words relative to the spike counts, such that the spike count of a neuron for word X corresponded to an embedding of a different word that was in the future or past, based on the index. To determine whether neural activity was predicted by each word index, the distribution of actual model performance (y-yhat spearman correlation) from every neuron was compared to that of a shuffled model (chance distribution) - a model trained on shuffled firing rates of that neuron. For each word index, these distributions were compared using the one-sided Wilcoxon signed-rank test ($P < 0.05$). Model performance was significantly greater than chance for up to six words in the past and three words in the future.

Single neuron encoding of words

Semantic encoding of each neuron was quantified using a ROC (receiver operating characteristic) analysis, a standard approach that has previously been used to characterize neural responses during behavior (Britten et al., 1992; Li et al., 2017). Here, the firing rate of each

neuron is the classifier that is evaluated on its ability to distinguish instances of a word from all other words. For statistical reasons, since we need to compare two distributions, we performed the analysis on every neuron for every word that occurred at least 4 times (283 words, on average each word occurred 20 times, the most was 407).

Upon application of a binary threshold to firing rates and comparison with a binary event vector denoting word label (1 for the current word, 0 for other words), word detection based on neural activity can be measured using the true positive rate (TPR) and the false positive rate (FPR). Plotting the TPR against the FPR over a range of binary thresholds, spanning the minimum and maximum values of the neural signal, yields an ROC curve that describes how well the neural signal detects that word at each threshold. With the `perfcurve` function in Matlab 2023b, we used the area under the ROC curve (AUC) as a metric for how strongly neurons are modulated by each word (**Figure 2A-C**). The number of word occurrences in the 'Current word' class were balanced with the number of words in "Other" class by subsampling words from the "Other" class over 100 iterations of the analysis. For each neuron, AUC was computed, then averaged and compared to a null distribution across iterations. For each neuron, the observed AUC was compared to a null distribution of 1,000 AUC values generated from constructing ROC curves over randomly permuted neural signals (that is, firing rate permuted across words, thus destroying the relationship between neural responses and word labels). The p-value is calculated as the proportion of permuted AUC values that are greater than or equal to the observed AUC and a neuron was considered significantly responsive if $P < 0.05$. To better visualize encoding magnitude, we standardized each AUC value by subtracting chance (0.5) from every value, resulting in values between 0 and 0.5.

Contextual embeddings

In an excel file of our transcribed words, each row corresponded to an individual token, including punctuation and sentence terminators such as ".", "?", and "!". We reconstructed sentences by iterating through the file and appending words until encountering a recognized sentence delimiter. This process yielded multi-word segments that more accurately captured natural linguistic boundaries. Any remaining words following the last delimiter were grouped into a final sentence. Notably, punctuation tokens were preserved to maintain contextual fidelity but were tracked carefully to avoid introducing sub-word alignment errors during tokenization.

To extract GPT-2 embeddings, we employed the "gpt2-large" model (Radford et al., 2019) from the Hugging Face Transformers library (Wolf et al., 2020). This version of GPT-2 features a total of 37 layers (1 initial embedding layer plus 36 transformer layers), each producing 1280-dimensional hidden states, with a maximum context length of 1024 tokens. We extracted hidden states from all 37 layers using a progressive context approach designed to incorporate the autoregressive nature of GPT-2. Specifically, for each sentence, we began with an empty context and added words sequentially. After adding each new word, we tokenized the updated text string via GPT-2's Byte Pair Encoding (BPE) tokenizer, ran the token sequence through the model in a no-grad (inference-only) mode, and retrieved the hidden states. We then identified the newly appended word's sub-word tokens and averaged their corresponding vectors to obtain a single word-level embedding. Because GPT-2 often splits words into multiple sub-word fragments (otherwise known as "tokens", e.g., "computa" + "tion"), averaging those fragments preserves consistency with the transcript's original word boundaries. The GPT-2 outputs resulted in 1280-dimensional embeddings for each layer.

