nature neuroscience

Article

https://doi.org/10.1038/s41593-026-02436-4

A population code for semantics in human hippocampus

Received: 26 August 2025

Accepted: 31 July 2026

Published online: 30 September 2026

Check for updates

Melissa Franch[1] ✉, Elizabeth A. Mickiewicz[1], James L. Belanger[1], Brad Joiner[1], Kalman A. Katlowitz[1], Hanlin Zhu[1], Ana G. Chavez[1], Assia Chericoni[1], Alica M. Goldman[2], Vaishnav Krishnan[2], Atul Maheshwari[2], Danika L. Paulo[1], Eleonora Bartoli[1,3], Suzanne Kemmer[4], Steven T. Piantadosi[5], Nicole R. Provenza[1,3,6], Jay A. Hennig[3,6,7], Sameer A. Sheth[1,3,6,7,8] & Benjamin Y. Hayden[1,3,4,6,7]

Motivated by the successes of large language models (LLMs), we hypothesized that semantic information in the brain may be expressed in distributed patterns of neural activity. We recorded responses of hippocampal neurons while individuals listened to narrative speech. Here, using an encoding model approach, we find robust encoding of semantics, even after controlling for phonemics and grammar. Individual words had complex selectivities that included multiple words in multiple semantic categories. Similar to embedding vectors in LLMs, distance between neural population responses correlated with semantic distance. For 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, neural population activity aligned most closely with GPT-2 embeddings, and variation in response patterns correlated with LLM-derived polysemy measures, indicating that semantic encoding is largely contextualized. Ultimately, these results provide a neurocomputational account for understanding how hippocampal neurons encode word meaning during language listening.

The principles by which neurons encode meaning remain poorly understood. One hypothesis is that the brain uses a population code, so that word meanings are represented in patterns of activation across a large number of units[1–4]. Population coding facilitates solutions to several important problems including similarity judgments, analogies, generalization, hierarchical reasoning, pattern completion and efficient learning[5]. For example, if population codes are structured so that words with more similar meanings evoke more similar neural responses, then similarity is the distance between words' evoked responses.

In areas outside of language, population coding has attracted growing interest[6,7]. We hypothesized that population patterns of neural responses collectively instantiate a representation of word meaning that is reliable across repeated contexts but varies as the context of the word changes its specific meaning. Moreover, we hypothesized that many neurons participate in encoding any given concept, and individual neurons encode a range of distinct concepts. One difficulty in assessing the population coding hypothesis with linguistic stimuli is that language encoding is contextual[8–11]. For example, the word 'sharp' means something different in 'sharp knife' and 'sharp mind'. Contextual

[1]Department of Neurosurgery, Baylor College of Medicine, Houston, TX, USA. [2]Department of Neurology, Baylor College of Medicine, Houston, TX, USA. [3]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. [4]Department of Linguistics, Rice University, Houston, TX, USA. [5]Department of Psychology, University of California, Berkeley, Berkeley, CA, USA. [6]Neuroengineering Initiative, Rice University, Houston, TX, USA. [7]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA. [8]Gordon and Mary Cain Pediatric Neurology Research Foundation Laboratories, Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA. ✉e-mail: melissa.franch@bcm.edu

embedders capture these nuances of contextual language[3,12–16]. We hypothesized that incorporating context would improve fit quality to neural representations of semantics.

We are especially interested in the hippocampus because of its well-established role in semantic and conceptual coding in humans[17–22]. Studies find language-specific responses in the hippocampus[19,23] and implicate the medial temporal lobe in semantic processing[24,25]. In addition, hippocampal neurons track conceptual relationships during language comprehension[20,26,27]. This form of conceptual coding may support the retrieval of semantic associations and facilitate recall[17,28]. Finally, hippocampal populations track narrative structure[29,30]. Together, these findings establish the hippocampus as a central node for representing semantic content and conceptual knowledge.

We recorded responses of hippocampal neurons in epilepsy patients listening to narrative speech. We found that individual neurons have complex, multiword tunings and that lexical concepts are encoded by large and heteromorphic populations of hippocampal neurons. Moreover, semantic embedding distance correlates with neural embedding distance. Neural embedding variance correlates with polysemy, changing activity with different instances of the same word. Finally, we find evidence for a contrastive code for highly similar words. Together, these results endorse the hypothesis that the hippocampus uses population coding principles for semantics during natural language comprehension.

Results

Neural encoding of meaning

Ten native English-speaking patients (six males and four females) undergoing neural monitoring for epilepsy listened to 47 min of English speech (six monologues taken from the Moth podcast) (Methods; Fig. 1a,b). This stimulus set consisted of 7,346 words (1,351 unique). We collected responses of isolated single neurons in hippocampus ($n = 356$ neurons, Extended Data Fig. 1) from ten patients. Neurons showed a clear response to the onset of individual words, often with a ramping to a peak, followed by a gradual decline in firing, although there was a great deal of diversity across neurons (Fig. 1c–e and Extended Data Fig. 2). Across words of different durations (short ~100 ms, middle ~200 ms, long ~500 ms), neurons show a similar rise in firing rate shortly after word onset (Extended Data Fig. 3). We defined firing rate as the average number of spikes in 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. Rates were normalized by word duration.

A subset of patients ($n = 4$) also listened to monologues containing nonsense words ('jabberwocky stimuli', 1,890 words; Methods) following the podcast episodes[31]. The nonsense narratives included words modified from real ones to create meaningless, novel terms, such as 'toom' or 'skook'. We found that 57% of neurons (75/125) exhibited a significant change in firing rate to nonword stimuli, with most cells showing reduced activity for nonwords relative to words (Fig. 1d and Extended Data Fig. 2b). Words and nonwords were linearly decodable from hippocampal neural population activity using a support vector machine classifier (decoding accuracy 75%, chance 50%; Extended Data Fig. 2c).

We next predicted each neuron's spike count for each word from dimensionality-reduced Word2Vec embeddings[32], word duration, and their interactions (Fig. 1f). Word2Vec represents each word as a fixed (that is, noncontextual) 300-dimensional vector. The Poisson linear model (cross-validated, ridge regression) fit the responses of 62% ($n = 222/356$) of neurons, showing a significant and positive log-likelihood difference between actual and shuffled models and positive adjusted R^2 values (Fig. 1f,g and Extended Data Fig. 4c; 62% of model-fit neurons exceeded the 5% change expectation, one-sided binomial test, $P < 2.2 \times 10^{-308}$). Semantic encoding models explained a median 26% of the neural variance. Adding semantic predictors to a duration-only model significantly improved fit, raising adjusted R^2 by a median of 0.01 (Wilcoxon $z = 11.35$, $P < 10^{-9}$) and improving model performance by an average of 23% across neurons (Extended Data Fig. 4d). Likewise, semantic information significantly increased the Pearson correlation between actual and predicted spiking activity by 0.02 (semantic model median $r = 0.68$; Wilcoxon signed-rank $z = 5.99$, $P < 10^{-9}$). Together, these effects indicate that semantics captures variance in neural responses. We found comparable results with contextual embeddings (BERT embeddings fit 62% of neurons; GPT-2 layer 37 fit 54% of neurons; the observed fraction of model-fit neurons was significantly greater than chance, one-sided binomial test, $P < 2.2 \times 10^{-308}$). Across patients, the model fit an average of 60% of recorded neurons, with all patients contributing fitted neurons (mean of 10% recorded neurons per patient). Among neurons successfully fit by the model, significant predictors included word duration, embedding dimensions and their interactions. Notably, the first eight embedding dimensions were commonly selected across neurons, with the number of significant predictors gradually decreasing across higher dimensions (Extended Data Fig. 4a,b). Removing function words and pronouns produced a modest but significant drop in model fit, yet their exclusion did not alter any main findings (Extended Data Fig. 5a,b).

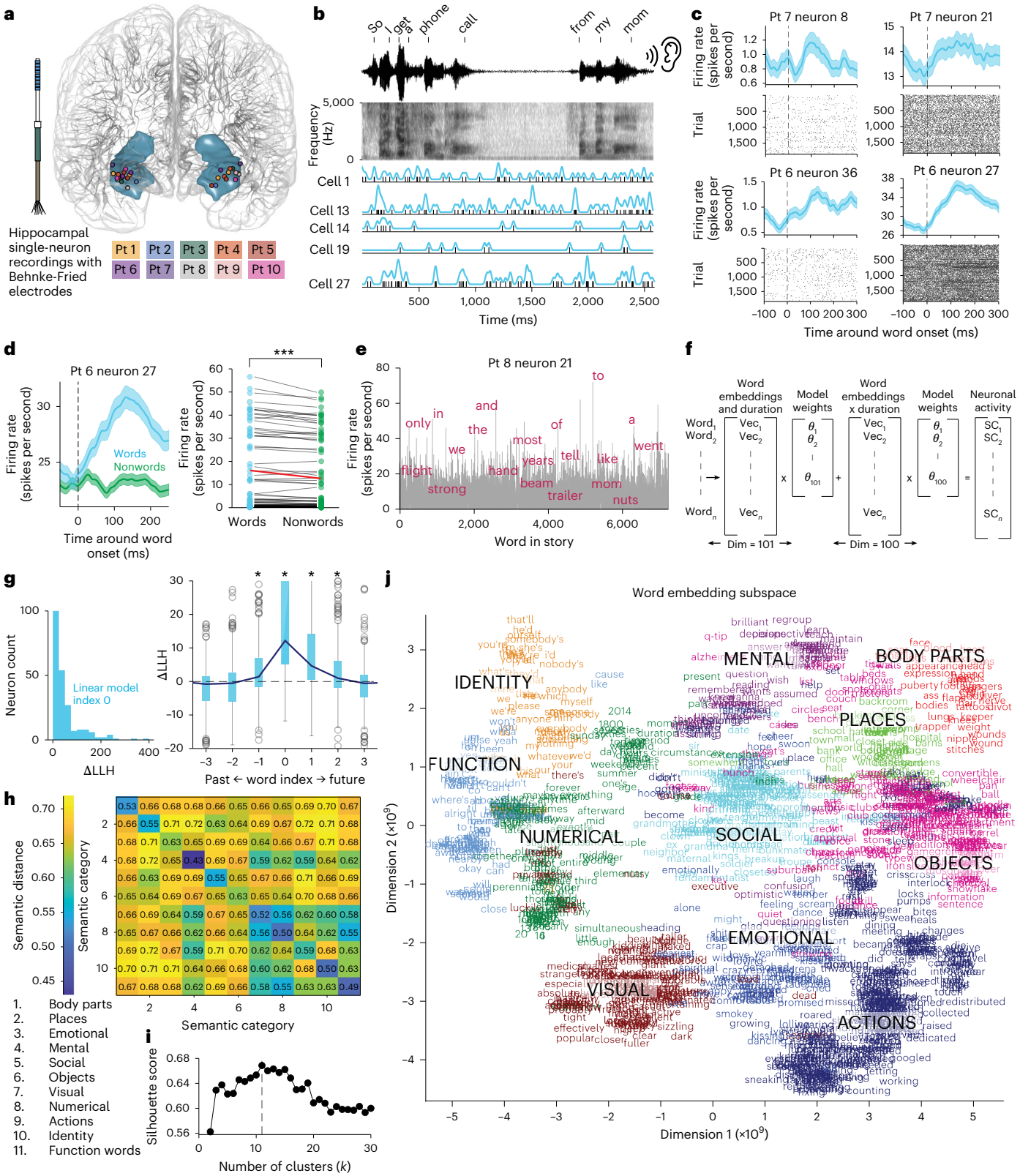
To investigate whether neurons encode the semantics or phonetics of a word, we fit neural tuning models with the phonemes of each word as predictors instead of their embedding values (Methods). Phonetic models fit fewer neurons than semantic, fitting 42% ($n = 150/356$) of neurons. In additon, for neurons fit by both models (40%, $n = 142/356$), the semantic model had significantly higher performance than the phonetic model (median phonetic−semantic $\Delta AIC = 136$, $P < 0.001$ and Extended Data Fig. 4e,f). We find that, across neurons, adding semantic

Fig. 1 | Hippocampal neurons track spoken words during language comprehension. a, Electrode recording location of hippocampal neurons from ten patients (represented as Pt). **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 patients 6 (bottom row) and 7 (top row). Dashed lines represent word onset. Plots display mean firing rate ± s.e.m. shaded. **d**, Left: peri-stimulus time histogram during words and jabberwocky (nonsense) nonwords. Plots display mean firing rate ± s.e.m. Right: 71 of the 125 neurons from 4 patients that exhibit significantly different firing rates between real and nonwords (two-sided Wilcoxon rank-sum test with FDR correction, $P < 0.05$). $P = 5.6 \times 10^{-5}$, two-sided Wilcoxon signed-rank test between neurons' means. Red line represents the mean firing rate from words and nonwords (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**, Regression equation used to fit an individual neuron's spike count from word embeddings, word duration and their pairwise interactions. Dim reflects the number of predictors including the first 100 PCs of the word embeddings and word duration. SC is a neuron's spike count for each word. θ represents regression coefficients. **g**, Left: histogram of log-likelihood improvements (LLH diff; actual−shuffled models) for fitted neurons ($n = 222$) from Poisson ridge regression predicting neuronal spike counts using the equation in **f** for the current word. Right: log-likelihood improvement of all neurons from the current word duration and embeddings (index 0) and embeddings of past and future words. In each boxplot, the center line marks the median, box edges show the 25th and 75th percentiles, whiskers extend to the most extreme nonoutliers and 'o' symbols mark outliers. The navy curve is fit to the median of each box. The y axis is truncated for visual clarity. Asterisks reflect significant P values from a one-sided Wilcoxon signed-rank test for LLH diff greater than 0 (lag −1, $P = 2.17 \times 10^{-10}$; lag 0, $P = 3.61 \times 10^{-54}$; lag 1, $P = 1.49 \times 10^{-36}$; lag 2, $P = 5.85 \times 10^{-7}$). **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.

https://doi.org/10.1038/s41593-026-02436-4

predictors significantly improves model fit relative to a phonetic-only model (median $\Delta R^2 = 0.003$, sign-rank $z = 12.4$, $P < 0.001$). By contrast, adding phonetic predictors to a semantic model did not improve performance and slightly but significantly reduced explained variance (median $\Delta R^2 = -0.001$, $z = -2.7$, $P = 0.006$; Extended Data Fig. 4f). Together, these results indicate that single-neuron activity reflects the meaning rather than phonetics.

To understand the relationship between hippocampal activity and semantic, phonetic and syntactic features, we applied the same ridge regression encoding models with a total of 61 linguistic features. Specifically, models included (1) full phonetic embeddings (43 International Phonetic Alphabet (IPA) phones; count coded)[33], (2) ten principal components (PCs) of semantic embeddings (Word2Vec, BERT or GPT-2), (3) word order in sentence, (4) word order in clause,



(5) opening node number, (6) closing node number, (7) word frequency (SUBTLEX-US), (8) pitch, (9) a scalar representing degree of polysemy[34–36] and (10) word duration. We found that one or more semantic embeddings were significant predictors for a mean 64% of the neurons fit by the model, followed by phonetic embeddings as the next most common significant predictor across neurons, fitting 41% of neurons on average (Extended Data Fig. 4g). Semantic embeddings also had the highest weights, followed by phonetic embeddings and opening and closing nodes (Extended Data Fig. 4h). Linguistic features such as pitch, word position in clause and word frequency were significant predictors for only 18% of neurons on average and had lower weights than semantic and phonemic predictors. In summary, even with a variety of linguistic features accounted for in the model, semantic predictors have the strongest and most widespread influence on neuronal activity, indicating that meaning remains the dominant driver of neural responses in the hippocampus.

Next, we used the same linear modeling to predict neural activity from up to five words in the past and future. We found that neural responses to the current word could be predicted from word embeddings up to one word in the past and up to two words in the future (LLH differences were always significantly greater than 0; Methods). However, prediction accuracy showed a characteristic decay in decodability as a function of distance forwards and backwards in time (Fig. 1g, right). Thus, it appears that hippocampal neurons keep track of recently heard words and also show some evidence of encoding upcoming words, just as other brain regions have been shown to do[9,10,37,38]. These findings suggest that hippocampal neuron responses integrate information across multiple words and could contribute to representing contextual semantics during natural speech. Note that these effects could also arise from overlapping neural responses across adjacent words or from correlations between nearby words in natural speech. In addition, whether this future word encoding represents a true prediction is a debated issue; our results do not resolve that debate[39].

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 such as 'dad' and 'father' have a smaller semantic distance (cosine distance) than unrelated words like 'dad' and 'emergency'[32,40]. We leveraged this fact to build semantic categories[37,41] (Fig. 1h). Silhouette criterion analysis[42] (Fig. 1i) and *k*-means clustering of the word embeddings (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] (Fig. 1j). Note that the categorization here is based on a static Word2Vec embedding rather than a contextual embedding, such as that provided by BERT or GPT-2. In this case, clustering of single words based on contextual embeddings wouldn't produce meaningful categories, because word meanings are determined by their local contexts. We show these categories in Extended Data Fig. 6a,b. We will cover contextual embeddings below.