We conducted a parallel extraction procedure with the "bert-base-cased" model (Devlin et al., 2018), also accessed via the Hugging Face Transformers library (Wolf et al., 2020). In contrast to GPT-2, whose progressive context reflects its decoder-only design, BERT uses a masked language modeling objective and an encoder-only architecture. We therefore segmented the data into complete utterances - often spanning multiple sentences - and tokenized them up to BERT's 512-token limit. To ensure that these representations aligned with the transcript words, we again averaged sub-word embeddings whenever a single word was split by the tokenizer. However, in this instance, we extracted only the last layer of hidden states from BERT, yielding one 768-dimensional vector per token, which was then aggregated to the word level. All code was implemented in Python using PyTorch (Paszke et al., 2019) as the underlying deep learning framework, and final embeddings were saved in Excel spreadsheets for subsequent analyses.

Embedding distance

The semantic embedding vector for each word was its full vector of embedding values extracted from either Word2Vec (300 in length), BERT (768 in length), or GPT-2 (1280 in length) language models. For each word, it's neural embedding vector was every neuron's firing rate response for that word (356 in length). Neurons were combined across all 10 patients, making a pseudo population totaling 356 neurons. We used the function `pdist`, Matlab 2023b, to compute cosine distance between semantic embedding vectors of all 26,978,185 word pairs, known as semantic distance. We repeated the procedure, computing cosine distance between neural embedding vectors for all word pairs, termed neural cosine distance. "All word pairs" included pairs of the same word except in Word2Vec analysis, since cosine distance between two instances of the same word will always be 0 (since embedding values are the same). This procedure was followed for word pairs within and between each semantic category to create the semantic distance similarity matrix in **Figure 1H**, for the neural and semantic distances within each category in **Figure 3L**, and for the neural distance similarity matrix in **Figure 4D**. For **Figure 3C**, we computed the semantic distance from BERT embeddings for all word pairs within each semantic category, then computed neural distance for all word pairs within each semantic category and correlated (Pearson) these two distance vectors to compute the values shown on the y-axis. Similarly, we correlated semantic distance and neural distance from all word pairs (unless otherwise specified in the caption) to compute the correlation coefficients in **Figure 3G-J**. In **Figure 4E-F**, neural cosine distance was always computed between word pairs of multiple instances (minimum 2) of the same word. We took the median of these distances to extract one distance value per word.

$$\text{Cosine distance } (w_1, w_2) = 1 - \frac{W1 * W2}{||W1||\ ||W2||}$$

Cosine distance between neural or semantic embeddings of word 1 and word 2 is the 1 minus the dot product of embedding vectors for each word, divided by the magnitudes (norms) of each word vector.

Support Vector Machine Decoder

We used a SVM decoder (Bishop, 2006) with a nonlinear radial-based function kernel to determine whether firing rates during words carry information about semantics by distinguishing each category from all others (**Figure 4B**). We also repeated this analysis with a linear kernel, and we found equivalent results to Figure 4B except performance was much lower, only 2-5% above chance level (results not shown). For each semantic category, we used the firing rate responses from every neuron (n = 356) for each word in the category (class 1) and all rates for all other words (class 2) to predict the class for each word from these neural data. The number of words in each class were always balanced – across comparisons and across semantic categories through random subsampling of words from the larger class. Random selections of class observations were repeated for 100 iterations, giving us the average classification accuracy over 1,000 test splits of the data for each session. To train and test the model, we used a tenfold cross-validation. In brief, the data were split into ten subsets, and in each iteration the training consisted of a different 90% subset of the data; the testing was done with the remaining 10% of the data. We used the default hyperparameters as defined in `fitcsvm`, MATLAB 2023b, with hyperparameter $C = 1$ and z-scored normalization of firing rates. Decoder performance was calculated as the percentage of correctly classified test trials. We compared model performance for predicting train and test data to check for overfitting. In each iteration, we trained a separate decoder with randomly shuffled class labels. The performance of the shuffled decoder was used as a null hypothesis for the statistical test of decoder performance.

Decoding jabberwocky stimuli. We used the same methods described above on a linear SVM model (default hyperparameters) to decode whether a word was a real word from podcast stories or a non-word from the jabberwocky stimuli from neural population activity. We excluded the function words from jabberwocky stimuli in this analysis since those words are real words but used in a nonsensical context.

Sparseness

The firing rate of each neuron during the word duration (see Methods, *Firing rate responses to words)* was used to compute sparseness values. We applied the method used by Vinje and Gallant (Vinje & Gallant, 2000) which is a slight modification of the method first proposed by Treves and Rolls (Rolls & Tovee, 1995).

Lifetime sparseness. For a given neuron, for each semantic category, its lifetime sparseness was computed using:

$$s = \frac{1 - \frac{1}{N}\frac{\left(\sum_i r_i\right)^2}{\sum_i r_i^2}}{1 - \frac{1}{N}} \quad ,$$

where N is the number of words in the semantic category and r_i is the response of the neuron to the ith stimulus. The sparseness value ranges from 0 (when the neuron is least sparse and responds equally to all the words in the category) to 1 (when the neuron is maximally sparse and responds only to 1 word). Lifetime sparseness was averaged across all neurons in each category and plotted in **Figure 4C -** left.

Population sparseness. For a given word in each semantic category, its population sparseness was computed using:

$$s = \frac{1 - \frac{1}{M}\frac{\left(\sum_j r_j\right)^2}{\sum_j {r_j}^2}}{1 - \frac{1}{M}} \quad ,$$

where M is the number of neurons and r_j is the response of the jth neuron to the word. This sparseness value also ranges from 0 (when the word is least sparse and elicits equal responses from all the neurons) to 1 (when the word is maximally sparse and elicits response from only 1 neuron). For each semantic category, population sparseness was averaged across words in that category and plotted in **Figure 4D** - right. Lifetime and population sparseness results in **Figure 4D** are consistent even when balancing the number of words across semantic categories by subsampling the larger category (data not shown).

Polysemy

To compute polysemy for each word, we computed cosine distance between each BERT embedding vector of all occurrences of a given word (at least 2 occurrences, 594 words total). We took the median of this distribution to obtain one polysemy value per word (**Figure 4E-F**). The distribution of polysemy values for all words in each semantic category is plotted in Figure 4E – top. We repeated the polysemy analysis using neural embeddings instead of BERT embeddings to compute the neural cosine distance values in Figure 4E – bottom and Figure 4F. We also computed polysemy for each word as the total number of synsets (i.e., distinct senses) associated with each word in WordNet (Fellbaum, 1998). These polysemy values yielded the same results shown in Figure 4E-F (results not shown).

REFERENCES

Bausch, M., Niediek, J., Reber, T. P., Mackay, S., Boström, J., Elger, C. E., & Mormann, F. (2021). Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts. *Nature Communications*, *12*(1), 6164. https://doi.org/10.1038/s41467-021-26327-3

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York : Springer

Blanchard, T. C., Piantadosi, S. T., & Hayden, B. Y. (2018). Robust mixture modeling reveals category-free selectivity in reward region neuronal ensembles. *Journal of Neurophysiology*, *119*(4), 1305–1318. https://doi.org/10.1152/jn.00808.2017

Blank, I. A., Duff, M. C., Brown-Schmidt, S., & Fedorenko, E. (2016). Expanding the language network: Domain-specific hippocampal recruitment during high-level linguistic processing. *BioRxiv*, 091900. https://doi.org/10.1101/091900

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of Neuroscience*, *12*(12), 4745. https://doi.org/10.1523/JNEUROSCI.12-12-04745.1992

Bromley, J., Guyon, I., Lecun, Y., Sickinger, E., Shah, R., Bell, A., & Holmdel, L. (1994). *Signature Verification using a "Siamese" Time Delay Neural Network*.