A 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[37]. However, this categorical coding regime could be biologically inefficient, as it would require many neurons—enough needed for all the concepts learned across one's life. We hypothesized instead that word coding would be highly distributed, with neurons exhibiting mixed selectivity—meaning one word should be encoded by many different neurons and a single neuron should encode many, semantically unrelated words[43–45].

To quantify individual neuron encoding of words, we performed a receiver operating characteristic (ROC) analysis (Fig. 2a,b), a standard approach that has previously been used to characterize neural responses during behavior[46,47]. 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 (Methods). Every single one of these 283 words was encoded by at least three neurons (Fig. 2c–e). The words with the least encoding (that is, most sparsely coded) were 'was' and 'and', encoded by exactly three neurons; the most dense word, 'hit', was encoded by 67 neurons (0.8 and 18.8% of our sample, respectively; Fig. 2d,e). On average, a single word is encoded by 37 neurons (10.4%, Fig. 2d), suggesting a distributed code for semantics. 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 six words (Fig. 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 (Fig. 2g). Indeed, of the 28 neurons encoding 'father' and 19 encoding 'dad', only 7 encode both words (Fig. 2g). This low overlap reflects the contextual nature of hippocampal semantic coding during natural language comprehension—neurons encode the meaning of a word as it is used in a specific narrative context, rather than responding to membership in a broad semantic category. 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 five unique semantic categories. Overall, a given neuron's response field spanned 4 of our 11 semantic categories (recall that these proportions refer to all sampled neurons, not just a preselected subset of selective neurons) (Fig. 2i).

Finally, for each neuron that significantly encoded at least two words ($n = 334$), we quantified its semantic range by computing the cosine distance between embeddings of all pairs of words it encoded. We compared this distribution to a null one that represents the full semantic space of the word data (Methods). On average across embedding models, 88% of eligible neurons encoded words that spanned at least 70% of the null distribution (mean 295/334 neurons across models; GPT-2: 327/334, 98%; BERT: 333/334, 99%; Word2Vec: 226/334, 68%). Neurons covered significantly more semantic range than expected by chance (one-sided Wilcoxon signed-rank test, GPT-2 median of 0.99, $P = 8.49 \times 10^{-57}$; BERT median of 0.99, $P = 8.80 \times 10^{-57}$; Word2Vec median of 0.87, $P = 7.63 \times 10^{-4}$). These results demonstrate that hippocampal neurons generally encoded semantically diverse sets of words rather than narrowly restricted semantic clusters, supporting broad, mixed semantic tuning.

We also wondered whether more specific concepts, such as proper nouns, were encoded in this distributed manner. Only three proper nouns appeared frequently enough in our data to perform statistics: 'Mark', 'God' and 'Michelle'. These three are also encoded across neurons in the population, with 'God' showing the highest density at 60 neurons (Fig. 2h). A total of 89 neurons (25%) encoded these three proper nouns, with minimal neuronal overlap between them (14 neurons encoded all three 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 distributed—encoded by many neurons (Fig. 2j). We found that such densely coded words had significantly higher classification performance (area under the curve, AUC) compared to sparsely coded

https://doi.org/10.1038/s41593-026-02436-4

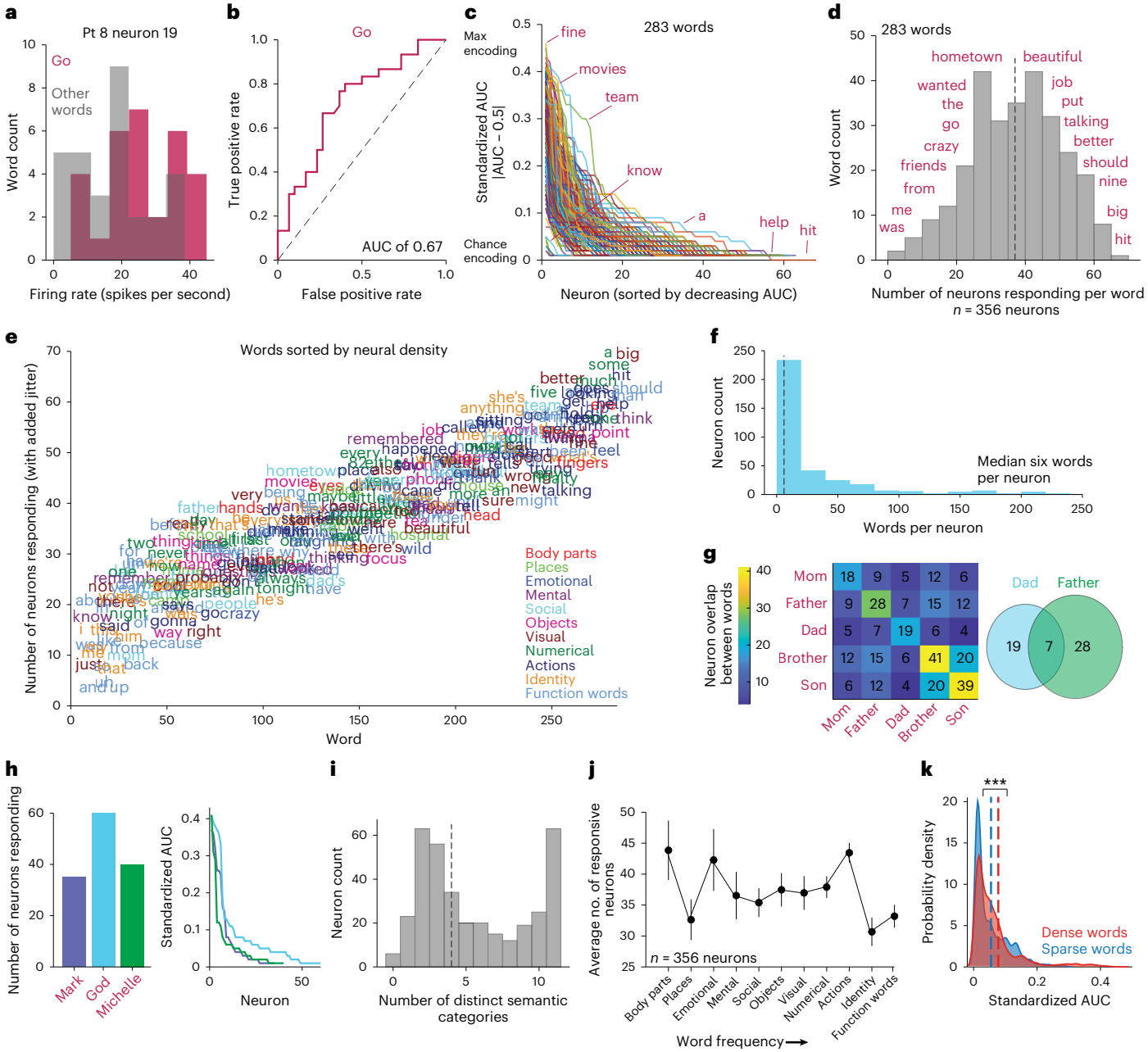


Fig. 2 | Hippocampal neurons represent words with a distributed code. **a**, Example firing rates from one neuron during all occurrences of the word 'go' and a random subset of other words. **b**, ROC curve for the same neuronal activity and words in **a**. **c**, For each word (line), the standardized AUC from each neuron that classified that word from all others (Methods). For each word, neurons encoding that word are sorted on the *x* axis from greatest to lowest encoding performance. **d**, Histogram of the number of neurons encoding each word. Every word was encoded by at least three neurons. **e**, The same as in **d**, but all 283 word texts 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. Plots display mean values ± s.e.m. **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). $P = 1.64 \times 10^{-9}$, two-sided Wilcoxon rank-sum test.

words, suggesting neurons exhibit more distinct firing rates for these less frequent words that are critical for understanding (Fig. 2k).

Hippocampal population activity reflects semantic relationships

In semantic embedders such as Word2Vec, words are represented as points in a high dimensional space, and distance in this space corresponds to semantic distance (Fig. 3a). Thus, synonyms such as 'mom' and 'mother' are located close to each other in embedding space, whereas unrelated terms such as 'mom' and 'tarmac' are located far apart. Here, we follow the standard approach of using angular (cosine) distance between embedding vectors to quantify vectorial distance[15,32,48]. 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 (Fig. 3b). Overall, we find that contextual embedding distance between randomly

selected word pairs is correlated with their neural embedding distances (Fig. 3c–e). Consider, for example, the contextual embedder BERT[49] (using unidirectional embeddings[50–52]; Pearson $r = 0.0650$, $P < 0.0001$; Fig. 3d) (note that each datapoint represents the cosine distance (on the *y* axis) of neural responses between two single words, with no averaging across repetitions). Using GPT-2, we also found a significant positive correlation between neural distance and semantic distance ($r = 0.0921$, $P < 0.0001$; Fig. 3j and Extended Data Fig. 7a).

Note that while the size of this correlation is low, this is inevitable given the extremely high variance of single-neuron firing rates and the fact that each datapoint reflects a single occurrence of a given word pair. Indeed, we quantified the maximum correlation achievable between the actual semantic embeddings and their altered version that mirrors neural data (Methods) and found that it was 0.122 and 0.119 for BERT and GPT-2 respectively. Therefore, the neural-BERT correlation of 0.06 captures 54% of the noise-limited ceiling, showing that over half of the model's representational structure is recoverable despite the high variability shown by single neurons for single repeats of a word (Extended Data Fig. 7d). The neural-GPT-2 correlation of 0.09 is, relatively, even stronger, reaching 76% of the ceiling (Extended Data Fig. 7b–d). Moreover, neural–semantic correlations exceed the entire shuffle-based null distribution where neural responses were shuffled across words, indicating that the effect reflects true neural–semantic structure rather than an artifact of having many datapoints (Extended Data Fig. 7d, gray histograms).

This problem of high noise in neurons' firing rates 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, Fig. 3f). While the brain needs to recognize their similarity, it also needs to maintain their semantic separation as well. Thus, we hypothesized that, at the finest scale, neuronal embeddings would have an additional feature of contrastive coding[53–57] (sometimes called repulsive coding), 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 (Fig. 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 (Fig. 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, Fig. 3f,g). 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 such as '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 (Fig. 3f) and $r = +0.032$ ($P = 0.45$) for identical pairs (Fig. 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 37 embeddings[15] (notice the similarity between Fig. 3g,i and Extended Data Fig. 7a–d). Interestingly, we found a negative correlation between semantic and neural distance for early GPT-2 layers (1–13) and for Word2Vec embeddings (Fig. 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, noncontextual embeddings from models such as Word2Vec. Therefore, the neural coding of semantics is both vectorial and highly contextual. Note that all results in Fig. 3 were replicated when performed with populations of neurons from each patient individually (Extended Data Fig. 7e–g). In addition, removing function words and pronouns slightly increased neural–semantic-distance correlations, suggesting that these words add context-driven variability while content words yield more consistent semantic responses (Extended Data Fig. 5c). Because this does not change our results qualitatively, we retain them in the analysis.

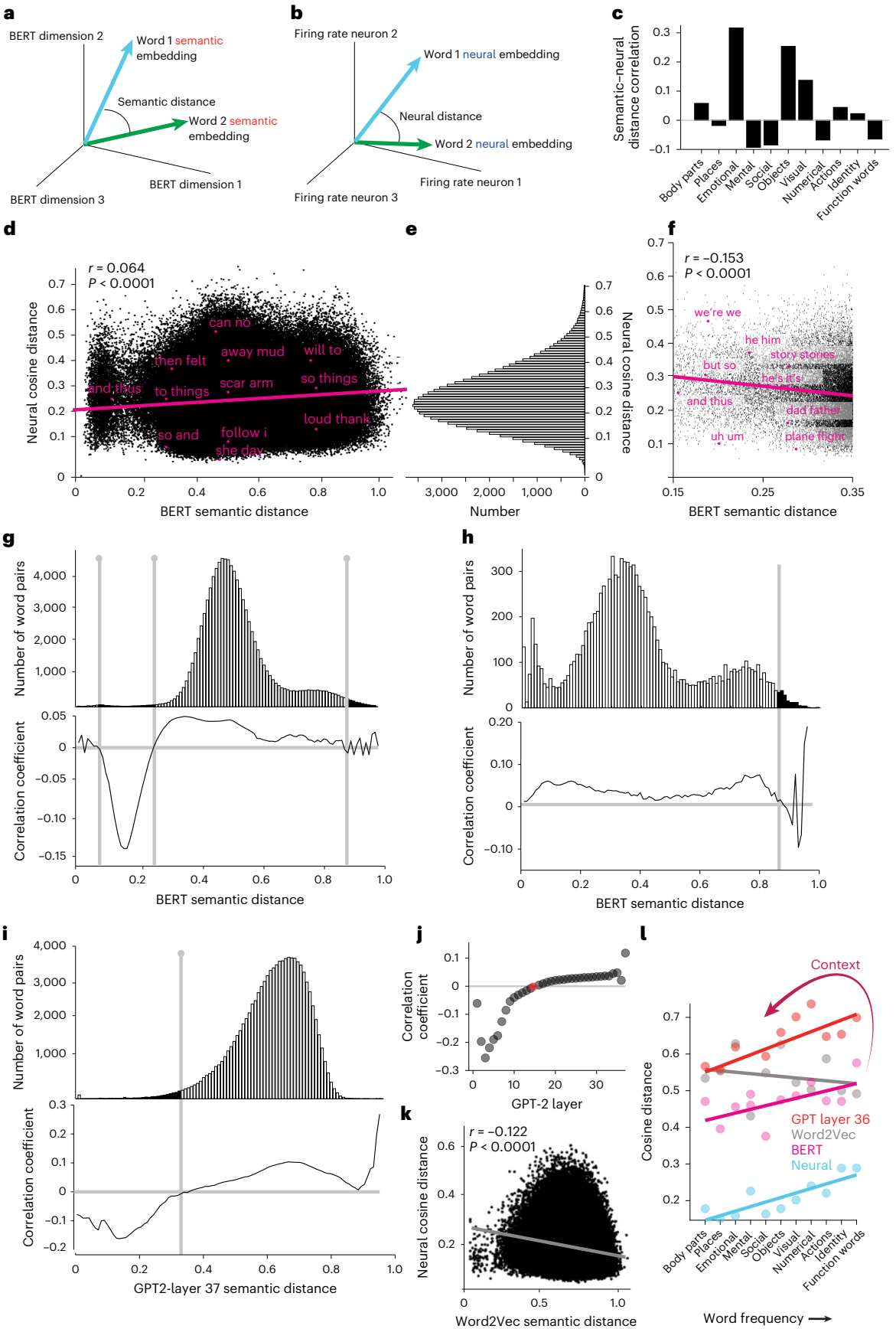
Interestingly, the distance between neural embeddings of words within a semantic category is a function of increasing word frequency (Fig. 3l, blue; this is the case for inter-category distances as well). Meaning, low-frequency body part words such as 'hand' and 'finger' are closer in both neural space and the semantic space of contextual embeddings, while high-frequency identity words such as 'everyone' and 'they' show greater distance, as seen in Fig. 3l (red, pink and blue). In other words, 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 such as Word2Vec (Fig. 3l, gray). This shift, influenced by context, aligns with the pattern of category and frequency-dependent neural distance, perhaps 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.

Semantic coding is contextual

The meaningful correspondence observed between neural and semantic distances led us to test a central prediction: that a word's semantic representation should be decodable from its neural population activity.

Fig. 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 11 semantic categories. Although most are positive, some categories have negative correlations, probably due to contrastive coding. **d**, Scatter plot illustrating the correlation between neural and semantic embedding distances between all 26,736,328 word pairs. Two-sided Pearson correlation ($r = 0.0635$, 95% confidence interval (CI) 0.0631 to 0.0638, $t = 328.99$, $P < 2.2 \times 10^{-308}$). The pink line represents the 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 0.15–0.35; this range represents 2.16% of the data ($n = 578{,}548$ word pairs). In this range, the correlation between neural and semantic distances is negative. Two-sided Pearson correlation ($r = -0.1536$, 95% CI −0.1561 to −0.1510, $t = -118.23$, $P < 2.2 \times 10^{-308}$). **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 subrange 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**, The 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**, The same data as in **g**, but using GPT-2 layer 37 rather than BERT. The 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 noncontextual 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 noncontextual and negatively correlate with neural distance. Two-sided Pearson correlation ($r = -0.1222$, 95% CI −0.1243 to −0.1201, $t = -113.16$, $P < 2.2 \times 10^{-308}$). **l**, Mean 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. Two-sided Pearson correlation, Word2vec $r = -0.22$, $P = 0.52$; BERT $r = 0.62$*, $P = 0.035$; GPT-2 $r = 0.72$*, $P = 0.01$; neural $r = 0.83$*; $P = 0.001$; regression lines are plotted.

https://doi.org/10.1038/s41593-026-02436-4



https://doi.org/10.1038/s41593-026-02436-4

We applied cross-validated reduced rank regression (RRR) to assess how well neural population activity predicted 30-dimensional semantic embeddings from Word2Vec, BERT and GPT-2[58,59]. Model performance was quantified as the cross-validated variance explained (1 − normalized mean squared error (NSE)) across predictive subspace dimensionalities (ranks, 0–15). All models showed near-zero performance at rank 0 (predicting the mean), confirming an unbiased baseline. Across ranks, predictive performance increased monotonically before reaching a plateau. GPT-2 achieved the highest accuracy (patient-averaged 1 − NSE = 0.06), followed by BERT (0.01) and Word2Vec (0.003), with a significant effect of model type (Fig. 4a; $F(2,30) = 98.74$, $P = 6.35 \times 10^{-14}$, partial $\eta^2 = 0.868$). In addition, performance plateaued at different subspace dimensionalities for each language model: median rank of 12 for GPT-2, 9 for BERT and 3 for Word2Vec (Fig. 4a, dashed lines). These optimal ranks reflect the dimensionality of the predictive neural subspace required to account for semantic structure in each embedding space. The higher optimal rank and superior performance for the contextual models (GPT-2 > BERT > Word2Vec) indicate that neural population activity shares a richer, more multidimensional linear structure with GPT-2 embeddings than with Word2Vec. This pattern implies that semantic information is distributed across multiple neural dimensions rather than confined to a single feature axis, consistent with a population-level, vector-like code. By contrast, the lower optimal rank for Word2Vec indicates that noncontextual embeddings capture a narrower, lower-dimensional component of the neural representation. The fact that GPT-2 and BERT outperform Word2Vec even at rank 1 suggests that the primary neural dimension is contextual, not categorical—if it reflected a single feature such as 'people versus places', all models should perform similarly at rank 1 (or Word2Vec would have the highest performance) because the decoder is limited to expressing only that one contrast. Together, these results demonstrate that contextual language models provide a substantially better and higher-dimensional description of the semantic information expressed in single-neuron population activity. Findings were consistent across patients (Fig. 4b,c).