Cervera, R. L., Wang, M. Z., & Hayden, B. Y. (2020). Systems neuroscience of curiosity. Current Opinion in Behavioral Sciences, 35, 48-55.

Chanales, A. J. H., Oza, A., Favila, S. E., & Kuhl, B. A. (2017). Overlap among Spatial Memories Triggers Repulsion of Hippocampal Representations. *Current Biology*, *27*(15), 2307-2317.e5. https://doi.org/https://doi.org/10.1016/j.cub.2017.06.057

Chaure, F. J., Rey, H. G., & Quian Quiroga, R. (2018). A novel and fully automatic spike-sorting implementation with variable number of features. *Journal of Neurophysiology*, *120*(4), 1859–1871. https://doi.org/10.1152/jn.00339.2018

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). *Evaluating vector-space models of analogy*. http://arxiv.org/abs/1705.04416

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/https://doi.org/10.1006/nimg.1998.0395

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, *16*(6), 693–700. https://doi.org/https://doi.org/10.1016/j.conb.2006.10.012

Deniz, F., Tseng, C., Wehbe, L., Dupré la Tour, T., & Gallant, J. L. (2023). Semantic Representations during Language Comprehension Are Affected by Context. *The Journal of Neuroscience*, *43*(17), 3144. https://doi.org/10.1523/JNEUROSCI.2459-21.2023

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

Dijksterhuis, D. E., Self, M. W., Possel, J. K., Peters, J. C., van Straaten, E. C. W., Idema, S., Baaijen, J. C., van der Salm, S. M. A., Aarnoutse, E. J., van Klink, N. C. E., van Eijsden, P., Hanslmayr, S., Chelvarajah, R., Roux, F., Kolibius, L. D., Sawlani, V., Rollings, D. T., Dehaene, S., & Roelfsema, P. R. (2024). Pronouns reactivate conceptual representations in human hippocampal neurons. *Science*, *385*(6716), 1478–1484. https://doi.org/10.1126/science.adr2813

Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W., & Hayden, B. Y. (2019). Tonic exploration governs both flexibility and lapses. PLoS computational biology, 15(11), e1007475.

Eisenreich, B. R., Hayden, B. Y., & Zimmermann, J. (2019). Macaques are risk-averse in a freely moving foraging task. Scientific reports, 9(1), 15091.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, *6*(10), 635–653. https://doi.org/https://doi.org/10.1002/lnco.362

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press; Bradford Books.

Fletcher, P. C., Frith, C. D., & Rugg, M. D. (1997). The functional neuroanatomy of episodic memory. *Trends in Neurosciences*, *20*(5), 213–218. https://doi.org/https://doi.org/10.1016/S0166-2236(96)01013-2

Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A., & Rogers, T. T. (2023). Decoding semantic representations in mind and brain. *Trends in Cognitive Sciences*, *27*(3), 258–281. https://doi.org/10.1016/j.tics.2022.12.006

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Nastase, S. A., Zada, Z., Ham, E., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Reichart, R., Friedman, D., Brenner, M., Hassidim, A., … Hasson, U. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to

common geometric patterns. *Nature Communications*, *15*(1), 2768. https://doi.org/10.1038/s41467-024-46631-y

Goldstein, A., Ham, E., Schain, M., Nastase, S., Zada, Z., Dabush, A., Aubrey, B., Gazula, H., Feder, A., Doyle, W. K., Devore, S., Dugan, P., Friedman, D., Reichart, R., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., Levy, O., & Hasson, U. (2023). *The Temporal Structure of Language Processing in the Human Brain Corresponds to The Layered Hierarchy of Deep Language Models*. https://doi.org/10.1101/2022.07.11.499562

Goldstein, A., Wang, H., Niekerken, L., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S. A., Gazula, H., Schain, M., Singh, A., Rao, A., Choe, G., Kim, C., Doyle, W., Friedman, D., Devore, S., Dugan, P., Hassidim, A., Brenner, M., … Hasson, U. (2023). *Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations*. https://doi.org/10.1101/2023.06.26.546557

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., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Groppe, D. M., Bickel, S., Dykstra, A. R., Wang, X., Mégevand, P., Mercier, M. R., Lado, F. A., Mehta, A. D., & Honey, C. J. (2017). iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *Journal of Neuroscience Methods*, *281*, 40–48. https://doi.org/https://doi.org/10.1016/j.jneumeth.2017.01.022

Hartmann, R. R. K., & James, G. (1998). *Dictionary of Lexicography*. Psychology Press.