Given this evidence for a high-dimensional neural–semantic space, we next examined how semantic categories are arranged within this population geometry. For each patient separately, we performed principal component analysis (PCA) of neurons' firing rates for every word (7,346 words by patient's neurons), reducing the dimensionality of their neurons' responses to two components. To visualize neural organization of semantics across patients, we first projected each patient's neural data into the first two PCs. These two components explained 44% of the total neural variance, on average. Next, we computed the centroids of each semantic category in PCs 1 and 2, and then aligned this data across patients into a common reference frame, yielding a shared low-dimensional subspace (Methods). Within individual patient subspaces and within this shared one, we noticed the first axis (PC1) separated social, objects and places, from identity, mental and action words. Notably, there appeared to be an organization based on word frequency, where high-frequency words clustered near the center and low-frequency words were more spread out (Fig. 4d). Indeed, when we computed, for each patient, their semantic category centroid to subspace center distance, we found a significant decrease in distance with increasing word frequency ($r = -0.764$, $P = 0.006$; Fig. 4e), meaning high-frequency words were more overlapping and central in the subspace.

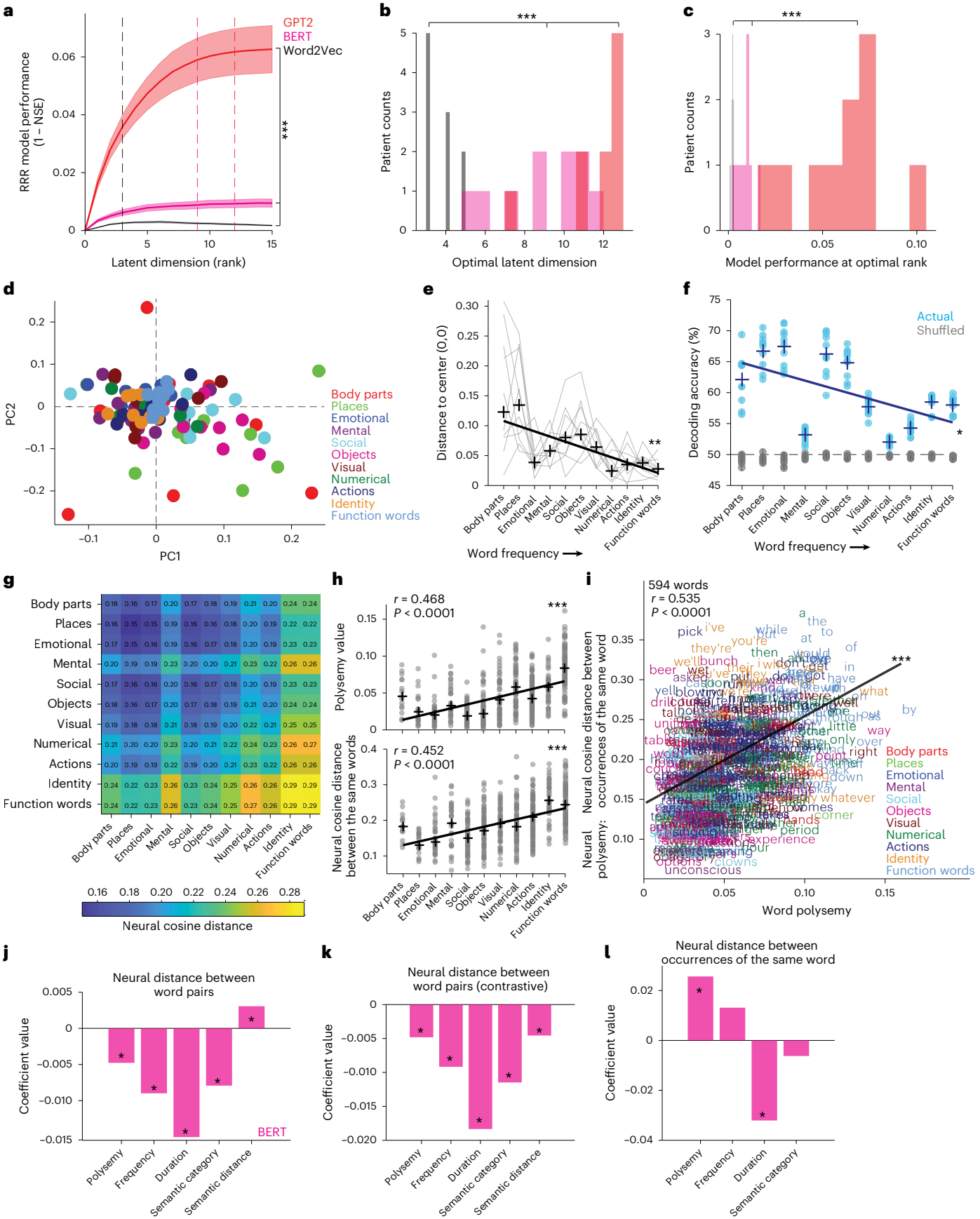
Notably, this representational geometry aligned with decoding performance, meaning semantic categories that were further from the center in the subspace were more easily decodable compared to geometrically clustered categories ($r = -0.54$, $P = 0.02$; Pearson correlation). Each semantic category can be decoded from the others using population activity (Methods), but high-frequency words such as numbers, actions and function words had reduced decoding performance compared to low-frequency words (Fig. 4f). Together, these findings suggest, as do others above, that neural coding principles differ markedly for high-frequency and low-frequency words.

Indeed, we also observe a trend in word frequency for neural distance. For word pairs within each semantic category, neural distance increases with word frequency (Fig. 3l (blue) and Fig. 4g (diagonal)), meaning neural firing rate population patterns between two common words are more distinct than patterns between two low-frequency words. In fact, high-frequency words exhibited the greatest neural separation both among themselves and relative to words from other semantic categories (Fig. 4g). Although PCA projects high-frequency words close to the center (because their distinctions lie in low-variance axes not captured by PC1 anf PC2; Fig. 4d), vectorial distance shows that they are broadly separated across the full high-dimensional neural space.

Why do common words cluster in neural state space, have reduced decoding and increased vectorial neural distance (within-category separability) compared to other, less spoken words? We think this is due, in large part, to the fact that common words are more polysemous[60] (Fig. 4h, top), 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

Fig. 4 | Population coding of words across frequency and context. a, Model performance across 15 latent dimensions (ranks) from RRR analysis predicting semantic embeddings from neural firing rate responses to words. Models were fit separately for each patient, and then averaged across. Lines represent mean performance ± s.e.m. Model performance was significantly different between models with Word2Vec, BERT and GPT-2 embeddings; one-way repeated-measures analysis of variance (ANOVA) across RRR ranks ($F(2,30) = 98.74$, $P = 6.35 \times 10^{-14}$, partial $\eta^2 = 0.868$). **b**, The optimal latent dimension for each patient and model. Optimal rank differed between models ($F(2,18) = 133.22$, $P = 1.63 \times 10^{-11}$, partial $\eta^2 = 0.937$). **c**, Model performance at the optimal rank for each patient and model. Performance differed between models; one-way repeated-measures ANOVA across patients ($F(2,18) = 59.83$, $P = 1.12 \times 10^{-8}$, partial $\eta^2 = 0.869$). **d**, Centroids of words within semantic categories from each patient plotted in a 2D shared semantic subspace. **e**, Mahalanobis distance of each semantic category centroid to the origin (0,0), computed separately for each patient (gray lines); crosses represent median distance. Distance significantly decreases across semantic categories; two-sided Pearson correlation ($r = -0.764$, $P = 0.006$). **f**, For each patient, 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). Crosses represent patient-averaged decoding accuracy. Plots display mean prediction accuracy on test observations (±s.e.m.). $P = 0.002$ for each category, two-sided Wilcoxon signed-rank test comparing actual and shuffled accuracy across patients. Decoding accuracy significantly decreases across semantic categories; two-sided Pearson correlation ($r = -0.54$, $P = 0.02$). **g**, Similarity matrix of vectorial neural distance between all word pairs within and across each semantic category. The diagonal of this matrix is plotted in Fig. 3l. **h**, Top: polysemy values for each unique and repeated word within each semantic category. Two-sided Pearson correlation ($r = 0.4689$, $P = 8.19 \times 10^{-34}$). Bottom: neural distance between different occurrences of the same word, within each semantic category. Two-sided Pearson correlation ($r = 0.4516$, $P = 3.41 \times 10^{-31}$). **i**, 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. Two-sided Pearson correlation ($r = 0.5352$, $P = 2.51 \times 10^{-45}$). **j**, Weights from a ridge regression model that predicted neural distance between all word pairs from the difference in polysemy, word frequency, word duration and semantic category for each word pair, and the BERT semantic distance between them. Asterisks represent significant predictors. **k**, The same as in **j** but restricted to word pairs with semantic distances within the contrastive coding range, 0.14–0.4. **l**, Weights from a ridge regression model that predicted, for every word, the median neural distance between all occurrences of that word from that word's polysemy (computed with BERT embeddings), frequency, duration and semantic category. The asterisks represent significant predictors. Lines in **e**, **f**, **h** and **i** are regression lines.

https://doi.org/10.1038/s41593-026-02436-4



distance between BERT embedding vectors of repeated occurrences of the same word[34–36] (n = 594 distinct and repeated words; Methods). We then computed 'neural polysemy' or the neural distance between all occurrences of the same word, and for each word, took the Pearson correlation between its median neural distance and polysemy value. We find a strong significant correlation between neural polysemy and word polysemy (r = 0.54, $P = 2.51 \times 10^{-45}$; Fig. 4i). This positive correlation remains even when function words and pronouns are removed (r = 0.49, Extended Data Fig. 5d).

We also computed polysemy for each word as the total number of synsets (that is, distinct senses) associated with each word in WordNet[61], and these polysemy values yielded the same results shown in Fig. 4h,i (Pearson r = 0.514, $P < 0.001$). Polysemy is probably driving the larger neural distance between words occurring within high-frequency categories compared to low-frequency ones (Figs. 4g 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 (Figs. 4g 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). Indeed, when words are clustered using neural embeddings rather than semantic embeddings, the resulting groups do not align with clear semantic categories, consistent with the idea that word meaning in the brain is highly context-dependent (Extended Data Fig. 8). Given that neural distance positively correlates with semantic distance in contextual models (Fig. 3), it follows that neural activity patterns also strongly correlate with polysemy (Fig. 4h,i, bottom). 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.

Because multiple linguistic features covaried with neural distance and polysemy, we next asked which of them primarily drove the observed changes in neural activity patterns. For each word pair, we used semantic distance, differences in duration, frequency and polysemy, and a binary indicator of whether the words belonged to the same semantic category (where 0 is same and 1 is different) as predictors in a cross-validated ridge regression model to predict neural distance (Methods). We found that all predictors were statistically significant, but semantic distance (from contextual models such as BERT or GPT-2; Extended Data Fig. 9) was the only one with a positive coefficient, meaning that greater semantic dissimilarity was associated with greater neural separation (Fig. 4j; see Extended Data Fig. 9 for semantic-distance results were reversed in Word2Vec). Limiting the model to the contrastive-coding range (semantic distance 0.14–0.4) produced a sign reversal for semantic distance, whereas all other predictors remained negative (Fig. 4k). This behavior identifies semantic distance as the sole predictor whose relationship to neural geometry systematically tracks the emergence of contrastive coding. In additon, we used the same model architecture to predict neural polysemy from word polysemy, word frequency, word duration and semantic category. We found that only word polysemy and duration were significant predictors of neural polysemy, with polysemy associated with greater context-driven neural variability and longer duration associated with more stable neural responses (Fig. 4l). Repeating the polysemy and related analyses using only neurons that remained semantically selective after controlling for other linguistic features produced similar results, consistent with a distributed population code for semantics, where semantic relationships are carried by patterns of activity across many neurons rather than by a small group of highly selective units (Extended Data Fig. 10).

Discussion

Here, we show that the hippocampus exhibits a distributed 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 (Fig. 2). While a single neuron can distinguish individual words, the semantic relationship between words (even multiple occurrences of the same word) is reflected in neural population activity (Figs. 3 and 4). 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. These results, then, support the idea that neuronal coding of word meanings is similar to that observed in word models and large language models (LLMs), in which single vector entries do not correspond to specific concepts, and may not, on their own, have natural interpretations. Instead, meaning emerges from the collective pattern of activity across the population.

In many prior studies, semantics has been operationalized in terms of discrete lexical categories[37,41] (for example, actions, animals or people). However, our results suggest that hippocampal neurons encode aspects of contextual semantics—how a word's meaning is modulated by its linguistic context—rather than fixed lexical semantic categories during natural language comprehension. Notably, the distributed semantic coding and mixed selectivity in single hippocampal neurons we observed appears to contrast the semantic coding regime in the left posterior middle frontal gyrus (pmFG), where neurons respond selectively to specific semantic domains[37]. Most pmFG neurons (30%) were selective for words related to actions, and most neurons that exhibited such semantic selectivity (84%) were selective for one semantic domain and no others[37]. Both that study and ours examined semantic coding as participants listened to natural sentences. Why would single neurons in two different brain regions, the pmFG and the hippocampus, encode semantics so differently? We theorize pmFG favors selective coding of literal word meanings because of its proximity and anatomical connectivity to speech production regions, whereas the hippocampus exhibits a flexible encoding of relational and context-dependent semantics, given its role in memory[62,63]. Extending single-neuron studies to other brain areas will reveal how broadly—and in what form—semantics is encoded across the human brain.

Surprisingly, at short semantic distances (~0.14–0.4), we observed an anti-correlation between neural responses and linguistic embeddings for the most similar 2–4% of words. We propose that this reversal reflects a contrastive code that helps the brain distinguish highly similar words (Fig. 3f and Extended Data Fig. 7e,g). For similar words, such as 'house' and 'home' neural coding is more different than expected from the broader pattern across randomly sampled words. This effect appears driven by semantic embedding distance, not other linguistic features such as word duration, frequency, polysemy or shared semantic category (Fig. 4j,k). It is also present in each patient individually and is highly statistically significant (Extended Data Fig. 7e,g). This pattern is not observed in word models or LLM embeddings, suggesting that it may reflect constraints specific to the brain. One possible constraint is neural noise: highly similar meanings may be more easily confused biologically than in silico. Yet, fine distinctions can be crucial. Words such as 'brother' and 'sister', 'hamburger' and 'veggieburger', or 'eighteen' and 'nineteen' have similar embeddings but distinct meanings that matter in everyday contexts (for example, when buying birthday gifts, when feeding a vegetarian friend and when paying for a parking spot). Contrastive coding may also support context-dependent distinctions, such as sincere versus ironic uses of words such as 'generous' or 'fun'. Because word models and LLMs probably have different noise properties than the brain, they may not require this additional separation. Thus, the brain may use distance-based coding at most scales, while adding contrastive elements for the most confusable words and relying on a complementary strategy at the finest scales.

Such a strategy may combine high shattering dimensionality with strong generalization, allowing simultaneous separation and abstraction[64]. This contrastive code would produce the pattern observed here, although follow-up studies with more controlled datasets are needed to fully test this hypothesis.