Hayden, B. Y. (2019). Why has evolution not selected for perfect self-control?. Philosophical Transactions of the Royal Society B, 374(1766), 20180139.

Hong, Z., Wang, H., Zada, Z., Gazula, H., Turner, D., Aubrey, B., Niekerken, L., Doyle, W., Devore, S., Dugan, P., Friedman, D., Devinsky, O., Flinker, A., Hasson, U., Nastase, S. A., & Goldstein, A. (2024). Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language. *BioRxiv*, 2024.06.12.598513. https://doi.org/10.1101/2024.06.12.598513

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. Neurobiology of Language, 5(1), 43-63.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Jamali, M., Grannan, B., Cai, J., Khanna, A. R., Muñoz, W., Caprara, I., Paulk, A. C., Cash, S. S., Fedorenko, E., & Williams, Z. M. (2024). Semantic encoding during language comprehension at single-cell resolution. *Nature*, *631*(8021), 610–616. https://doi.org/10.1038/s41586-024-07643-2

Jazayeri, M., & Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, *70*, 113–120. https://doi.org/https://doi.org/10.1016/j.conb.2021.08.002

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/https://doi.org/10.1006/nimg.2002.1132

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/https://doi.org/10.1016/S1361-8415(01)00036-6

Johnston, W. J., Fine, J. M., Yoo, S. B. M., Ebitz, R. B., & Hayden, B. Y. (2024). Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. *Nature Neuroscience*, *27*(11), 2218–2230. https://doi.org/10.1038/s41593-024-01758-5

Johnston, W. J., Palmer, S. E., & Freedman, D. J. (2020). Nonlinear mixed selectivity supports reliable neural computation. *PLOS Computational Biology*, *16*(2), e1007544-. https://doi.org/10.1371/journal.pcbi.1007544

Joshi, A., Scheinost, D., Okuda, H., Belhachemi, D., Murphy, I., Staib, L. H., & Papademetris, X. (2011). Unified Framework for Development, Deployment and Robust Testing of Neuroimaging Algorithms. *Neuroinformatics*, *9*(1), 69–84. https://doi.org/10.1007/s12021-010-9092-8

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *ArXiv*, *abs/1607.01759*. https://api.semanticscholar.org/CorpusID:1210515

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Kolibius, L. D., Roux, F., Parish, G., Ter Wal, M., Van Der Plas, M., Chelvarajah, R., Sawlani, V., Rollings, D. T., Lang, J. D., Gollwitzer, S., Walther, K., Hopfengärtner, R., Kreiselmeyer, G., Hamer, H., Staresina, B. P., Wimber, M., Bowman, H., & Hanslmayr, S. (2023). Hippocampal neurons code individual episodic memories in humans. *Nature Human Behaviour*, *7*(11), 1968–1979. https://doi.org/10.1038/s41562-023-01706-6

Kommers, C., Ustun, V., Demski, A., & Rosenbloom, P. (2015). *Hierarchical Reasoning with Distributed Vector Representations*. http://cs.fit.edu/mmahoney/compression/textdata.html.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2024). *Shared functional specialization in transformer-based language models and the human brain*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

LeBel, A., Jain, S., & Huth, A. G. (2021). Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual. *Journal of Neuroscience*, *41*(50), 10341–10355. https://doi.org/10.1523/JNEUROSCI.0118-21.2021

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. Journal of neuroscience, 31(8), 2906-2915.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, *264*, 47–56. https://doi.org/https://doi.org/10.1016/j.jneumeth.2016.03.001