Another striking result is that semantic predictors had the strongest and most pervasive influence on neuronal activity, even in models with many linguistic features, despite modest absolute effect sizes: 26% variance explained in encoding models and 1–6% in population decoding models. This modest variance is expected because our naturalistic design differs fundamentally from paradigms using repeated presentations of discrete stimuli, as is common in single-neuron studies of language processing[17,33,37,65]. Our stimuli were not repeated identically but occurred once within continuous, context-rich narrative speech. Because word meaning in natural language is strongly context-dependent, averaging across repetitions would collapse distinct contextual states. Our study adds to growing work using semantic embeddings and encoding models to study neural semantic representations. Prior studies of naturalistic language have reported similarly modest effects across neural signals, including BOLD responses[41] and intracranial field potentials[9,10,38]. Human single-neuron studies using visually presented stimuli have also found comparable relationships between hippocampal activity and semantic similarity, although with different metrics and paradigms[25]. Importantly, after accounting for the sparsity and Poisson-like variability of single-unit activity, our results represent a substantial fraction of the noise-limited ceiling for semantic relationships between words: the neural semantic-distance correlation of 0.06 for BERT reaches 54% of its estimated ceiling, and GPT-2 reaches 76%. Thus, even under the demanding constraints of single-presentation naturalistic stimuli, neural population activity captures a surprisingly large proportion of recoverable semantic structure, showing that modest absolute variance can reflect robust and theoretically meaningful relationships.

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, Fig. 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 (static) 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[53–57]. Note that, contrary to our findings, another study shows a positive correlation between single neuron and semantic noncontextual embeddings in pmFG, highlighting the distinct coding repertoire used by the hippocampus[37].

If hippocampal semantic coding is distributed, then it shares key representational features with semantics in LLMs. This does not imply that the brain functions such as an LLM, and our data highlight clear differences between the two systems—for example, anticorrelations between neural and embedding distances for noncontextual embeddings and for very close words. Rather, the parallel lies in their use of high-dimensional, distributed population structure: individual units (neurons or embedding dimensions) are not interpretable on their own, yet distances in the population space reliably reflect semantic relationships. This shared geometry does not suggest a shared algorithm; it indicates that both biological and artificial systems can converge on similar representational organization. LLMs therefore provide a useful comparative framework for interpreting neural population structure, not because they mirror the brain's mechanisms but because they offer a high-dimensional representational model against which neural data can be contextualized.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-026-02436-4.

References

1. Mitchell, J. & Lapata, M. Vector-based models of semantic composition. In *Proc. ACL-08: HLT* (eds Moore, J. D. et al.) 236–244 (Association for Computational Linguistics, 2008).
2. Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A. & Rogers, T. T. Decoding semantic representations in mind and brain. *Trends Cogn. Sci.* **27**, 258–281 (2023).
3. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
4. Erk, K. Vector space models of word meaning and phrase meaning: a survey. *Lang. Linguist. Compass* **6**, 635–653 (2012).
5. Piantadosi, S. T. et al. Why concepts are (probably) vectors. *Trends Cogn. Sci.* **28**, 844–856 https://doi.org/10.1016/j.tics.2024.06.011 (2024).
6. Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron* **109**, 3055–3068 (2021).
7. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).
8. Deniz, F., Tseng, C., Wehbe, L., Dupré la Tour, T. & Gallant, J. L. Semantic representations during language comprehension are affected by context. *J. Neurosci.* **43**, 3144 (2023).
9. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
10. Goldstein, A. et al. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat. Commun.* **15**, 2768 (2024).
11. Kumar, S. et al. Shared functional specialization in transformer-based language models and the human brain. *Nat. Commun.* **15**, 5523 (2024).
12. Linzen, T. & Baroni, M. Syntactic structure from deep learning. *Annu. Rev. Linguist.* **7**, 195–212 (2021).
13. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl Acad. Sci. USA* **117**, 30046–30054 (2020).
14. Pavlick, E. Semantic structure in deep learning. *Annu. Rev. Linguist.* **8**, 447–471 https://doi.org/10.1146/annurev-linguistics-031120-122924 (2022).
15. Radford, A. et al. Language models are unsupervised multitask learners. *GitHub* https://github.com/codelucas/newspaper (2019).
16. Peters, M. E. et al. Deep contextualized word representations. *GitHub* http://allennlp.org/elmo (2018).
17. Dijksterhuis, D. E. et al. Pronouns reactivate conceptual representations in human hippocampal neurons. *Science* **385**, 1478–1484 (2024).
18. Courellis, H. S. et al. Abstract representations emerge in human hippocampal neurons during inference. *Nature* **632**, 841–849 (2024).
19. Blank, I. A., Duff, M. C., Brown-Schmidt, S. & Fedorenko, E. Expanding the language network: domain-specific hippocampal recruitment during high-level linguistic processing. Preprint at *bioRxiv* https://doi.org/10.1101/091900 (2016).

20. Piai, V. et al. Direct brain recordings reveal hippocampal rhythm underpinnings of language processing. *Proc. Natl Acad. Sci. USA* **113**, 11366–11371 (2016).
21. Duff, M. C. & Brown-Schmidt, S. The hippocampus and the flexible use and processing of language. *Front. Hum. Neurosci.* **6**, 69 (2012).
22. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
23. Katlowitz, K. A. et al. Plasticity and language in the anaesthetized human hippocampus. *Nature* **654**, 714–723 (2026).
24. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
25. Karkowski, K. et al. Semantic tuning of single neurons in the human medial temporal lobe. Preprint at *bioRxiv* https://doi.org/10.1101/2025.10.21.682935 (2025).
26. Katlowitz, K. A. et al. Attention is all you need (in the brain): semantic contextualization in human hippocampus. Preprint at *bioRxiv* https://doi.org/10.1101/2025.06.23.661103 (2025).
27. Yan, X. et al. Shared neural geometries for bilingual semantic representations in human hippocampal neurons. *Cell* https://doi.org/10.1016/j.cell.2026.05.020 (2026).
28. Kolibius, L. D. et al. Hippocampal neurons code individual episodic memories in humans. *Nat. Hum. Behav.* **7**, 1968–1979 (2023).
29. Cohn-Sheehy, B. I. et al. The hippocampus constructs narrative memories across distant events. *Curr. Biol.* **31**, 4935–4945.e7 (2021).
30. Baldassano, C. et al. Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5 (2017).
31. Keuleers, E. & Brysbaert, M. Wuggy: a multilingual pseudoword generator. *Behav. Res. Methods* **42**, 627–633 (2010).
32. Mikolov, T., Chen, K., Corrado, G. S. & Dean, J. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations* (2013).
33. Khanna, A. R. et al. Single-neuronal elements of speech production in humans. *Nature* **626**, 603–610 (2024).
34. Cruse, D. A. *Lexical Semantics* (Cambridge Univ. Press, 1986).
35. Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMO, and GPT-2 embeddings. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (eds Inui, K. et al.) 55–65 https://doi.org/10.18653/v1/d19-1006 (2019).
36. Gorfein, D. S. *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity* (American Psychological Association, 2001).
37. Jamali, M. et al. Semantic encoding during language comprehension at single-cell resolution. *Nature* **631**, 610–616 (2024).
38. Zada, Z. et al. A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron* **112**, 3211–3222.e5 (2024).
39. Schönmann, I., Szewczyk, J., de Lange, F. P. & Heilbron, M. Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding during natural listening. *eLife* https://doi.org/10.7554/elife.106543.1 (2025).
40. Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. Bag of tricks for efficient text classification. Preprint at https://arxiv.org/abs/1607.01759 (2016).
41. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
42. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
43. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
44. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
45. Johnston, W. J., Palmer, S. E. & Freedman, D. J. Nonlinear mixed selectivity supports reliable neural computation. *PLoS Comput. Biol.* **16**, e1007544 (2020).
46. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745 (1992).
47. Li, Y. et al. Neuronal representation of social information in the medial amygdala of awake behaving mice. *Cell* **171**, 1176–1190.e17 (2017).
48. Pennington, J., Socher, R. & Manning, C. D. GloVe: global vectors for word representation. *Stanford Natural Language Processing Group* https://nlp.stanford.edu/projects/glove/ (2014).
49. Devlin, J., Chang, M.-W., Lee, K., Google, K. T. & Language, A. I. BERT: pre-training of deep bidirectional transformers for language understanding. *GitHub* https://github.com/tensorflow/tensor2tensor (2018).
50. Lyu, B., Marslen-Wilson, W. D., Fang, Y. & Tyler, L. K. Finding structure during incremental speech comprehension. *eLife* **12**, RP89311 (2024).
51. Madureira, B., Kahardipraja, P. & Schlangen, D. When only time will tell: interpreting how transformers process local ambiguities through the lens of restart-incrementality. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (eds Ku, L.-W. et al.) Vol. 1, 4722–4749 (Association for Computational Linguistics, 2024).
52. Kahardipraja, P., Madureira, B. & Schlangen, D. Towards incremental transformers: an empirical analysis of transformer models for incremental NLU. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 1178–1189 (Association for Computational Linguistics, 2021).
53. Bromley, J. et al. Signature verification using a 'Siamese' time delay neural network. *Adv. Neural Inf. Process. Syst.* **6**, 737–744 (1994).
54. Chanales, A. J. H., Oza, A., Favila, S. E. & Kuhl, B. A. Overlap among spatial memories triggers repulsion of hippocampal representations. *Curr. Biol.* **27**, 2307–2317.e5 (2017).
55. Prince, J. S., Alvarez, G. A. & Konkle, T. Contrastive learning explains the emergence and function of visual category-selective regions. *Sci. Adv.* **10**, eadl1776 (2025).
56. Radford, A. et al. Learning transferable visual models from natural language supervision. Preprint at https://arxiv.org/abs/2103.00020 (2021).
57. Tian, Y., Krishnan, D. & Isola, P. Contrastive multiview coding. In *Computer Vision—ECCV 2020* (eds Vedaldi, A. et al.) 776–794 (Springer, 2020).
58. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical areas interact through a communication subspace. *Neuron* **102**, 249–259.e4 (2019).
59. Izenman, A. J. Reduced-rank regression for the multivariate linear model. *J. Multivar. Anal.* **5**, 248–264 (1975).
60. Zipf, G. K. The meaning-frequency relationship of words. *J. Gen. Psychol.* **33**, 251–256 (1945).
61. Fellbaum, C. *WordNet: An Electronic Lexical Database* (The MIT Press, 1998).
62. Squire, L. R. & Zola-Morgan, S. The medial temporal lobe memory system. *Science* **253**, 1380–1386 (1991).

63. Rutishauser, U., Reddy, L., Mormann, F. & Sarnthein, J. The architecture of human memory: insights from human single-neuron recordings. *J. Neurosci. Off. J. Soc. Neurosci.* **41**, 883–890 (2021).
64. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967.e21 (2020).
65. Xu, D. et al. Human precentral gyrus neurons link speech sequences from listening to speaking. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.27.630563 (2025).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2026, modified publication 2026

Methods

Patient recruitment

This research complies with all relevant ethical regulations; experiments were conducted according to protocol guidelines approved by the Institutional Review Board for Baylor College of Medicine and Affiliated Hospitals (H-18112). All the patients had drug-resistant epilepsy and were undergoing clinical standard-of-care intracranial seizure monitoring. Each patient provided written informed consent to participate in the study and were aware that participation was voluntary and would not affect their clinical course. Patients were provided with compensation (US$100 per research participation day). Included patients' age ranged from 18–45 years old (average, 29.1), with four female and six male patients.

Human intracranial neurophysiology

All experiments were performed in the Epilepsy Monitoring Unit at Baylor St. Luke's Hospital. Single-neuron data were recorded from stereo-electroencephalography electrodes with microwires extending from the tip (AdTech Medical Behnke-Fried). Each patient had an average of three probes terminating in the left and right hippocampus. Electrode locations were verified by coregistered preoperative magnetic resonance imaging (MRI) and postoperative computed tomography (CT) scans. Each probe includes eight microwires, each with one contact, 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 Wave_clus sorting algorithm[66] 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 and trough-to-peak) and exponentially decaying interspike interval histogram with no interspike interval shorter than the refractory period (1 ms). The analyses here used all single and multiunit activity (Extended Data Fig. 1).

Electrode visualization

Electrodes were localized using the software pipeline intracranial Electrode Visualization[67] (iELVis) and plotted across patients on an average brain using Reproducible Analysis and Visualization of iEEG[68] (RAVE). For each patient, DICOM images of the preoperative T1 anatomical MRI and the postoperative stealth CT scans were acquired and converted to NIfTI format[69]. The CT was aligned to MRI space using Functional Magnetic Resonance Imaging of the Brain Software Library[70,71]. The resulting coregistered CT was loaded into BioImage Suite (version 3.5β1)[72], and the electrode contacts were manually localized. Electrodes coordinates were converted to patient native space using iELVis MATLAB functions[73] and plotted on the Freesurfer (version 7.4.1) reconstructed brain surface[74]. Microelectrode coordinates are taken from the first (deepest) macro contact on the Ad-Tech Behnke-Fried depth electrodes. RAVE was used to transform each patient's brain and electrode coordinates into MNI152 average space[68]. 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-min in duration) taken from *The Moth Radio Hour*, which totaled 47:25 min 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 min of jabberwocky stories (1,890 words) to assess neural specificity for semantics (Fig. 1d and Extended Data Fig. 2b,c). 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[31]. 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 artificial intelligence (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 an .xls and loaded into Matlab and Python for further analysis.

Semantic embedding extraction from language models

Word2Vec. We used the pretrained fastText Word2Vec model in MATLAB to extract word embeddings for all words in our dataset[32,40]. This pretrained 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 such as 'Applebee's' that did not have word embeddings were discarded from the analysis (these were rare in our sample).

GPT-2 Large. 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 multiword 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 subword alignment errors during tokenization.

To extract GPT-2 embeddings, we used the 'gpt2-large' model[15] from the Hugging Face Transformers library[75]. 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 1,024 tokens. Although GPT-2 supports a 1,024-token context window, all embeddings used in this study were computed within individual sentences only—that is, the model's context was reset whenever a sentence-ending delimiter (for example, '.', '!' and '?') was encountered. We extracted hidden states from all 37 layers using a progressive context approach designed to incorporate GPT-2's autoregressive nature. 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 tokenizer, ran the token sequence through the model in no-grad (inference-only) mode and retrieved the hidden states. We then identified the newly appended word's subword tokens and averaged

their corresponding vectors to obtain a single word-level embedding. Because GPT-2 often splits words into multiple subword fragments (otherwise known as 'tokens', for example, 'computa' + 'tion'), averaging those fragments preserves consistency with the transcript's original word boundaries. The GPT-2 outputs resulted in 1,280-dimensional embeddings for each layer, as seen in Fig. 3j. All analyses with GPT-2 used word embeddings from the last layer 37, unless otherwise specified, as in Fig. 3j.

BERT. We conducted a parallel extraction procedure with the 'bert-base-cased' model[49], also accessed via the Hugging Face Transformers library[75]. 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 that processes all input tokens in parallel, granting access to both past and future context. To align BERT's outputs more closely with the unidirectional processing of GPT-2, we applied an expanding-context procedure: beginning with an empty string, we appended one new word at a time, concatenated the running text (preserving punctuation for contextual fidelity) and re-tokenized the updated segment with BERT's tokenizer. This process continued until a sentence-ending delimiter (for example, '.', '!' or '?') was encountered, at which point the context was reset, ensuring that embeddings were computed strictly within sentence boundaries. Although BERT inherently considers bidirectional information, we extracted only the hidden states corresponding to the just-added word, thereby restricting interpretation to left-contextual information only. Note that we extract BERT embeddings in a left-to-right, unidirectional manner for all analyses in this study, as seen in Lyu et al. [50], Madureira et al. [51] and Kahardipraja et al. [52]. Subword vectors belonging to that word were averaged to yield one word-level embedding per layer. Hidden states from all 13 layers (768 dimensions each) were retained, and analyses reported in this paper used the final hidden layer 13. All code was implemented in Python using PyTorch[76] as the underlying deep-learning framework, and the resulting embeddings were saved to Excel spreadsheets for subsequent analyses.

Clustering of semantic embeddings

To identify the natural semantic categories present in our word data (Fig. 1h–j), all unique words were clustered into groups based on related meanings using a word embedding approach[32,40]. For each word, we used its 300-dimensional Word2Vec embedding. 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 two-dimensional (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[42]). 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[37,41]. We repeated this same method to create semantic clusters from neural data (Extended Data Fig. 8) and contextual embeddings (Extended Data Fig. 6).

Firing rate responses to words

We computed the duration for each word in the story and for each nonword in Jabberwocky stimuli. The firing rate of a neuron for each word was the number of spikes occurring during the word with an 80-ms delay after word onset to account for the approximate delay of information to hippocampus. This value was divided by word duration and then multiplied by 1,000 to compute spikes per second.

Regressing spike counts on word embeddings, phonemes and linguistic features

Word embeddings, such as those derived from Word2Vec or similar models, are typically high-dimensional and can exhibit substantial collinearity among dimensions—that is, many of the embedding features are correlated with one another. We first used PCA (`pca`, Matlab 2023b) on the full word embeddings to obtain uncorrelated features that still capture the dominant structure in the embedding space with reduced dimensionality. For each word, we used the first 100 PC from Word2Vec or BERT embeddings because they explained at least 60% of the variance of the original embedding vectors.

We modeled the spike count responses of individual neurons using a Poisson generalized linear model with a log-link function and ridge (L2) regularization (we confirmed that spike counts for each word follow a Poisson distribution; Fig. 1f,g). The model aimed to predict the number of spikes a neuron fired in response to each word (summed spikes across word duration), using the top 100 PCs of the word embedding vector, the duration of the word and the interaction between each PC and the word's duration as predictors (*z*-scored before model fitting). This resulted in a 201-dimensional feature vector for each word: 100 PC values, 1 duration value and 100 PC × duration interaction terms.

The Poisson generalized linear model assumes that the spike count y_i for word i is drawn from a Poisson distribution with mean λ_i, in which

$$y_i \approx \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + x_i^T\beta,$$

where $\lambda_{i,n}$, the expected spike count for word (i) and neuron n (n), is modeled by the linear regression coefficients (β), one for each of the 201 predictors (x_i^T), plus the y intercept (β_0).

For each neuron, we fit a Poisson ridge regression model using a nested cross-validation framework with five contiguous outer folds. Words were partitioned into five sequential test blocks of approximately equal size, such that in each outer fold one contiguous block of words served as the held-out test set and the remaining four blocks were used for training. This blocked outer-fold design was used to reduce temporal leakage arising from autocorrelation across nearby words in the continuous speech stream. Within each outer training set, the optimal regularization strength (α; alpha) was selected using fivefold inner cross-validation, drawn from a log-spaced range (10^{-3} to 10^{3}), and the value maximizing the mean validation log-likelihood across inner folds was selected. Using the selected α, a Poisson ridge model was then fit to the full outer training set and then evaluated on the held-out test set using several performance metrics, including log-likelihood and adjusted R^2. To assess model complexity, we estimated the effective degrees of freedom from the training data and used this to compute an out-of-sample AIC-like metric for the test set. Final summary statistics for each neuron were obtained by taking the median across the five outer folds.

To assess statistical significance relative to a null, we generated 100 shuffled predictor matrices per dataset, in which embeddings were shuffled while keeping word duration preserved. Each shuffle circularly shifted the embedding matrix along the observation axis by a randomly selected offset of at least ±100 samples (to avoid short-range autocorrelation artifacts while maintaining some temporal autocorrelation in the shuffled model, see Nastase et al. [77]). To determine if the actual model showed significant improvement from the shuffled, we performed a permutation test comparing the actual model LLH to the shuffled model LLH distribution. We then took the difference between the actual and shuffled model (mean) LLH (LLH diff; ΔLLH). A neuron was fit by the model if its ΔLLH was both positive and significant and if McFadden's pseudo-R^2 value was positive ($P < 0.05$, permutation test; Fig. 1g, left). Shuffled model R^2 values were significantly less than those from actual models ($P < 0.05$, Wilcoxon signed-rank test; Extended Data Fig. 4c).

Predictor significance was assessed using *P* values from a Poisson generalized linear model fit to the same predictors, whereas coefficient magnitudes were taken from the L2-regularized Poisson regression model used for prediction. As expected, word duration was always a significant predictor, followed by the first eight embedding dimensions and their interactions with duration, with the number of significant predictors gradually decreasing across higher dimensions (Extended Data Fig. 4a,b,g,h).

We used the same Poisson linear ridge regression model to predict neural activity from past and future word embeddings (Fig. 1g, right). 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 and word duration 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 (ΔLLH) from every neuron was compared to that of a null model (chance distribution, described above). For each word index, these distributions were compared using the one-sided Wilcoxon signed-rank test ($P < 0.05$).

Phonemic regression. We built a fixed-length feature space over the 43 IPA phonemes observed in our corpus. Here, IPA refers to the International Phonetic Alphabet, a standardized set of symbols that denote speech sounds independent of spelling. The podcast transcript was processed word by word: tokens were lower-cased and converted from graphemes to phonemes using the Python phonemizer package (version 3.0.1) with the eSpeak backend as the grapheme-to-phoneme (G2P) model (https://pypi.org/project/phonemizer/3.0.1/). For each word, we took the canonical IPA transcription returned by phonemizer. Phonemes were then manually inspected and corrected by linguists (J.B. and B.J.) in Praat. Each word was then represented as a 43-dimensional count vector whose entries reflect the number of occurrences of each IPA phoneme in that word (unigram counts; no duration or stress information). This approach is similar to Khanna et al. [33], in that words are represented by the presence of their constituent phonemes; however, we did not group phonemes into broader classes, in order to preserve the full identity of the 43 phones observed in our data and maintain the embedding's informational richness and dimensionality[33]. To better differentiate words that contain the same phoneme set but in different orders (for example, dog versus god), we double-counted the first phoneme in every word while counting all subsequent phonemes once. The resulting word-by-phoneme matrix (*N*words × 43 features) was used in subsequent analyses relating phonetic composition to neural responses. We used this word-by-phoneme matrix, word duration and their interactions in the same Poisson regression described above for predicting spike counts of the current word. The data folds used for training and testing were preserved across semantic and phonetic models and both models had the same number of words. The results in Extended Data Fig. 4e,f were replicated even when using the same number of predictors (the first 43 PCs of semantic embeddings compared to 43 phonemes).

We used the same Poisson ridge regression model to predict an individual neuron's spike count for each word from 61 linguistic features: (1) full phonemic embeddings (43 IPA phones; count coded, compare to Khanna et al. [33]), (2) ten PCs of semantic embeddings, (3) word order in sentence, (4) word order in clause (Stanford CoreNLP), (5) opening node number (Stanford CoreNLP), (6) closing node number (Stanford CoreNLP), (7) word frequency (SUBTLEX-US), (8) pitch, (9) a scalar representing degree of polysemy and (10) word duration[34,35,78] (Extended Data Fig. 4g,h).

Single-neuron encoding of words

Semantic encoding of each neuron was quantified using a ROC analysis, a standard approach that has previously been used to characterize neural responses during behavior[46,47]. 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 four 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 and the false-positive rate. Plotting the true-positive rate against the false-positive rate 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 (Fig. 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 and took the absolute value of that difference. This results in values between 0 and 0.5 (Fig. 2c).

Semantic range

Each neuron encodes a given set of words determined from ROC analysis ('Single-neuron encoding of words' section). For each neuron that encoded at least two words, we quantified its semantic range as the distribution of pairwise cosine distances among embeddings of the words it encoded. For contextual embeddings, including GPT-2 and BERT, distances were computed using the contextualized embeddings for all occurrences of each encoded word. For Word2Vec, which provides a static embedding for each word, distances were computed using one embedding per encoded word type.

We compared each neuron's empirical semantic range to a matched null distribution generated from the ROC-testable word set. For each neuron, we randomly sampled the same number of word embeddings as in that neuron's empirical encoded set and computed all pairwise cosine distances among the sampled embeddings. This procedure was repeated 1,000 times per neuron. Thus, each neuron's null distribution was matched to its empirical distribution size and reflected the semantic-distance structure of the testable vocabulary. To quantify how much of the null semantic-distance distribution was covered by each neuron's empirical range, we computed the proportion of null distances that fell between the empirical minimum and maximum distance for that neuron. This value, termed the null-coverage proportion, ranges from 0 to 1, with larger values indicating broader semantic coverage. We classified neurons as broadly tuned if their empirical range covered at least 70% of the matched null semantic-distance distribution. At the population level, we tested whether the median null-coverage proportion across neurons exceeded this 70% broad-coverage criterion using a one-sided Wilcoxon signed-rank test. We performed this analysis separately for each embedding model (Word2Vec, GPT-2 (layers 37 and 25) and BERT (last layer)).

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 (1,280 in length) language models. For each word, its neural embedding vector was every neuron's firing rate response for that word (356 in length). Neurons were combined across

all ten 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, as 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 Fig. 1h, for the neural and semantic distances within each category in Fig. 3l and for the neural distance similarity matrix in Fig. 4g. For Fig. 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 Fig. 3g–j. In Fig. 4h,i, 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

$$\text{cosine distance}\,(w_1, w_2) = 1 - \frac{W1 \times W2}{||W1||\,||W2||}.$$

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.

Estimating a neural noise ceiling for semantic embeddings

To estimate a noise ceiling constrained by the variability observed in the neural data, we generated noisy versions of the GPT-2 (or BERT or Word2Vec) embedding space by injecting Gaussian noise whose magnitude was scaled according to the firing rate statistics of the recorded neurons. Let $G \in R^{W\times D}$ denote the clean GPT/linguistic embedding matrix (rows = words, columns = embedding dimensions) and $R \in R^{W\times N}$ the neural firing rate matrix (rows = words, columns = neurons).

Because we empirically confirmed that hippocampal firing rates across words follow Poisson-like statistics, with each neuron's variance closely matching its mean, we used a Poisson noise model in which the variance-to-mean ratio was fixed at $\phi = 1$ for all neurons. This provides a conservative, interpretable baseline for quantifying how neural-level variability would distort the representational geometry of the language model embeddings. Then, we injected additive Gaussian noise into the embedding space using a subset of embedding dimensions equal in number to the recorded neurons ($D = 356$).

For each word w and embedding dimension d

$$\tilde{G}_{w,d} = G_{w,d} + \varepsilon_{w,d}, \varepsilon_{w,d} \approx N(0, \phi\,\mu_{w,n}),$$

where $\mu_{w,n}$ is the mean firing rate of neuron n (taken from R) corresponding to embedding dimension d, and $\phi = 1$. Thus, the additive noise variance scaled linearly with the neurons' mean firing rates, emulating Poisson variability across words. This model assumes that noise enters independently of the embedding magnitude, perturbing all points in embedding space by a constant variance level proportional to the corresponding neuron's mean firing rate.

For each noise model, pairwise cosine-distance matrices were computed (using the function, `pdist`, in Matlab 2023b) for the noisy language embeddings (D_{noisyGPT}), the original language embeddings (D_{GPT}) and the neural data (D_{neural}). The noise-limited ceiling was defined as the Pearson correlation between D_{neural} and D_{GPT}

$$r_{\text{ceiling}} = \text{corr}(D_{\text{noisyGPT}}, D_{\text{GPT}}).$$

The observed brain–model correspondence was computed as

$$r_{\text{real}} = \text{corr}(D_{\text{GPT}}, D_{\text{neural}})$$

The ratio $r_{\text{real}}/r_{\text{ceiling}}$ quantifies how much of the model's representational structure is recoverable in the neural data, relative to the theoretical maximum allowed by intrinsic neural variability.

RRR

To quantify how hippocampal population activity predicts high-dimensional semantic embeddings of heard words, we applied RRR using the Matlab code and following the framework of Semedo et al.[58]. For each patient, neurons served as predictors and individual words as observations (Fig. 4a–c).

Neural data preprocessing: For each patient, we computed per word firing rates normalized by word duration ('Firing rate responses to words' section). The resulting matrix contained one observation per word and one column per neuron (X). Word-level confounding variables—the log of word frequency and position—were regressed out of the neural activity matrix, and the residuals were used in analysis.

The first 30 PCs of semantic embeddings from Word2Vec, BERT layer 13 or GPT-2 layer 37 were used in the regression, forming a word by semantic embedding matrix (Y). RRR estimates a mapping from neural activity X to semantic embeddings Y that is constrained to lie in a low-dimensional shared subspace. We used the Semedo et al.[58] `ReducedRankRegress` function, which implements the closed-form solution for

$$Y = XB + \epsilon B = B_{\text{OLS}} V_r V_r^{\top},$$

where V_r are the top r eigenvectors of $Y^{\top}XX^{\top}Y$. For each patient, we computed: B, the ranked predictive components, V: the eigenvectors describing predictive dimensions.

These define the semantic subspace captured by the neural population. To determine the number of predictive dimensions, we evaluated models with 0–15 RRR dimensions using tenfold cross-validation, following Semedo et al.[58]. Model performance (averages across folds) was quantified as the cross-validated variance explained (1 − NSE) across predictive subspace dimensionalities (ranks = 0–15). The optimal dimensionality (rank) can be defined as the number of dimensions needed for the prediction accuracy to plateau, selected using the Semedo et al.[58] function, `ModelSelect`. These RRR results quantify how many dimensions of semantic variation can be linearly predicted from hippocampal population activity, as well as the geometry of the neural to semantic mapping.

PCA

We computed components (scores) from `pca`, MATLAB 2023b, on the word by neuron firing rates matrix for each patient (neuron count ranges from 11–54 per patient). To visualize and compare semantic representations across patients, we first constructed a common two-dimensional subspace (Fig. 4d). For each patient, word-by-neuron firing rate matrices were reduced with PCA and the first two PCs (PC1–PC2) were retained. Because the scale and covariance structure of these axes can vary across patients, we whitened each patient's PC1–PC2 scores using their within-patient covariance matrix. This whitening procedure ensures that Euclidean distances in the transformed space are equivalent to Mahalanobis distances in the original coordinates.

Semantic category centroids were then computed as the mean of the whitened word representations belonging to each category. To align patients into a common reference frame, we applied an orthogonal Procrustes analysis to the configuration of category centroids, restricting the transformation to a pure rotation (no reflection, scaling or translation). This yielded a shared coordinate system in which all patients' category centroids could be visualized and compared.

Within this common space, we quantified the distance of each category centroid to the origin (0,0). Because whitening centers each

patient's space at their mean response, the origin corresponds to the population-average baseline. Distances were computed as the Euclidean norm of the whitened coordinates, which is equivalent to the Mahalanobis distance to the patient's mean response in the original PC1–PC2 space. These values provide a standardized measure of how strongly each semantic category deviates from the center of the shared semantic subspace. Results in Fig. 4d,e were replicated using Euclidean distance between centroid and center (0,0) within each patient's individual 2D subspace.

SVM decoder

We used a support vector machine (SVM) decoder[79] 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 (Fig. 4f). For each patient, each semantic category, we used the firing rate responses from every neuron ($n = 11–54$ across patients) 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 $C = 1.3$ 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 nonword from the jabberwocky stimuli from neural population activity (Extended Data Fig. 2c). We excluded the function words from jabberwocky stimuli in this analysis since those words are real words but used in a nonsensical context.

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 two occurrences, 594 words total). We took the median of this distribution to obtain one polysemy value per word (Fig. 4h,i). The distribution of polysemy values for all words in each semantic category is plotted in Fig. 4h, top. We repeated the polysemy analysis using neural embeddings instead of BERT embeddings to compute the neural cosine distance values in Fig. 4h, bottom, and Fig. 4i.

Modeling neural distance and polysemy from linguistic features

We used the same cross-validated, linear Ridge Regression architecture described in the 'Regressing spike counts on word embeddings, phonemes and linguistic features' section to predict neural distance or neural polysemy from different linguistic features. Models fit in Fig. 4j,k used word pairs as observations and semantic distance, differences in duration, frequency and polysemy, and a binary indicator of whether the words belonged to the same semantic category (where 0 is same and 1 is different) as predictors in a cross-validated ridge regression model to predict neural distance. In Fig. 4l, the model used each word that was repeated at least twice as observations, and the word's polysemy, frequency, duration and semantic category (as defined by Word2Vec: categories ranged from 1 to 11, treated as a categorical), were predictors for neural polysemy of that word. Ridge regularized coefficients/weights from these models are plotted in Fig. 4j–l.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available on the Data Archive for the Brain Initiative at https://dabi.loni.usc.edu/projects/W2ZHP46ITR2D. Source data are provided with this paper.

Code availability

The code underlying this work relies on Python 3.11.7 and Matlab 2023b or higher. The custom code written to analyze the data and generate Figs. 1–4 and Extended Data Figs. 1–10 is available via Github at https://github.com/mcfranch/PopulationCodeForSemanticsHippocampus.

References

66. Chaure, F. J., Rey, H. G. & Quian Quiroga, R. A novel and fully automatic spike-sorting implementation with variable number of features. *J. Neurophysiol.* **120**, 1859–1871 (2018).
67. Groppe, D. M. et al. iELVis: an open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *J. Neurosci. Methods* **281**, 40–48 (2017).
68. Magnotti, J. F., Wang, Z. & Beauchamp, M. S. RAVE: comprehensive open-source software for reproducible analysis and visualization of intracranial EEG data. *Neuroimage* **223**, 117341 (2020).
69. Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* **264**, 47–56 (2016).
70. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**, 143–156 (2001).
71. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
72. Joshi, A. et al. Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics* **9**, 69–84 (2011).
73. Yang, A. I. et al. Localization of dense intracranial electrode arrays using magnetic resonance imaging. *Neuroimage* **63**, 157–165 (2012).
74. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).
75. Wolf, T. et al. Transformers: state-of-the-art natural language processing. *GitHub* https://github.com/huggingface/ (2020).
76. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. Preprint at https://arxiv.org/abs/1912.01703 (2019).
77. Nastase, S. A., Gazzola, V., Hasson, U. & Keysers, C. Measuring shared responses across subjects using intersubject correlation. *Soc. Cogn. Affect. Neurosci.* **14**, 667–685 (2019).
78. Manning, C. et al. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (eds Bontcheva, K. & Zhu, J.) 55–60 https://doi.org/10.3115/v1/P14-5010 (Association for Computational Linguistics, 2014).
79. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).

Acknowledgements

We thank J. Johnston, J. Adkinson, J. Fine, O. Buhlak, V. Gates, R. Mathura and G. Kokalas for invaluable assistance.