Li, Y., Mathis, A., Grewe, B. F., Osterhout, J. A., Ahanonu, B., Schnitzer, M. J., Murthy, V. N., & Dulac, C. (2017). Neuronal Representation of Social Information in the Medial Amygdala of Awake Behaving Mice. *Cell*, *171*(5), 1176-1190.e17. https://doi.org/https://doi.org/10.1016/j.cell.2017.10.015

Linzen, T., & Baroni, M. (2025). *Syntactic Structure from Deep Learning*. *21*, 43. https://doi.org/10.1146/annurev-linguistics-032020

Magnotti, J. F., Wang, Z., & Beauchamp, M. S. (2020). RAVE: Comprehensive open-source software for reproducible analysis and visualization of intracranial EEG data. *NeuroImage*, *223*, 117341. https://doi.org/https://doi.org/10.1016/j.neuroimage.2020.117341

Manley, J., Lu, S., Barber, K., Demas, J., Kim, H., Meyer, D., Traub, F. M., & Vaziri, A. (2024). Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. *Neuron*, *112*(10), 1694-1709.e5. https://doi.org/https://doi.org/10.1016/j.neuron.2024.02.011

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

Mejri, M., Amarnath, C., & Chatterjee, A. (2024). *LARS-VSA: A Vector Symbolic Architecture For Learning with Abstract Rules*. http://arxiv.org/abs/2405.14436

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:5959482

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. http://arxiv.org/abs/1310.4546

Mitchell, J., & Lapata, M. (2008). *Vector-based Models of Semantic Composition*.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/https://doi.org/10.1016/0006-8993(71)90358-1

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. http://arxiv.org/abs/1912.01703

Pavlick, E. (2025). *Semantic Structure in Deep Learning Keywords*. https://doi.org/10.1146/annurev-linguistics-031120

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. http://nlp.

Peters, M. E., Neumann, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. http://allennlp.org/elmo

Piantadosi, S. T., Muller, D. C. Y., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. In *Trends in Cognitive Sciences* (Vol. 28, Issue 9, pp. 844–856). Elsevier Ltd. https://doi.org/10.1016/j.tics.2024.06.011

Poli, D., Thiagarajan, S., Demarse, T. B., Wheeler, B. C., & Brewer, G. J. (2017). Sparse and specific coding during information transmission between co-cultured dentate gyrus and CA3 hippocampal networks. *Frontiers in Neural Circuits*, *11*. https://doi.org/10.3389/fncir.2017.00013

Prince, J. S., Alvarez, G. A., & Konkle, T. (2025). Contrastive learning explains the emergence and function of visual category-selective regions. *Science Advances*, *10*(39), eadl1776. https://doi.org/10.1126/sciadv.adl1776

Pulvermüller, F. (1999a). . *Words in the brain: Where? Why? How?* http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-63798

Pulvermüller, F. (1999b). Words in the brain's language. *Behavioral and Brain Sciences*, *22*(2), 253–279. https://doi.org/DOI: 10.1017/S0140525X9900182X

Quian Quiroga, R. (2023). An integrative view of human hippocampal function: Differences with other species and capacity considerations. *Hippocampus*, *33*(5), 616–634. https://doi.org/https://doi.org/10.1002/hipo.23527

Quian Quiroga, R., Kraskov, A., Koch, C., & Fried, I. (2009). Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain. *Current Biology*, *19*(15), 1308–1313. https://doi.org/10.1016/j.cub.2009.06.060

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. https://doi.org/10.1038/nature03687

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, *abs/2103.0*. https://arxiv.org/abs/2103.00020

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://github.com/codelucas/newspaper

Rey, H. G., Panagiotaropoulos, T. I., Gutierrez, L., Chaure, F. J., Nasimbera, A., Cordisco, S., Nishida, F., Valentin, A., Alarcon, G., Richardson, M. P., Kochen, S., & Quian Quiroga, R. (2025). Lack of context modulation in human single neuron responses in the medial temporal lobe. *Cell Reports*, *44*(1). https://doi.org/10.1016/j.celrep.2024.115218

Rogers, T. T., & Mcclelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*.