Author contributions

M.F. and B.Y.H. conceptualized the study and designed the experimental procedure and downstream analysis. B.Y.H. and S.A.S. supervised the work. M.F. led the analysis and generated the figures. E.A.M., J.L.B., B.J., K.A.K., H.Z., A.G.C. and A.C. all helped to collect, curate and manage the data. S.A.S. and D.L.P. performed the stereotactic surgery. A.M.G., V.K. and A.M. all worked closely with the scientific team to help identify patients for study participation. E.B., S.K., S.T.P., N.R.P. and J.A.H. all contributed to data interpretation and paper drafting and editing and their framing within the existing literature.

Funding

This research was supported by the McNair Foundation (grant nos. NIH R01 MH129439 and NIDCD F32 DC023126).

Competing interests

S.A.S. has consulting agreements with Boston Scientific, Zimmer Biomet, Koh Young, Abbott Laboratories and Neuropace. S.A.S. is a cofounder of Motif Neurotech. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41593-026-02436-4.

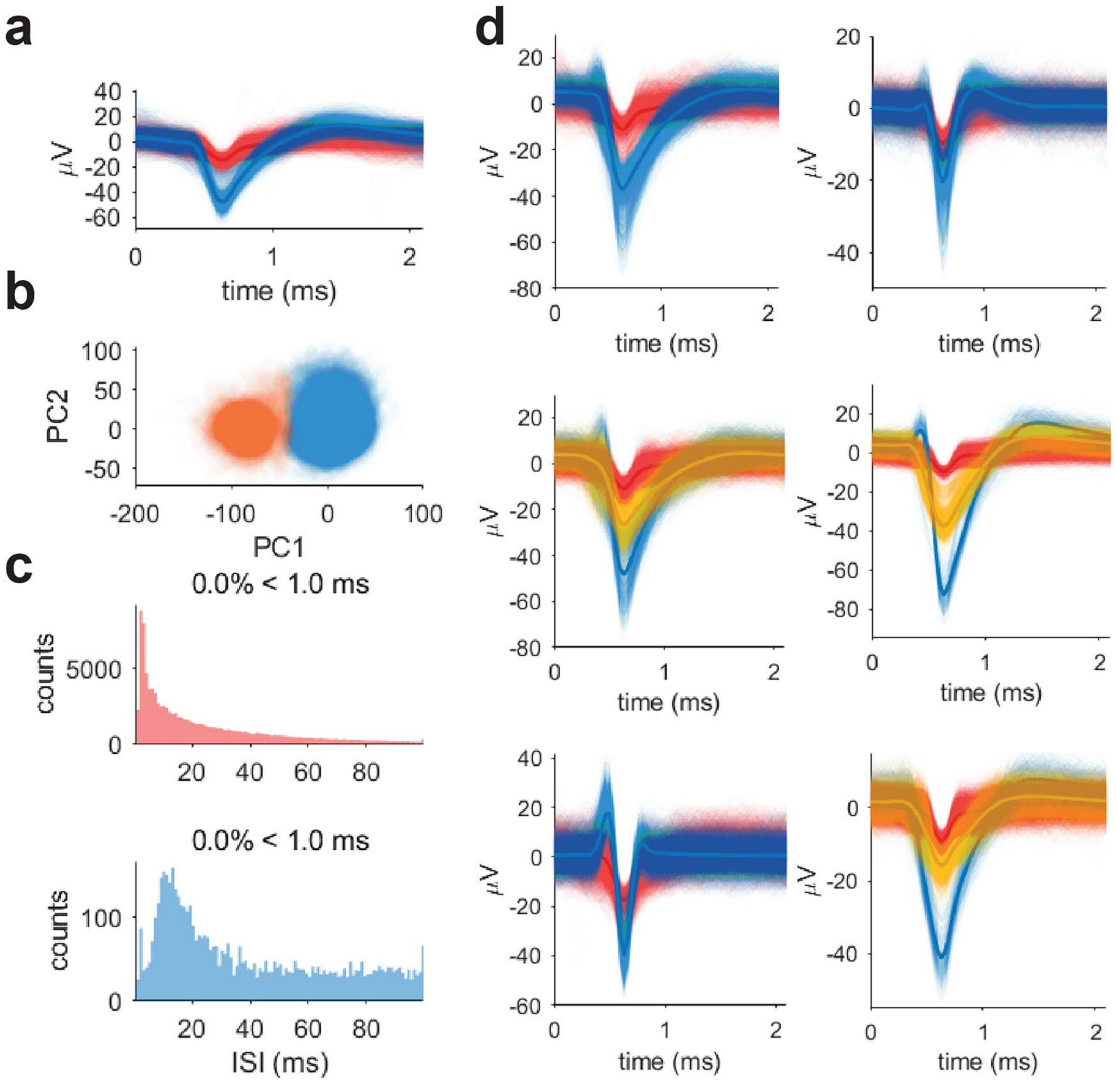
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-026-02436-4.

Correspondence and requests for materials should be addressed to Melissa Franch.

Peer review information *Nature Neuroscience* thanks Andrea Martin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

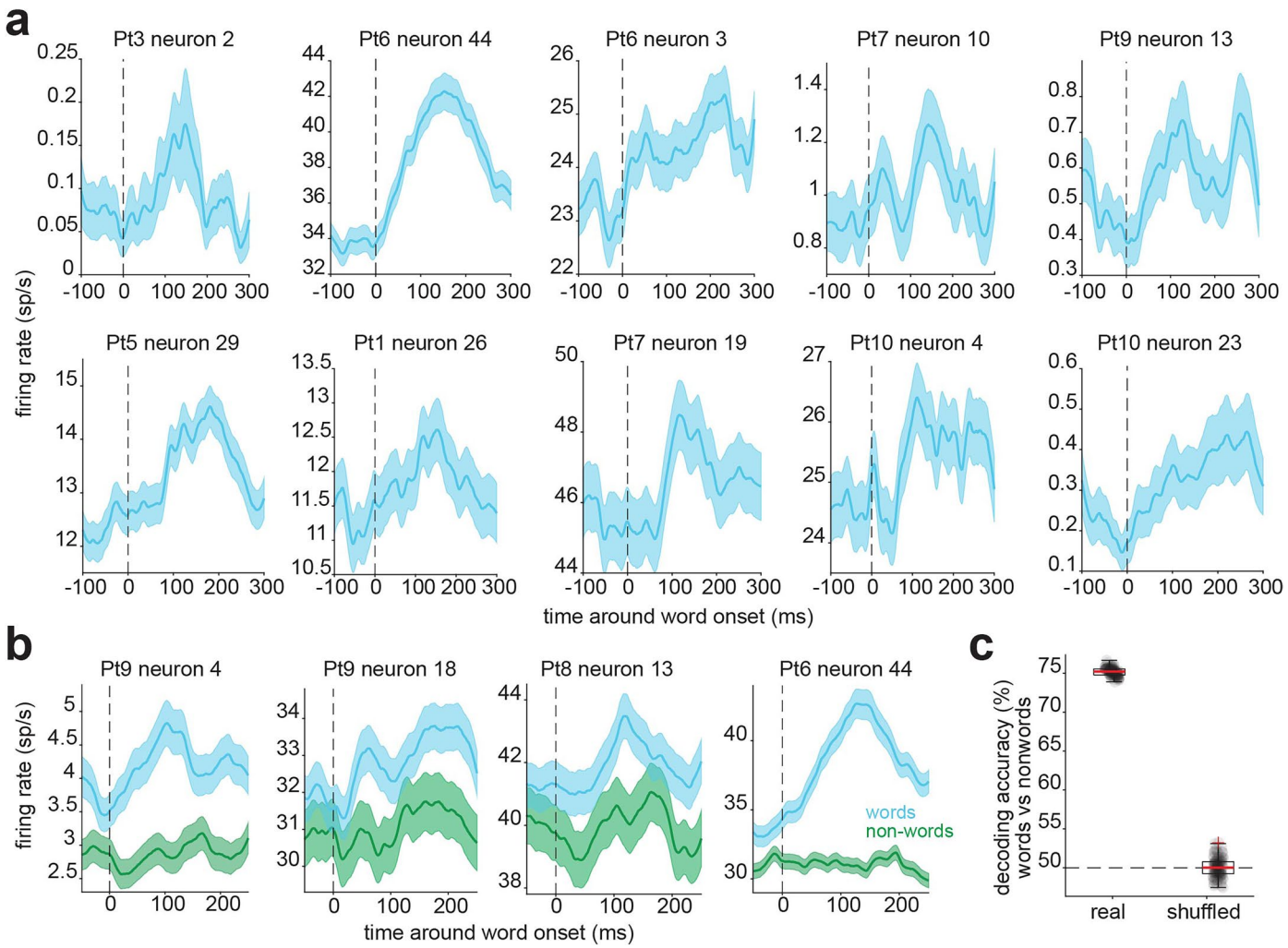
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41593-026-02436-4



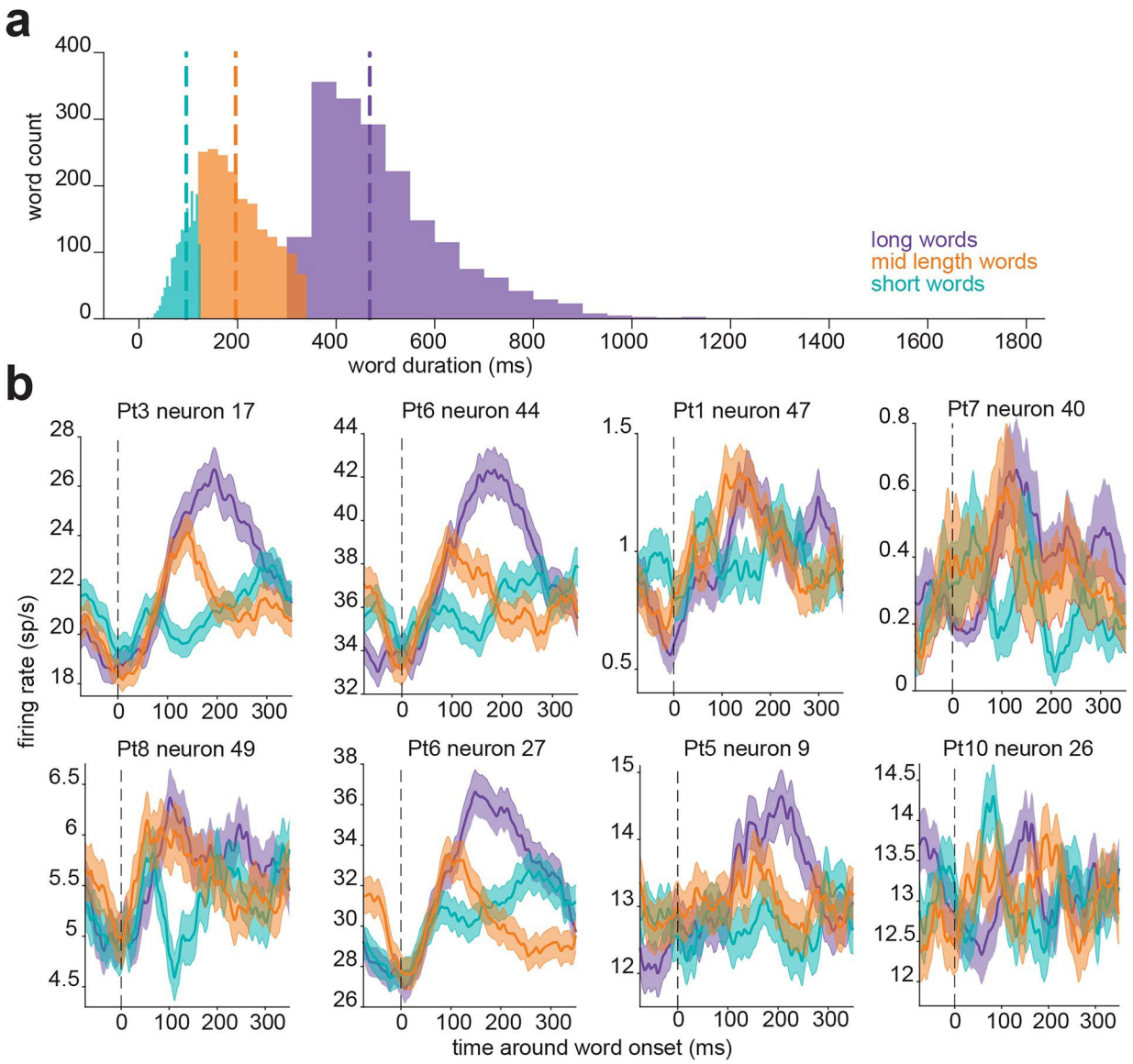
Extended Data Fig. 1 | Single and multi-unit detection and isolation. **a**, A random sample of 5,000 individual waveforms from a single hippocampal neuron (blue) and multi unit cluster (red) from patient 8 recorded on one electrode channel. Bolded lines represent average waveforms. Noisy signals were already removed using *Wave_clus* sorting algorithms and manual inspection. **b**, All waveforms from each cluster in **a** are well isolated in Principal Component space. **c**, Interspike interval (ISI) histograms from the neurons in **a**. Single and multi-unit clusters always have refractory period violations below 5% (as shown in figure titles). **d**, Example hippocampal single (yellow and blue) and multiunit (red) waveforms on individual electrode channels from four patients that were identified and sorted following the criteria in **b-c** using *Wave_clus* software and manual inspection.

https://doi.org/10.1038/s41593-026-02436-4



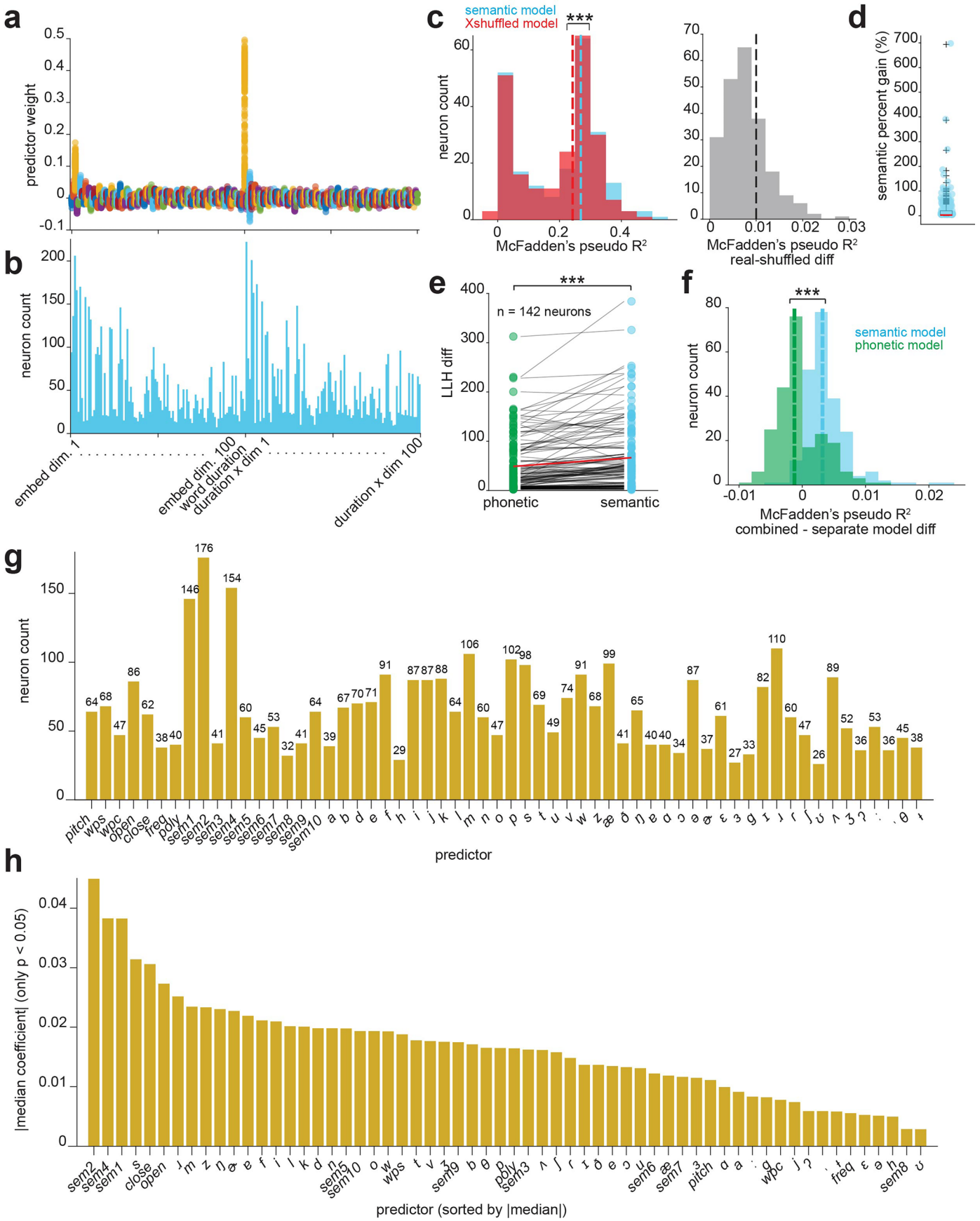
Extended Data Fig. 2 | Hippocampal neuron responses to words and non-words. a, Peri-stimulus time histograms of firing rate responses to words from example hippocampal neurons in seven different patients during language comprehension. Plots display mean firing rate ±SEM shaded. **b**, Peri-stimulus time histograms of firing rate responses of neurons exhibiting significant changes in firing rate to words and non-words (Wilcoxon rank-sum test with FDR correction, $P < 0.05$). Example hippocampal neurons are shown from 3 different patients. Plots display mean firing rate ±SEM shaded. **c**, Linear decoding accuracy for words and nonwords (jabberwocky stimuli) from neural population activity comprised of all neurons across four patients (n = 125 neurons), for models using real and shuffled data. In each boxplot, the center red line marks the median, box edges show the 25th and 75th percentiles, whiskers extend to the most extreme non-outliers, and model accuracies from each iteration are scattered on top as points. Shuffled model performance is significantly lower than actual (mean paired difference = 23.54, 95% bootstrap CI = [23.2965, 23.7780], n = 100 paired model iterations, P = 9.999e-05, one-sided paired sign-flip permutation test).