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, *73*(2), 713–726. https://doi.org/10.1152/jn.1995.73.2.713

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., & Cukur, T. (2017). *Semantic Structure and Interpretability of Word Embeddings*. https://doi.org/10.1109/TASLP.2018.2837384

Squire, L. R., & Zola-Morgan, S. (1991). The Medial Temporal Lobe Memory System. *Science*, *253*(5026), 1380–1386. https://doi.org/10.1126/science.1896849

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*(5), 858–866. https://doi.org/10.1038/s41593-023-01304-9

Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive Multiview Coding. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 776–794). Springer International Publishing.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., ... & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. Nature Human Behaviour, 8(3), 544-561.

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, *53*(Volume 53, 2002), 1–25. https://doi.org/https://doi.org/10.1146/annurev.psych.53.100901.135114

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science (New York, N.Y.)*, *287*(5456), 1273–1276. https://doi.org/10.1126/science.287.5456.1273

Wang, M. Z., & Hayden, B. Y. (2021). Latent learning, cognitive maps, and curiosity. Current Opinion in Behavioral Sciences, 38, 1-7.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2020). *Transformers: State-of-the-Art Natural Language Processing*. https://github.com/huggingface/

Xiao, J., Adkinson, J. A., Allawala, A. B., Banks, G., Bartoli, E., Fan, X., Mocchi, M., Pascuzzi, B., Pulapaka, S., Franch, M. C., Mathew, S. J., Mathura, R. K., Myers, J., Pirtle, V., Provenza, N. R., Shofty, B., Watrous, A. J., Pitkow, X., Goodman, W. K., … Hayden, B. Y. (2024). *Insula uses overlapping codes for emotion in self and others*. https://doi.org/10.1101/2024.06.04.596966

Xiao, J., Adkinson, J. A., Myers, J., Allawala, A. B., Mathura, R. K., Pirtle, V., Najera, R., Provenza, N. R., Bartoli, E., Watrous, A. J., Oswalt, D., Gadot, R., Anand, A., Shofty, B., Mathew, S. J., Goodman, W. K., Pouratian, N., Pitkow, X., Bijanki, K. R., … Sheth, S. A. (2024). Beta activity in human anterior cingulate cortex mediates reward biases. *Nature Communications*, *15*(1), 5528. https://doi.org/10.1038/s41467-024-49600-7

Yacoub, E., Grier, M. D., Auerbach, E. J., Lagore, R. L., Harel, N., Adriany, G., ... & Zimmermann, J. (2020). Ultra-high field (10.5 T) resting state fMRI in the macaque. Neuroimage, 223, 117349.

Yang, A. I., Wang, X., Doyle, W. K., Halgren, E., Carlson, C., Belcher, T. L., Cash, S. S., Devinsky, O., & Thesen, T. (2012). Localization of dense intracranial electrode arrays using magnetic resonance imaging. *NeuroImage*, *63*(1), 157–165. https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.06.039

Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Nastase, S. A., & Hasson, U. (2023). A shared linguistic space for transmitting our thoughts from brain to brain in natural conversations. *BioRxiv*, 2023.06.27.546708. https://doi.org/10.1101/2023.06.27.546708

Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Nastase, S. A., & Hasson, U. (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*, *112*(18), 3211-3222.e5. https://doi.org/10.1016/j.neuron.2024.06.025

Zeng, T., Li, X., & Si, B. (2019). *Learning Sparse Spatial Codes for Cognitive Mapping Inspired by Entorhinal-Hippocampal Neurocircuit*. http://arxiv.org/abs/1910.04590

Zipf, G. K. (1945). The Meaning-Frequency Relationship of Words. *The Journal of General Psychology*, *33*(2), 251–256. https://doi.org/10.1080/00221309.1945.10544509