https://doi.org/10.1038/s41593-026-02436-4



Extended Data Fig. 3 | Single neuron responses to words of different spoken durations. a, Histograms of the word durations for short (1,852 words), middle length (1,852 words randomly sampled to match other groups), and long words (1,842 words). Dashed lines represent the median word duration for each group. **b**, Peri-stimulus time histograms of firing rate responses to words of different durations from example hippocampal neurons in seven different patients during language comprehension. Dashed lines represent word onset. Plots display mean firing rate ±SEM shaded.

https://doi.org/10.1038/s41593-026-02436-4

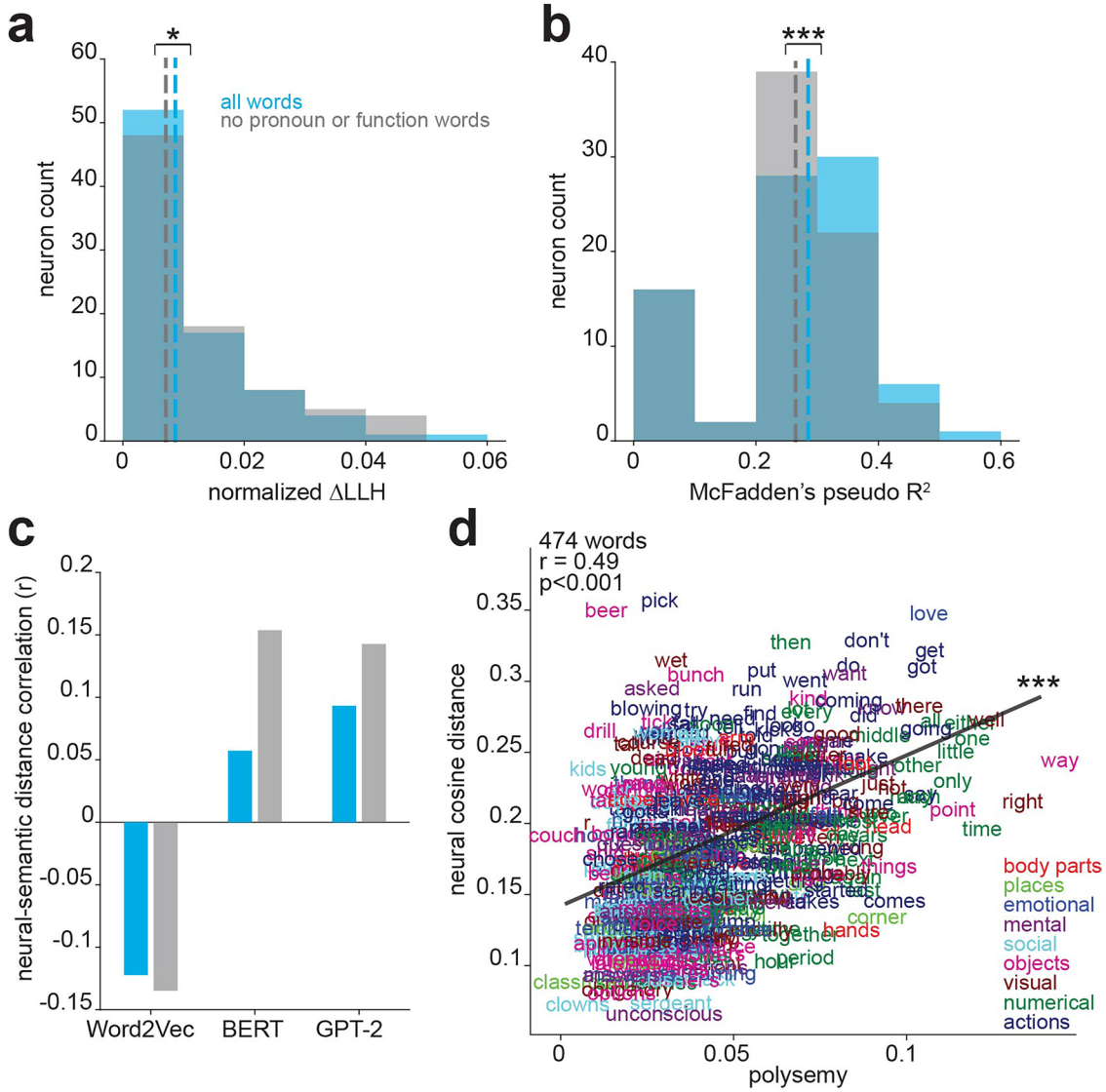


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41593-026-02436-4

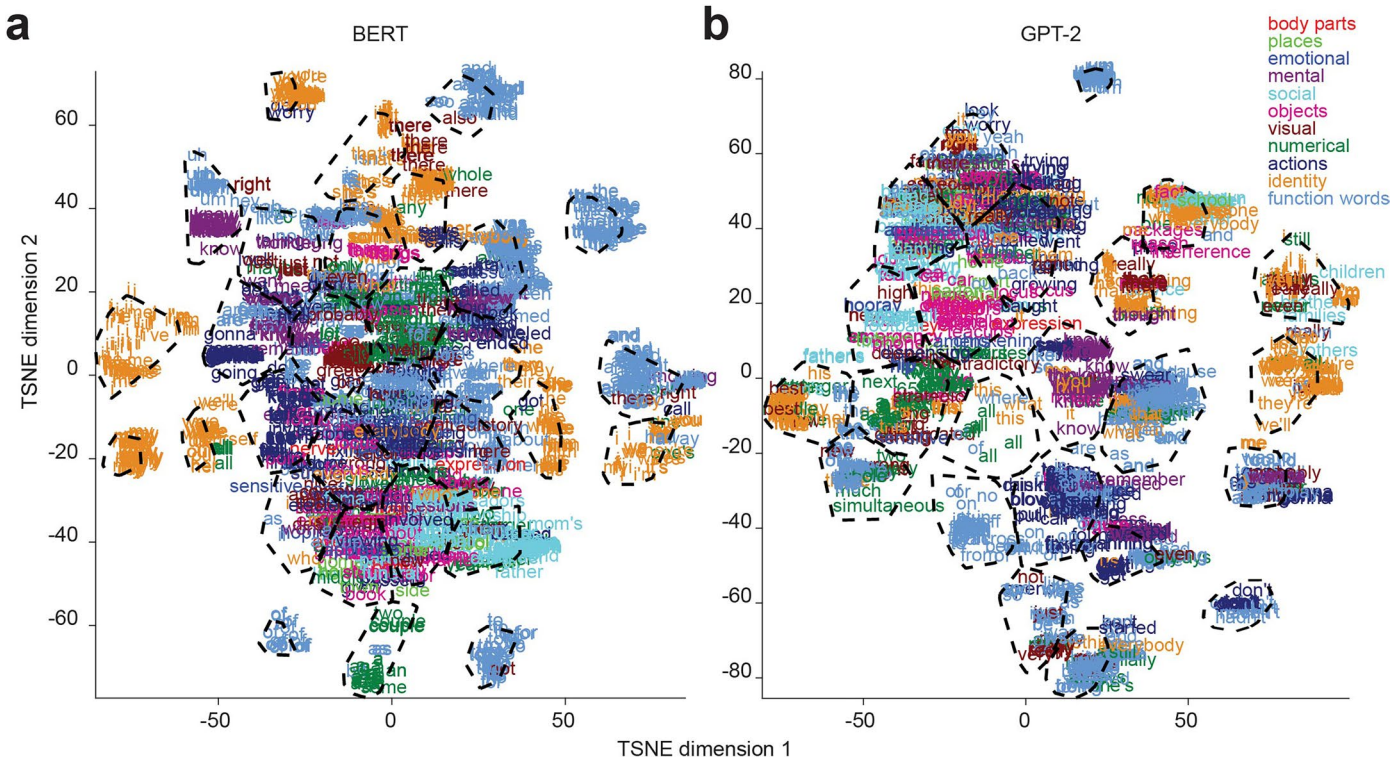
Extended Data Fig. 4 | Semantic, phonetic, and combined linguistic feature regression models. a, Semantic model (Word2Vec) coefficients of significant predictors of neurons fit by the model. Word duration and embeddings 1-8 had the largest absolute valued weights. **b**, The number of neurons for which each predictor (as seen in **a**) was significant. PC embeddings 1-8, word duration, and their interactions were significant predictors in most neurons fit by the semantic model. **c**, Left - McFadden's pseudo R-squared for neurons fit by the actual and embedding shuffled (X-shuffled) semantic model. Values from the real semantic model are significantly higher than shuffled; Two-sided Wilcoxon signed-rank test, P = 1.69e-13. Right- The difference in neurons' Pseudo R-squared values between the semantic and X-shuffled model distributions (left). **d**, For each neuron, the relative increase in pseudo-R^2 when semantic embeddings are added to a word duration-only model. Specifically, it is R^2 of the combined full model minus R^2 of the duration model, divided by R^2 of the duration model and multiplied by 100. **e**, Log likelihood difference between real and shuffled models for the same neurons fit by the phonetic and semantic embedding models (see Methods); Two-sided Wilcoxon signed-rank test, P = 8.54e-13. **f**, For each neuron, the difference in pseudo R-squared from a combined model with phonetic and semantic embeddings, and a separate model with either phonetic (green) or semantic (blue) embeddings. Semantic embeddings explain significantly more variance than phonetic embeddings. Two-sided Wilcoxon signed-rank test, P = 2.76e-27. **g**, The number of neurons for which each predictor was significant in the model is plotted on the y-axis, and the total number of neurons is also displayed on top of the bar plot. This model used semantic embeddings from Word2Vec, and models using BERT or GPT-2 embeddings had comparable results. **h**, The absolute value of the coefficient, or weight, of each predictor in the model is shown on the y-axis. Bars reflect the median predictor weight taken across neurons fit by the model. Predictors are sorted by decreasing weights. Poly = polysemy; Wps = word position in sentence; wpc = word position in clause; freq = word frequency; open and close represent opening and closing nodes; sem1-10 are semantic embeddings from Word2Vec.

https://doi.org/10.1038/s41593-026-02436-4



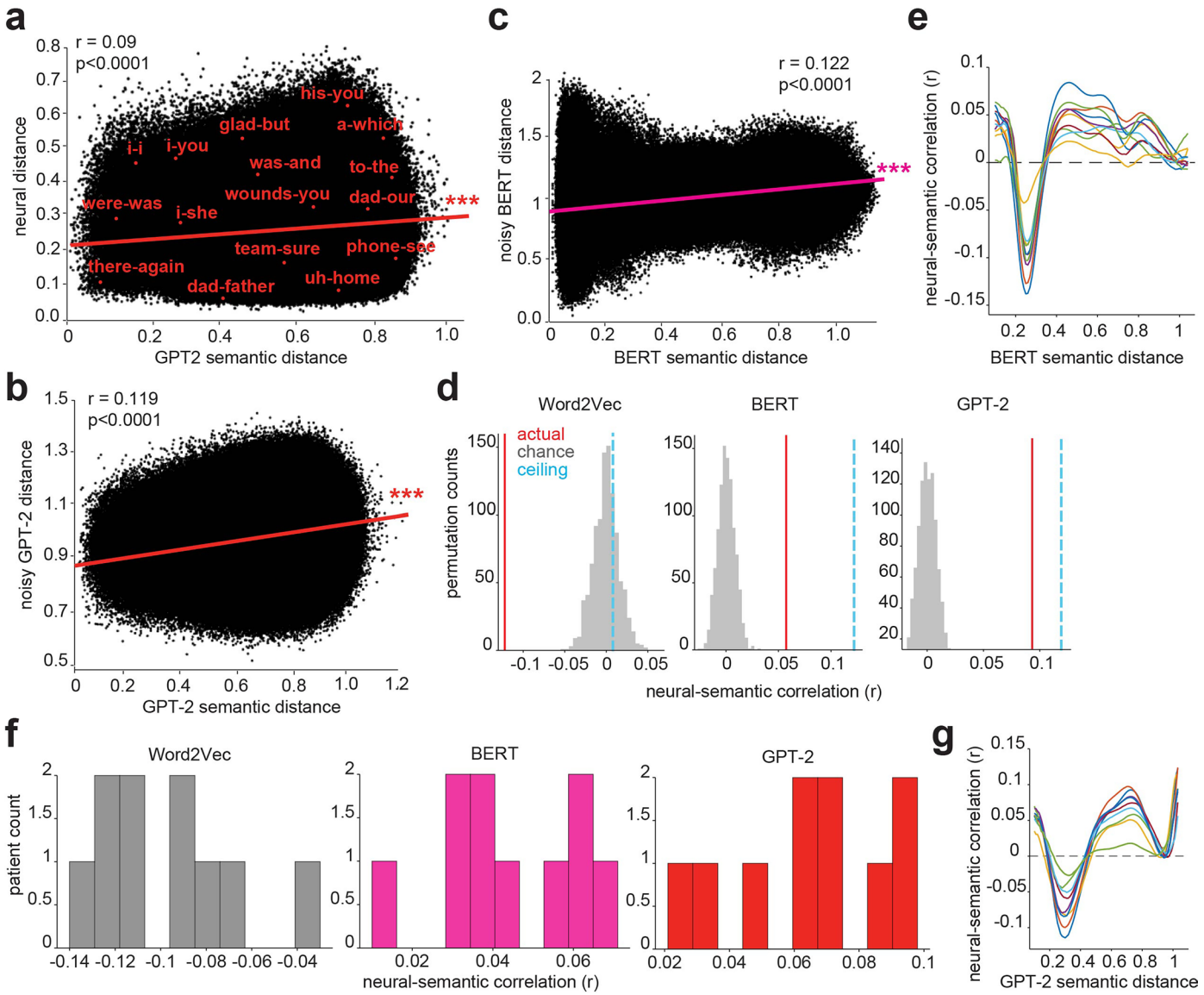
Extended Data Fig. 5 | Effects of function words and pronouns on semantic coding. a, The change in log-likelihood between the null and actual model for models fit with all words (blue) compared to the same measure in models fit without function words or pronouns (gray, FWE model). Delta LLH values were normalized by dividing each neuron's delta LLH value by the number of words used in the model. P = 0.0293, Two-sided Wilcoxon signed-rank test. **b**, The Pseudo R-squared values for the same neurons shown in A from models using all words or function words and pronouns removed. P = 9.97e-14, Two-sided Wilcoxon signed-rank test. **c**, The Two-sided Pearson correlation coefficient between neural and semantic distance for either all word pairs (blue) or word pairs excluding function words and pronouns (gray). All correlations are significant, P < 0.001. **d**, 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. Two-sided Pearson correlation, r = 0.4929, P = 4.93e-36.

https://doi.org/10.1038/s41593-026-02436-4



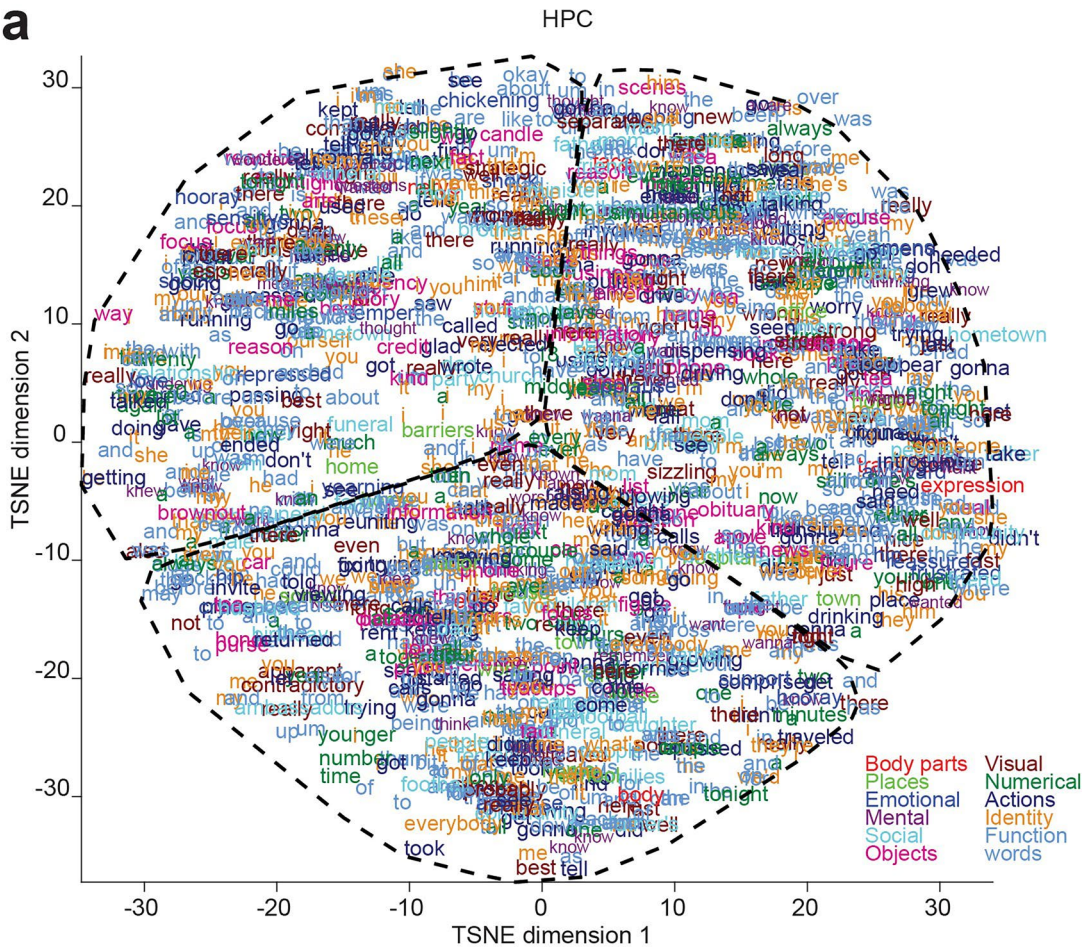
Extended Data Fig. 6 | Semantic categories from contextual embeddings. **a**, Words plotted in their dimensionality reduced BERT embedding space and colored according to their semantic category defined in Word2Vec. Words are grouped in 30 distinct categories defined from silhouette score and k-means clustering of BERT embeddings. Black dashed lines outline the identified categories. **b**, The same as in A but with GPT-2 layer 37 embeddings. 23 categories are identified. 1,808 words out of 7,346 are plotted here for visualization.

https://doi.org/10.1038/s41593-026-02436-4



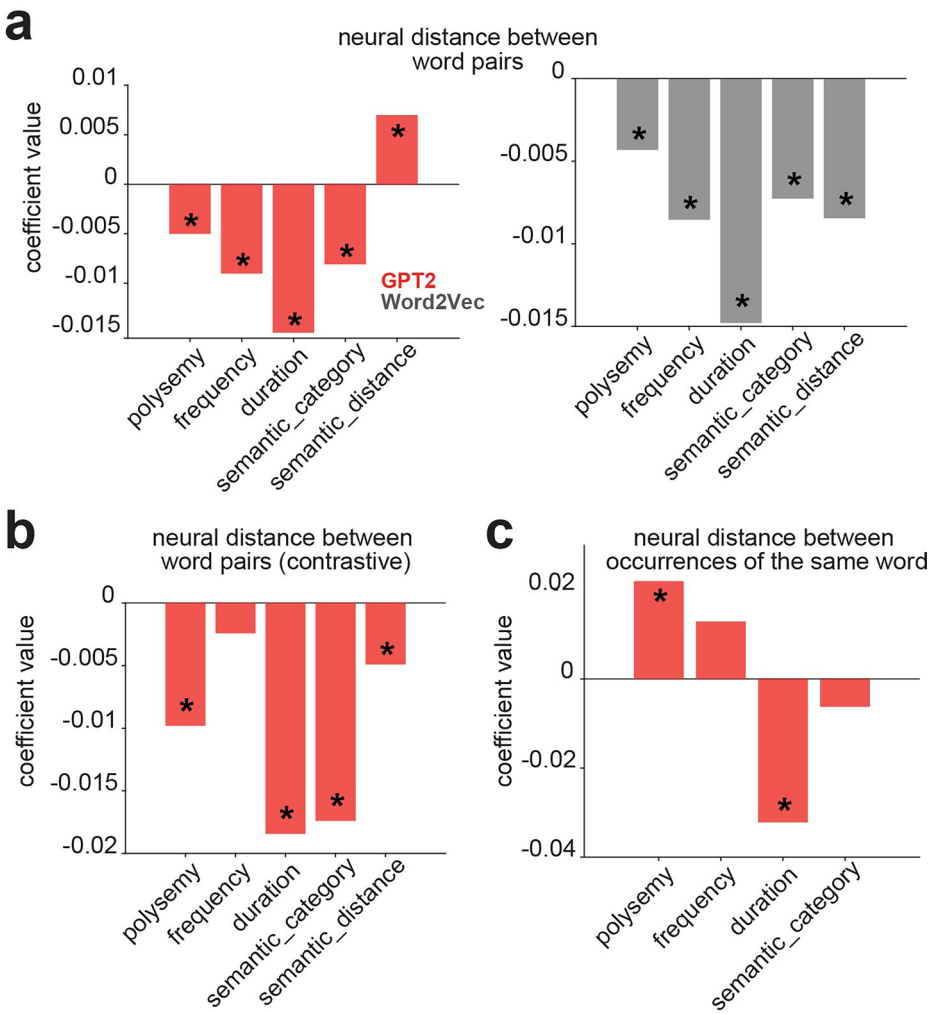
Extended Data Fig. 7 | The relationship between neural and semantic distance across language models, noise models, and individual patients. **a**, Scatter plot illustrating the correlation between neural and GPT-2 layer 37 semantic embedding distances between 26,154,528 word pairs. Two-sided Pearson correlation (r = 0.0932, 95% CI = [0.0928, 0.0935], t = 478.78, $P < 2.2 \times 10^{-308}$). Pink line: regression line. Each dot represents one randomly selected word pair; magenta text indicates example pairs. **b**, Scatter plot illustrating the correlation between GPT-2 layer 37 semantic embedding distances and GPT-2 embedding distances that were transformed with neural noise. Two-sided Pearson correlation (n = 26,154,528 word pairs, r = 0.1193, 95% CI = [0.1189, 0.1196], t = 614.61 $P < 2.2 \times 10^{-308}$). **c**, the same as in **b**, but with BERT embeddings from layer 13. Two-sided Pearson correlation (r = 0.1226, 95% CI = 0.1222, 0.1230], t = 631.42, $P < 2.2 \times 10^{-308}$). **d**, Plots for each language model show the distribution of Pearson correlations between neural and semantic distances obtained from shuffled neural activity (gray), the correlation from the real neural and semantic data (red), and the neural noise ceiling (blue dashed line), computed as the correlation between the semantic embeddings and their noise-injected versions. For BERT and GPT-2, the real neural correlation is well above chance and falls below - but close to - the noise ceiling, indicating that the models capture more than half of the explainable neural-semantic structure. In contrast, Word2Vec shows a negative correlation and a modest noise ceiling, consistent with its poor alignment with neural structure. **e**, For each patient (colored line), the semantic-neural correlation coefficient computed at increasing BERT semantic distances (binwidth 0.2, step size 0.01). While this line is generally positive in value, at especially short semantic distances (low x-values), it turns negative, reflecting contrastive coding in each patient. **f**, For each language model, histograms show the neural and semantic distance Pearson correlation using neural embeddings from each patient separately. **g**, Contrastive coding for each patient, as computed in panel **e** but using GPT-2 embeddings.

https://doi.org/10.1038/s41593-026-02436-4



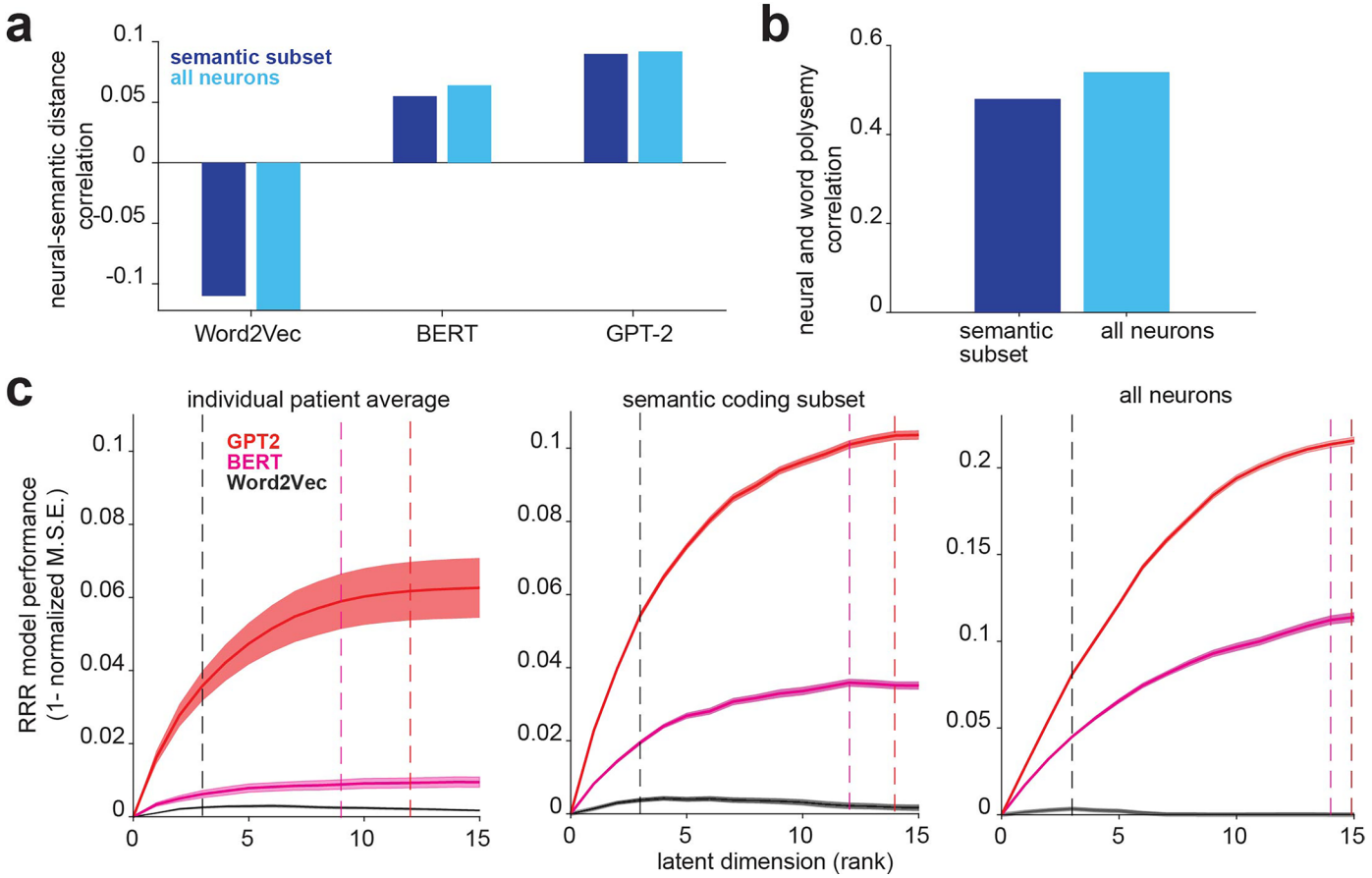
Extended Data Fig. 8 | Semantic categories from neural embeddings. a, Words plotted in their dimensionality reduced neural embedding space and colored according to their semantic category defined in Word2Vec. Words are grouped in 3 distinct categories defined from silhouette score and k-means clustering of neural embeddings. Black dashed lines outline the identified categories. 1,808 words out of 7,346 are plotted here for visualization.

https://doi.org/10.1038/s41593-026-02436-4



Extended Data Fig. 9 | The relationship between linguistic features and neural distance. a, Weights from a ridge regression model that predicted neural distance between all word pairs from the difference in polysemy, word frequency, word duration, and semantic category for each word pair, and the GPT-2 layer 37 (left) or Word2Vec (right) semantic distance between them. Asterisks represent significant predictors. **b**, The same as in A-left but restricted to word pairs with semantic distances within the contrastive coding range, 0.14 - 0.4. **c**, Weights from a ridge regression model that predicted, for every word, the median neural distance between all occurrences of that word from that word's polysemy (computed with GPT-2 layer 37 embeddings), frequency, duration, and semantic category. Asterisks represent significant predictors.

https://doi.org/10.1038/s41593-026-02436-4



Extended Data Fig. 10 | Comparison between semantic coding using the full population and a semantically selective subset of neurons. a, For semantic embeddings from Word2Vec, BERT layer 13, and GPT-2 layer 37, the Pearson correlation between neural and semantic distance using the full population of neurons (n = 356) or a subset of neurons whose residual activity was significantly predicted by semantic embeddings (n = 92). Two-sided Pearson correlation (semantic subset r = -0.1061, 0.0549, and 0.895 for word2vec, Bert, and GPT-2 respectively, all P < 2.2e-308). **b**, The Pearson correlation between word and neural polysemy, where neural polysemy was computed using a subset of semantically selective neurons or the full population. Two-sided Pearson correlation (semantic subset r = 0.4829, P = 1.36e-24). **c**, Reduced rank regression performance for predicting different language model embeddings from neural population activity using data from each patient individually (n ~ 30 neurons per patient), the subset of semantically selective neurons (n = 92) or the full population (n = 356). All regression models show the same trends in performance and optimal rank across the three language models.

nature portfolio

Corresponding author(s): Melissa Franch

Last updated by author(s): Jul 7, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Neural data was collected using Blackrock Neurotech recording system and analog audio channel inputs containing the speech from the podcast videos. Neural and audio data was recorded and synchronized in real time. We used Behnke-Fried stereoelectroencephalography (sEEG) probes (https://adtechmedical.com/epilepsy) and a Blackrock Neuroport system (https://blackrockneurotech.com/products/neuroport/) to record single units in the hippocampus synchronized to audio recordings.
Data analysis	Neural data was analyzed using programming and built-in functions in Matlab and Python. Audio data was analyzed transcribed in Praat, Montreal Forced Aligner and manual inspection. Spike sorting was performed with WaveClus (https://github.com/csn-le/wave_clus). Audio was transcribed with AssemblyAI (https://www.assemblyai.com/) and corrected manually using PRAAT (Boersma, 2011). Custom code written in MATLAB (R2024b) and Python (3.12.9) was used for analysis. Software packages included CoreNLP (https://stanfordnlp.github.io/CoreNLP/depparse.html), Wuggy (https://wuggycode.github.io/wuggy/), FOOOF (https://fooof-tools.github.io/fooof/), sklearn (https://scikit-learn.org/stable/), torch (https://pytorch.org/) and RAVE (https://github.com/rave-ieeg)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data availability
The data that support the findings of this study are available on the Data Archive for the Brain Initiative at this link: https://dabi.loni.usc.edu/projects/W2ZHP46ITR2D

Code availability
The code underlying this work relies on Python 3.11.7 and Matlab 2023b or higher. The custom code written to analyze the data and generate the figures is available on Github at: https://github.com/mcfranch/PopulationCodeForSemanticsHippocampus

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	the study contains data from 10 patients, 6 males and 4 females. Summary statistics are reported in the methods. Sex and gender were not considered in the analysis.
Reporting on race, ethnicity, or other socially relevant groupings	Races of the patients included White, Hispanic, and African American, however race, ethnicity, and other socially relevant groupings were not considered in the analysis.
Population characteristics	Summary statistics are reported in the methods. This feature was not considered in the analysis.
Recruitment	All patients diagnosed with drug resistant temporal lobe epilepsy and scheduled to undergo stereotactic encephalography from 2/1/2024 to 12/1/2025 were offered participation in our study. All patients provided written informed consent to participate in the study and were aware that participation was voluntary and would not affect their clinical course. There were no potential biases in the recruitment.
Ethics oversight	Experiments were conducted according to protocol guidelines approved by the Institutional Review Board for Baylor College of Medicine and Affiliated Hospitals, Houston TX (H-18112).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Patient recruitment was constrained by the availability of eligible neurosurgical cases and data were collected from all available participants during the study period. Data analysis was performed after completion of data collection. Within patients, units were included based on data quality, and only well-isolated neurons were analyzed. Sample sizes are consistent with, or larger than, most typically reported in human intracranial single-unit electrophysiology studies. Data were analyzed on an individual patient basis with replication across patients.
Data exclusions	All patients and all data were used for analysis.
Replication	10 or 5-fold cross validation was used to ensure the reproducibility of all classification and regression analyses. Key results were replicated within and across patients
Randomization	Patients were assigned to maximize the dataset across stimuli. No covariates would be expected between the tasks performed and patient

Randomization	specific features. Additionally, the majority of analyses were performed within patients, with replication across patients demonstrating generalization.
Blinding	Blinding was not possible due to the features of data collection and analysis. However, all methods were equally applied across patients and across neurons, so blinding is not relevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA