

Polysemanticity in human hippocampal neurons

Xinyuan Yan¹, Ji-An Li², Melissa Franch¹, Hanlin Zhu¹, Rhiannon Cowan³, James Belanger¹, Ana G. Chavez¹, Assia Chericoni¹, Taha Ismail¹, Kalman A. Katlowitz¹, Luca D. Kolibius⁴, Elizabeth A. Mickiewicz¹, Danika Paulo¹, Eleonora Bartoli¹, Jay A. Hennig^{1,5,6,7}, Tomasz M. Frączek¹, Nicole R. Provenza^{1,5,6,7}, Shervin Rahimpour³, Ben Shofty³, Elliot Smith³, Joshua Jacobs⁴, Benjamin Y. Hayden^{*†1,5,6,7}, and Sameer A. Sheth^{*1,5,6,7}

1. Department of Neurosurgery, Baylor College of Medicine
2. Department of Psychology, New York University
3. Department of Neurosurgery, University of Utah
4. Department of Neurology, University of Chicago
5. Department of Electrical and Computer Engineering, Rice University
6. Department of Bioengineering, Rice University
7. Neuroengineering Initiative, Rice University

† **Correspondence:** Benjamin.Hayden@bcm.edu

* These authors contributed equally

Funding statement

This research was supported by the McNair Foundation and by NIH R01 MH129439, U01 NS121472, UE5NS070694-15, the SNS Allan Friedman RUNN Research Grant, the NLM Training Program in Biomedical Informatics & Data Science for Predoctoral & Postdoctoral Fellows, T15LM007093-33, Gordon and Mary Cain Pediatric Neurology Research Foundation

Competing interests

S.A.S has consulting agreements with Boston Scientific, Zimmer Biomet, Koh Young, Abbott, and Neuropace. SAS is Co-founder of Motif Neurotech.

Acknowledgements

We thank Joshua Adkinson, Justin Fine, Victoria Gates, Suzanne Kemmer, George Kokalas, and Steven Piantadosi for their assistance.

ABSTRACT

To comprehend language, the brain must navigate a high-dimensional semantic landscape while seamlessly contextualizing meaning. Inspired by recent advances in the mechanistic interpretability of large language models (LLMs), we hypothesized that the brain utilizes polysemanticity, a coding strategy wherein individual neurons represent multiple semantically unrelated features through high-dimensional superposition (Elhage et al., 2022; Olah et al., 2020). We recorded single-unit activity from the human hippocampus during podcast listening. We found that hippocampal neurons exhibit dense semantic codes characterized by multiple tuning peaks with an overdispersed, isotropic geometry. This geometry satisfies the theoretical requirements for interference minimization in superimposed codes. Furthermore, semantic responses are strongly modulated by lexical and speaker-identity context; nonetheless, the underlying population geometry remains stable. This coding strategy permits rapid contextualization without requiring specialized, context-specific neurons. Indeed, we show clear pattern separation of similar terms, along with pattern completion for held-out words. Together, these results demonstrate that the human brain leverages superposition to solve a universal computational problem: maximizing semantic capacity within a constrained representational space.

INTRODUCTION

Just as neurons in the visual system are tuned for visual stimuli, neurons in the hippocampus are tuned for semantic features of the words we hear (Dijksterhuis et al., 2024; Franch et al., 2025; Katlowitz et al., 2025; Yan et al., 2025). A pioneering body of work shows that concept cells in the hippocampus and other medial temporal lobe (MTL) structures respond to stimuli that evoke particular concepts (Quiroga, 2012), either in an all-or-none (Quiroga et al., 2005) or graded (Karkowski et al., 2025) manner. However, such narrow selectivity does not readily lend itself to shading of meaning by context (contextualization) or contextual differentiation of otherwise similar stimuli (pattern separation, Bausch et al., 2026; Quiroga, 2020; Rey et al., 2025; Suthana et al., 2021), processes that are hallmarks of hippocampal function (Bakker et al., 2008; Cayco-Gajic & Silver, 2019; Leutgeb et al., 2007; Yassa & Stark, 2011). Moreover, contextualization is likely to be particularly important for language use, where it is the rule, rather than the exception (Caucheteux & King, 2022; Goldstein et al., 2022; Katlowitz et al., 2025; Kumar et al., 2024).

One of the hallmarks of LLMs, and a feature that is critical for their flexibility, is **polysemanticity**. Individual coding units, especially in internal layers, often respond to a range of seemingly unrelated stimuli (Bricken et al., 2023; Olah et al., 2020; Radford et al., 2019). Polysemanticity can be either disjunctive (*or-gating*) or conjunctive (*and-gating*), and shares many features with mixed selectivity (Rigotti et al., 2013; Tye et al., 2024). Polysemantic tuning is often a consequence of **superposition**, a strategy that allows a network to represent more independent features than it has available neurons by utilizing the high-dimensional geometry of neural state space (Elhage et al., 2022; Olah et al., 2020). Superposition, therefore, represents a solution to a coding problem imposed by language, in which the space of potential lexical meanings is vast. Polysemanticity has another benefit: it provides a substrate for contextualization. Because a neuron's firing can be conditioned upon specific co-occurring words, the local linguistic environment can act as a gate for its activity. This coding in turn facilitates pattern separation by projecting overlapping semantic inputs onto nearly orthogonal coding vectors and facilitates pattern completion by leveraging the learned statistical dependencies between co-occurring features (cf. Babadi & Sompolinsky, 2014; Cayco-Gajic & Silver, 2019). We therefore hypothesized that hippocampal semantic coding during language listening would exhibit polysemanticity.

The hypothesis that the human hippocampus utilizes polysemantic coding is consistent with its spatial codes. While classical hippocampal place cells are typically thought of as having single stable firing fields, recordings in functionally large environments reveal multiple, spatially unrelated firing fields that lack obvious topographical relationships (Eliav et al., 2021; Fenton et al., 2008; Rich et al., 2014). Similarly, entorhinal grid cells represent space through repeating, multi-peaked firing matrices (Hafting et al., 2005). However, this periodic structure transitions into dense, non-hexagonal multi-field packing in higher-dimensional (3-D) or non-uniform environments (Ginosar et al., 2021; Grieves et al., 2021; Hayman et al., 2011). (Recall that semantic space is high-dimensional). Indeed, recent work demonstrates that many of the same coding principles used for place and grid cells apply to abstract concepts as well (Constantinescu et al., 2016; Knudsen & Wallis, 2021; Park et al., 2021). These findings in turn raise the possibility that semantic coding has the dispersed multi-field properties observed in spatial navigation.

RESULTS

We collected responses of isolated single neurons in the hippocampus as patients listened to 47 minutes of speech (n=498 neurons, 15 patients; **Methods, Figure 1A**; Franch et al., 2025). All analyses were performed only on well-isolated single units, identified manually by trained lab members (e.g., **Figure 1B/C**). We defined word-evoked firing rates as beginning 80 ms after the onset of each word and lasting the duration of the word plus 40 ms. (All results shown below are robust to specific analysis windows used).

Hippocampal semantic coding during language listening is distributed

We quantified semantic tuning with ridge regression (Caucheteux & King, 2022; Jamali et al., 2024; Zada et al., 2025), with performance defined as cross-validated log-likelihood improvement over an intercept-only baseline (ΔLL ; **Methods**). As predictors, we used the first 100 principal components of GPT-2 embeddings (Radford et al., 2019). Due to our interest in the encoding of word concepts, we analyzed only nouns. Layer 36 yielded stronger encoding performance than Layers 24 or fastText (word embeddings, Joulin et al., 2016). Layer 36 was therefore used for all subsequent analyses (**Figure 1D**). To quantify population encoding we decoded single words from the responses of all neurons (Tang et al., 2023)(**Figure 1E**). Many words were reconstructed perfectly; nearly all others were close (**Figure 1F, right**).

We asked whether the distances between words in neural population activity mirror the relational codes in LLM embedding space using representational similarity analysis (RSA; **Figure 1G**) (Nikolaus Kriegeskorte et al., 2008; Nili et al., 2014). Indeed, the Spearman rank correlation between neural and LLM distance matrices is significantly above chance across the population (**Figure 1H-I**). This finding extends our previously published results by showing that this correspondence monotonically increases with population size (**Figure 1J-M**; Franch et al., 2025). This scaling behavior suggests the semantic encoding is distributed (Kafashan et al., 2021) because it implies that individual neurons carry partial, overlapping information about many words that accumulate across the population. To confirm this intuition, we repeated the RSA decimation analysis on simulated concept cell populations (see below and **Figure S2, Methods**). Unlike real neurons, simulated concept cells showed no monotonically increasing RSA with population size and produced near-zero correlations with GPT-2 semantic geometry at all sparseness levels (**Figure S3**). To rule out the possibility that the scaling merely reflects noise

averaging across neurons with identical tuning, we constructed a control population by copying the single best-encoding neuron 498 times and adding independent noise matched to the empirical residuals (**Methods**). This copy-plus-noise population plateaued at moderate N , whereas the real population showed continued representational gains at large N , confirming that additional real neurons indeed contribute to semantic geometry (**Figure S1C**).

Lack of discrete functional categories of neurons with similar tuning

Response patterns of monosemantic neurons should roughly recapitulate the structure of semantic space, meaning that single neurons' responses should be semantically adjacent. This is true even for neurons that embody more abstract categories that cross basic animal/place/thing category lines, such as “Christmastime” (which would include reindeer, the North Pole, and candy canes). In either case a neuron's set of driving stimuli will necessarily have greater proximity than chance in some dimensions of embedding space. A polysemantic code, by contrast, predicts no such categorical grouping, indeed, grouping may even be overdispersed.

To quantify the internal coding structure within the population, we applied symmetric non-negative matrix factorization (NMF; Lee & Seung, 1999) to the neuron-by-neuron tuning similarity matrix (**Figure 1N**). NMF decomposes this similarity matrix into discrete components, each reflecting a distinct linear reweighting of the entire neural population. It produces neuronal “modules,” each representing a functionally coherent structure within the space of possible population tuning functions. We then ask whether these neurally-derived modules align with the structure of semantic space as determined by word models (fastText; **Methods**). This analysis revealed thirteen modules; this number was selected by computing reconstruction error and identifying the point of greatest curvature change in the error curve. We confirmed module validity by comparing its reconstruction error against a null distribution (permutation, $p < 0.05$). We defined semantic categories by performing k-means clustering on the fastText word embeddings to define discrete semantic categories (**Methods**). This method will find any categories that are latent in the training corpus, including both natural categories (e.g., animals) or abstract ones (e.g., “things you might encounter at the beach”). We found that NMF-driven modules do not map onto single fastText-derived semantic categories. Instead, each module recruits words from multiple semantic categories (**Figure 1O**; e.g., Modules 7 and 11 each span six different semantic categories).

To quantify this observation, we compared the mean pairwise semantic distance among each module's top-scoring words (top 100 words) against a null distribution constructed by randomly sampling the same number of words (500 permutations). Only one of thirteen modules showed significantly clustered semantics (module 3, $z=-2.03$, $p=0.020$); eight modules were indistinguishable from chance, and four were significantly more dispersed than chance (all $p<0.036$), meaning their top words were drawn from more distant regions of semantic space than expected by chance. This specific distribution is biased towards overdispersion, a pattern we explore in more detail below ($p<0.01$, binomial test). This pattern is replicated at the individual patient level (**Table S1**).

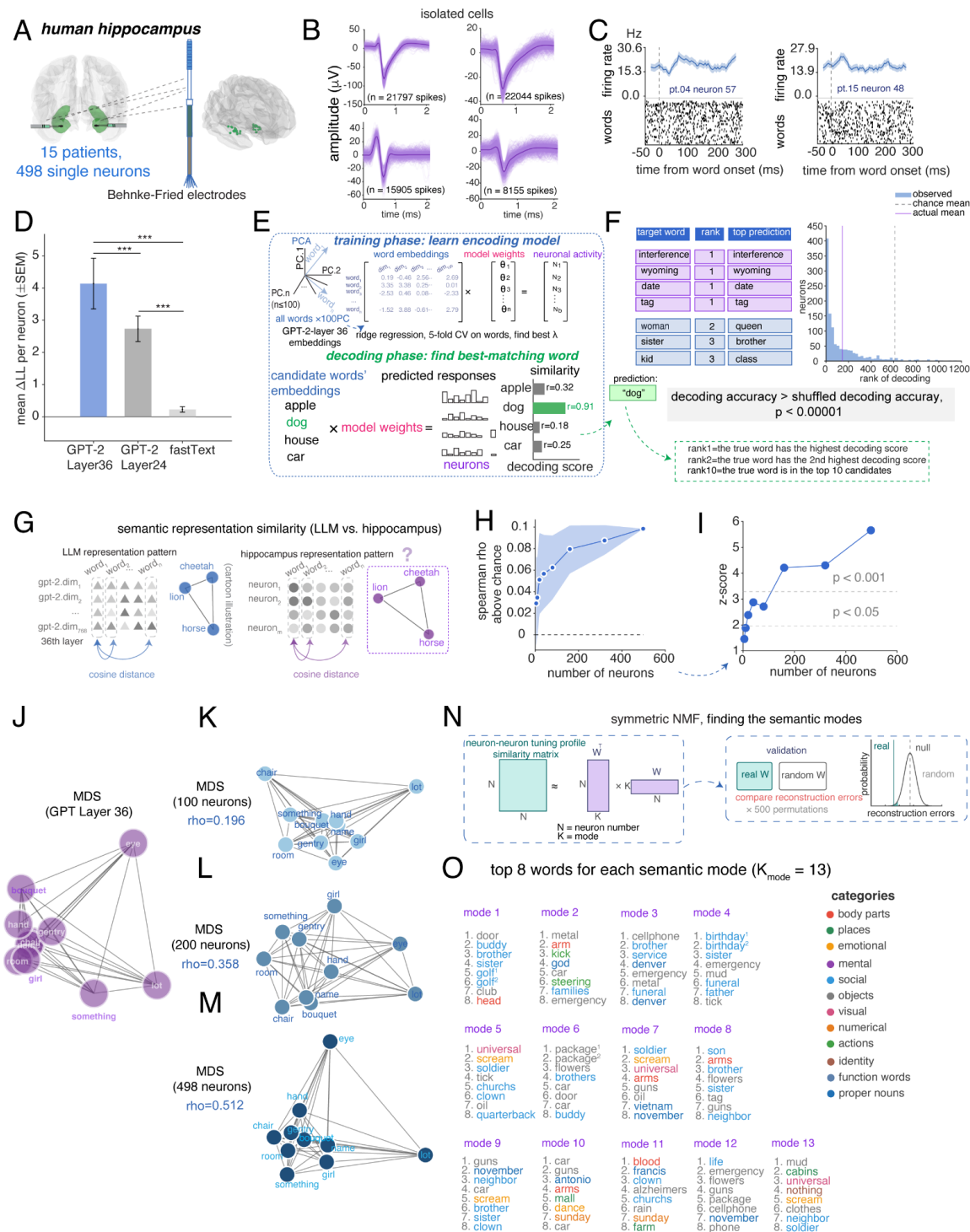


Figure 1. Hippocampal population recapitulates semantic embedding geometry.

(A) Recording setup. Single-neuron recordings were obtained with Behnke-Fried microelectrodes in the hippocampus.

(B) All analyses used well-isolated single units; four example cells shown.

(C) Example PSTHs from two neurons showing firing rate modulation across words.

- (D) Encoding model performance. Cross-validated Δ LL per neuron (mean $\pm$ SEM, $n = 498$, Method). GPT-2 Layer 36 outperformed Layer 24 and fastText (paired t-tests, FDR-corrected; $***p < 0.001$). Control analyses in **Figure S1A-B**.
- (E) Encoding-decoding framework. Top: ridge regression maps GPT-2 Layer-36 embeddings (100 PCs) to firing rates. Bottom: for each held-out word, the model ranks candidates by correlation with the observed population response (Methods).
- (F) Decoding performance. Left: examples of correct (purple) and semantically close (blue) recoveries. Right: recovery rank histogram is highly left-biased.
- (G) Representational similarity analysis (RSA) cartoon illustration.
- (H) RSA correlation scales monotonically with population size and does not saturate at $N = 498$. Shaded band: 95% CI of neuron-resampling distribution (Methods). Control analysis in **Figure S1C**.
- (I) Effect size of RSA at each population size.
- (J-M) Multidimensional Scaling (MDS) visualizations of semantic geometry. Multidimensional scaling plots are shown for (J) the GPT-2 Layer-36 reference embedding, and neural populations of (K) 100, (L) 200, and (M) 498 neurons. Spearman rho between neural and LLM distance matrices increases.
- (N) Symmetric NMF (Non-negative matrix factorization) schematic (Methods).
- (O) Neural modules span multiple word categories. The top 8 words for each of the 13 neural modules are shown (Methods). We colored the top 8 words in each module by their semantic categories (Methods).

Non-sparse semantic coding during language listening

We next characterize the coding density of individual neurons in four different ways.

Population sparseness quantifies how selectively a given word activates the neural population (**Figure 2A**; Rolls & Treves, 1997). Values near 1 indicate that only a single neuron responds, while values near 0 indicate uniform activation. We find population sparseness is moderate (mean $\pm$ std=0.581 $\pm$ 0.031). We compared this result to data generated from simulated concept cells that have metrics matching those of real concept cells (Waydo et al., 2006; **Figure S2A-E**; see **Methods**). Neurons in our dataset use a much denser code than the simulated concept cells (**Figure 2B**). Because Waydo et al. argue for a range of plausible sparseness parameters (0.20% to 1.0%), we confirmed this result using ten sparseness levels in this range (Wilcoxon rank-sum test, all $p < 0.001$).

Lifetime sparseness quantifies how selectively a given neuron responds across words (**Figure 2C**): values near 1 indicate firing to a single word, while values near 0 indicate uniform responding (Rolls & Treves, 1997). Lifetime sparseness of neurons in the hippocampus (mean $\pm$ std=0.369 $\pm$ 0.305) is considerably denser than that of the simulated concept cells (sparseness range: 0.20% to 1.0%, Wilcoxon rank-sum test, all $p < 0.001$, **Figure 2D**). We found 10 neurons with extreme high lifetime sparseness (i.e., responding to very few words) which met

the criteria for defining concept cells, but such neurons responded to several semantically unrelated words (**Figure S8**), suggesting that their sparse firing reflects polysemantic tuning rather than narrow concept selectivity.

Normalized Shannon entropy measures how uniformly a neuron distributes its firing across words (0 = maximally sparse, responding to a single word; 1 = maximally dense, responding uniformly to all words, **Methods**, Lehky et al., 2005). Neurons in our sample show relatively high entropy, that is, closer to responding to all words uniformly (mean $\pm$ std = 0.910 $\pm$ 0.130). Simulated concept cells produced much lower entropy values (sparseness range: 0.20% to 1.0%, Wilcoxon rank-sum test, all $p < 0.001$, **Figure 2E**).

A neuron could have high entropy (responds to many words) but carry very little information if its rate variations do not reliably distinguish between words. In other words, apparent density could be a by-product of high noise or variability. To test for this possibility, we computed the *information per word* (**Figure 2F**; Skaggs, McNaughton, et al., 1992). Neurons in our dataset convey substantially more information per word than simulated concept cells (**Methods**; Wilcoxon rank-sum test, all $p < 0.001$; **Figure 2F**). This result confirms that the broad tuning (high entropy) we observe is functionally meaningful and supports fine-grained discrimination among words. We replicated these results using narrower time windows, to control for the possibility that early responses to the next word may cause spurious coding density (**Figure S9A-D**).

Dense coding does not necessarily imply polysemanticity: a neuron could respond broadly within a single semantic category. We next asked whether the words that drive each neuron are semantically clustered.

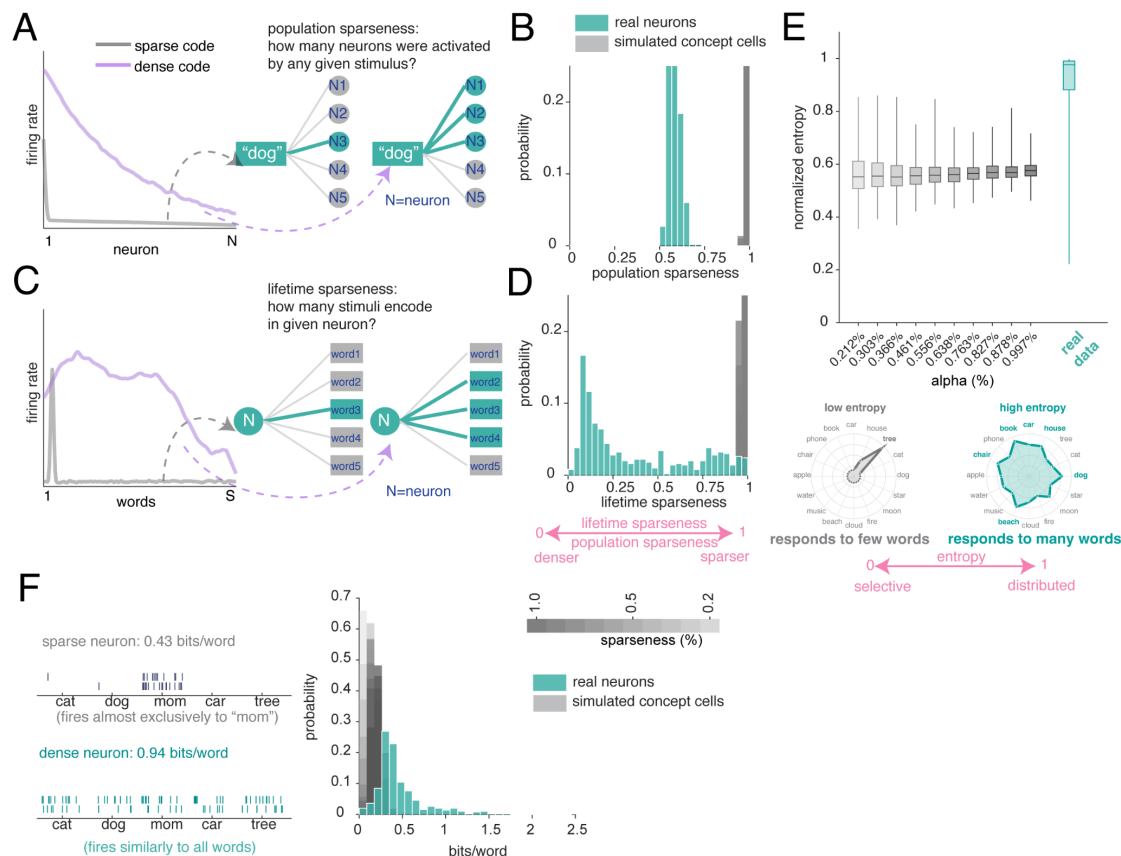


Figure 2. Hippocampal neurons use dense code to encode concepts.

(A) Cartoon illustration of population sparseness: under a sparse code (gray), a stimulus activates few neurons; under a dense code (purple), it activates many.

(B) Real neurons (teal) exhibit substantially lower population sparseness than simulated concept cells (gray). Simulation details in **Figure S2**.

(C) Cartoon illustration of lifetime sparseness: under a sparse code, a neuron responds to few stimuli; under a dense code, it responds to many.

(D) Real neurons exhibit substantially lower lifetime sparseness than simulated concept cells.

(E) Normalized Shannon entropy (0 = maximally selective; 1 = maximally uniform) across simulated sparseness levels, with real neurons (teal, far right) clustering near the high-entropy end. Bottom: radar plots contrasting a low-entropy neuron (left), which concentrates responses on few words, with a high-entropy neuron (right), which distributes responses broadly.

(F) Information per word (bits/word; Skaggs et al., 1992, Methods). Left: cartoon contrasting a sparse neuron firing almost exclusively to one word (0.43 bits/word) with a dense neuron firing broadly (0.94 bits/word), illustrating that broadly tuned neurons can carry more information. Right: real neurons (teal) carry substantially more information per word than simulated concept cells (gray).

Hippocampal neurons are polysemantic

One way to visualize a neuron's tuning function is to look at its predicted responses to a large corpus; we used a standard 50,000 noun corpus (Joulin et al., 2016). For illustration, we show the predicted response of three example neurons to these words (**Methods**). These show

complex patterns of activation, with responses to a wide range of stimuli along with local clusters of particularly strong activation (which we call “modes”, **Figure 3A**). For comparison, we calculated estimated responses of simulated concept cells (see above and **Methods**), which have the expected narrow peaks corresponding to their preferred concepts (**Figure 3B**).

We identified individual modes using DBSCAN clustering of high-firing words in t-SNE 2D semantic plane (cf. Cao et al., 2025; **Methods**), where density-based clustering is more effective than in high-dimensional spaces due to distance concentration (Aggarwal et al., 2001). For each mode, we then computed its centroid as the firing-rate-weighted mean of all the words within that mode in the original high-dimensional space (**Methods**). All subsequent geometric analyses (except the gridness score, see below) are based on the mode centroids in the original embedding space, rather than the reduced t-SNE space (**Methods**).

We assessed the regularity of inter-mode spacing by computing the coefficient of variation (CV) of all pairwise distances between mode centroids (**Figure 3C**). The population showed significantly more *regular* spacing than chance (mean $z \pm SD = -7.274 \pm 3.063$, one-sample t-test vs. 0: $p < 0.001$). We then tested whether nearest-neighbor distances between modes are uniform (**Figure 3D**). The population showed significantly more *equidistant* nearest neighbor spacing than chance (mean $z \pm SD = -8.130 \pm 3.582$, $p < 0.001$). We then examined the directional arrangement of modes by computing the CV of pairwise cosine similarities among direction vectors from the grand centroid to each mode centroid in the high-dimensional embedding space (**Figure 3E**). The population was significantly more *isotropic* than chance (mean $z \pm SD = -0.200 \pm 0.100$, $p < 0.001$), indicating that modes were more likely distributed uniformly across all directions in semantic space rather than concentrated along a few preferred directions, as grid cells would. Indeed, we found no evidence of hexagonal symmetry in the hippocampal population’s semantic tuning maps (**Figure 3F**). Specifically, *gridness* scores (Hafting et al., 2005; Langston et al., 2010) were negative (mean $z = -1.127 \pm 0.874$, one-sample t-test vs. 0: $p < 0.001$). In summary, hippocampal semantic representations are organized into regularly spaced, approximately equidistant, and isotropically distributed modes, without the periodic hexagonal symmetry characteristic of grid cells in two dimensions (equidistant and isotropic arrangements in high-dimensional spaces do not necessarily suggest hexagonal symmetry in two dimensions). However, this structure is reminiscent of grid cell geometry in

high (specifically, three) dimensional and non-uniform environments (Ginosar et al., 2021, 2023; Grieves et al., 2021; Krupic et al., 2015).

Note that PCA produces dense components, while ridge regression penalizes large weights without eliminating them. It is possible therefore that this pipeline may artifactually favor distributed codes by construction. To test for this possibility, we conducted a control analysis using a sparse-dominated alternative: non-negative matrix factorization (NMF, Lee & Seung, 1999) for embedding decomposition paired with elastic net regression (L1 ratio=0.9, see **Methods**), which promotes sparsity by forcing the coefficients of less relevant features to exactly zero. None of these results were affected by the choice of encoding model.

The isotropic geometric feature suggests that neuronal responses may be more semantically dispersed (*overdispersion*) than if their optimal driving stimuli were random (example neuron, **Figure 3G**). To test for overdispersion, for each neuron we identified the top N preferred nouns (by firing rate), computed their mean pairwise cosine distance in GPT-2 embedding space, and compared this distribution against a null distribution constructed from random nouns (**Figure 3H-I**). Across the population, the proportion of neurons whose preferred nouns are significantly more dispersed than chance (specifically, across a range of Ns, 14.4%-26.5%) substantially exceeds both the 5% false positive rate and the proportion showing significant clustering (i.e., across a range of Ns, 3.0%-2.5%). This pattern is not an artifact of the N we chose, as it holds across a range of Ns (specifically, all values of N from 10 to 100, **Figure 3J**). Quantitatively, the mean pairwise distance among each neuron's top N words is significantly greater than the null expectation at every top N tested (all $t > 8.400$, all $p < 0.001$; **Figure 3K**). This result can be replicated at the patient level (**Figure S4**).

One possible alternative explanation for these findings is that semantic overdispersion among each neuron's preferred words might be driven by noise in the tail of the firing rate distribution rather than by the neuron's strongest responses. To test for this possibility, we next performed a firing-rate-weighted control analysis (**Methods**). Each word pair is weighted by the product of their firing rates to ensure that the highest-firing words dominate the metric. This control analysis showed nearly identical results (all $t > 7.870$, all $p < 0.001$; **Figure S5**). We replicated these results using narrower time windows to control for the possibility that early responses to the next word may cause spurious polysemanticity (**Figure S9E-F**).

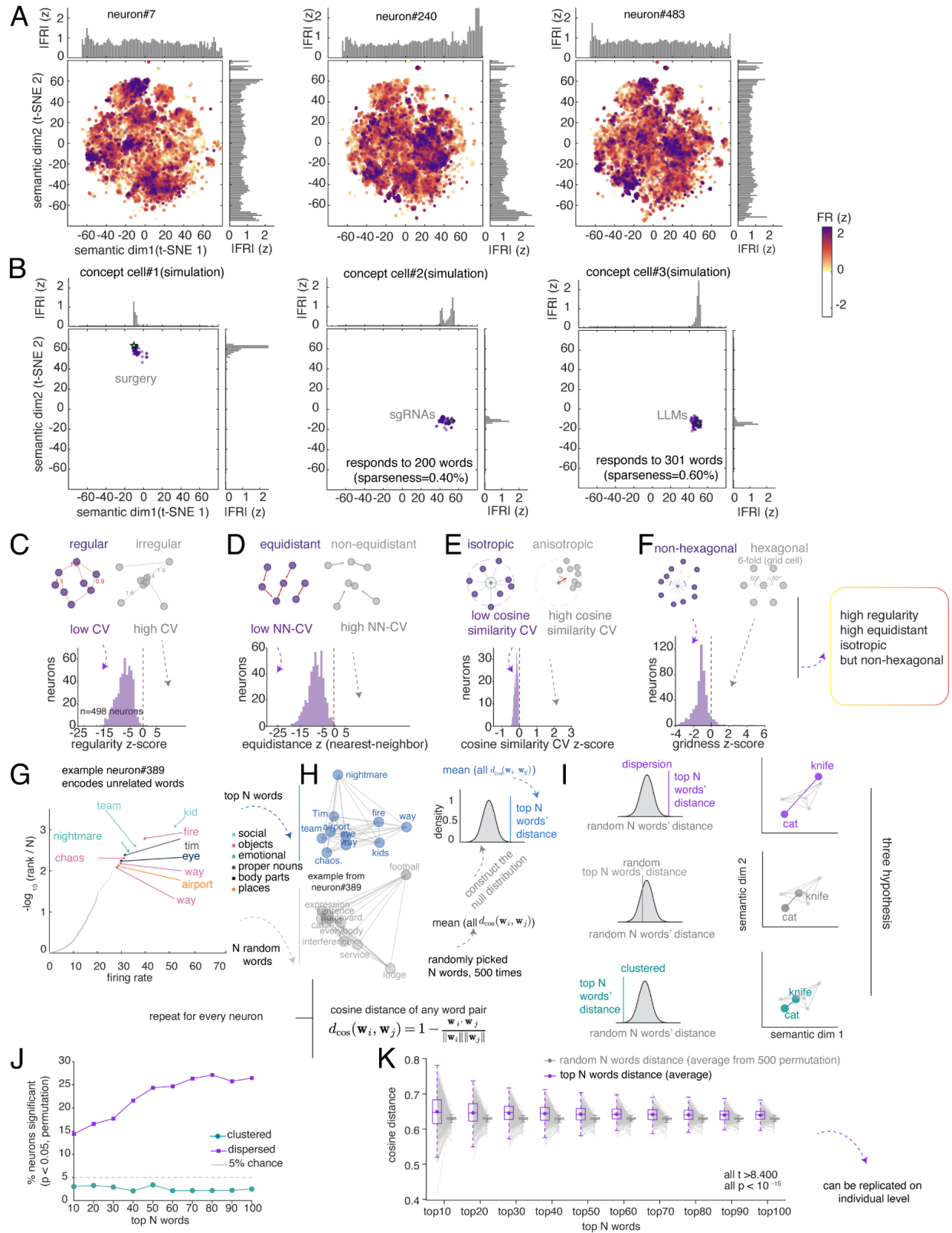


Figure 3. Single-neuron tuning landscapes are multi-moded and semantically overdispersed.

- (A) Single-neuron tuning landscapes. Predicted firing rates (z-scored) plotted on a t-SNE embedding of 50,000 nouns from an independent corpus (Methods). Three example neurons shown; dots colored by predicted firing rate (warm = high, white = low/zero); marginal histograms show mean |FR| per bin.
- (B) Simulated concept cells for comparison. Three concept cells (see Methods and Figure S2 for simulation details) at varying sparseness levels, displayed in the same t-SNE space.
- (C) Regularity. Top: schematic of regular vs. irregular arrangements. Bottom: z-score distribution shows tuning fields are more regularly spaced than chance.
- (D) Equidistance. Top: schematic. Bottom: nearest-neighbor distances are more uniform than chance.
- (E) Isotropy. Top: schematic. Bottom: tuning fields are approximately isotropically distributed.
- (F) Hexagonal symmetry. Top: schematic. Bottom: no evidence of six-fold symmetry, ruling out grid-like periodic tiling.
- (G) Example neuron (#389) encodes semantically unrelated words. Words ranked by firing rate with top 10 labeled by semantic category, spanning multiple unrelated categories.
- (H) Dispersion analysis schematic. Mean pairwise cosine distance among each neuron's top N words compared against a null from randomly sampled words (Methods).
- (I) Schematic of possible outcomes: dispersed, random, or clustered.
- (J) Proportion of significantly dispersed neurons exceeds both chance (5%) and clustered neurons at all N values.
- (K) Top-N words are significantly more dispersed than random words at every N (all $t > 8.4$, $p < 0.001$). Control analyses see **Figure S4** and **Figure S5**.

Context reshapes single-neuron tuning while preserving population geometry

We next asked whether hippocampal semantic tuning functions are modulated by speaker-identity context (i.e., story context; **Figure 4A**). For each neuron, we computed split-half tuning vectors from random halves (repeated 50 times with different random splits) of the same story and from different stories. Across-story cosine distances were greater than within-story distances (paired t-test, $t(497)=19.813$, $p<0.001$; **Figure 4B**), indicating that individual neurons adopt different semantic tuning depending on story. This change was individually significant in all fifteen patients (**Figure 4C**). This effect is not explainable by representational drift, as across-story tuning distance did not correlate with temporal separation between stories (Spearman $\rho=-0.074$, $p=0.794$; **Figure S6**).

Despite these context-dependent shifts in individual neurons' tuning, the population-level geometric arrangement of word representations was preserved across stories. For each pair of stories ($\binom{6}{2}=15$ pairs), we identified nouns that appeared in both stories and computed the pairwise distance (representational dissimilarity matrix, RDM) of these shared words separately within each story (**Figure 4D**; **Methods**). The correlation between story-specific RDMs was significantly higher than null distributions constructed from mismatched random words across all 15 story pairs (observed $r=0.206 \pm 0.126$, null $r = -0.001 \pm 0.010$; Wilcoxon signed-rank,

$p < 0.001$) (**Figure 4E**). In other words, sets of nouns maintain similar geometric relationships to each other regardless of their story context, even as neurons shift their tuning across stories.

We next examined how local lexical context affects semantic tuning (**Figure 4F**). We compared three nested linear models (**Figure 4G; Methods**): M1 predicted firing rates from static fastText word embeddings alone (word identity, no context); M2 added contextual GPT-2 embeddings of the preceding 1-20 words (i.e., context embeddings), allowing context to independently shift firing rates without changing how the neurons responds to word meaning; M3 further included the word-by-context interaction, allowing the context to directly modulate the way that neurons encode word meaning. To avoid arbitrarily selecting a single context window, we varied the number of preceding words from 1 to 20 and tested each window size independently. Model comparison using adjusted R^2 (which penalizes model complexity) revealed that M2 significantly outperformed M1 across all 20 context window sizes (all $p < 0.001$, Cohen's d range=[0.257, 0.363], **Figure S7A**), and M3 further improved over M2 at 19 of 20 window sizes (all $p < 0.05$ except window 1; Cohen's d range=[0.129, 0.220] for significant windows, **Figure S7A**), confirming that both context and its nonlinear interaction with word identity contribute explanatory variance beyond word identity alone (**Figure 4H**).

We further classified neurons by which predictors reached significance (**Figure 4I**). Under the additive model (M2), neurons are divided into word-only ($n = 57$), context-only ($n = 114$), and word-plus-context ($n = 22$) groups, with 305 neurons unclassified. When the interaction term was included (M3), a substantial population of interaction-specific neurons emerged ($n = 124/498$, 24.9%). These interaction-only (WxC) neurons are all from the unclassified neurons in M2, meaning that linear regression would define these as unmodified by either semantics or context. In total, 207 neurons (41.6%) involved the interaction term in M3, drawn from every category in M2 (**Figure 4J**). Together, these results demonstrate that apparent main effects of word identity and context in M2 are better explained as a word-by-context interaction in M3. In other words, local lexical context does not simply shift firing rates up or down in a gain-like process, but changes *how* neurons encode word meaning.

We next examined the relative magnitude of each predictor. We found that context embeddings (the GPT-2 representation of the preceding words) carried significantly larger standardized regression coefficients than word identity in M2 and M3 across all window sizes (all $p < 0.05$; **Figure S7B-C**). In M3, the interaction term was comparable to or larger than word

identity (**Figure S7D**) but consistently smaller than context main effects (**Figure S7E**, all $p < 0.05$), yielding a magnitude hierarchy of $|\beta_C| > |\beta_{W \times C}| \geq |\beta_W|$ (**Figure 4K**). Thus, local lexical context is the dominant driver of hippocampal firing rate variance, while its nonlinear modulation constitutes a secondary but substantial influence that exceeds the contribution of static word identity alone.

To assess the prevalence of nonlinear interaction effects, we tested all possible two-way ($W_i \times C_j$; 100 terms) and three-way ($W_i \times W_j \times C_k$ and $W_i \times C_j \times C_k$; 900 terms) interaction terms constructed from the first 10 principal components of the word (explaining 20.2% variance) and context embedding spaces (explaining 25.54% variance), yielding 1000 interaction models per neuron. For each neuron, we fit a simple regression of firing rate onto each interaction term independently and assessed significance at $\alpha = 0.05$. Under the null hypothesis of no interaction, the number of significant terms per neuron follows a binomial (1000, 0.05) distribution, with an expected value of 50 and a one-tailed significance threshold of 63. Across context window sizes 1-20, interactions per neuron ranged from 59.1 to 65.0, consistently exceeding chance. Overall, 39.8%-47.8% of neurons (198-238 out of 498 neurons) exhibited more interaction than chance (binomial test, $p < 0.05$; proportion stable across context window sizes).

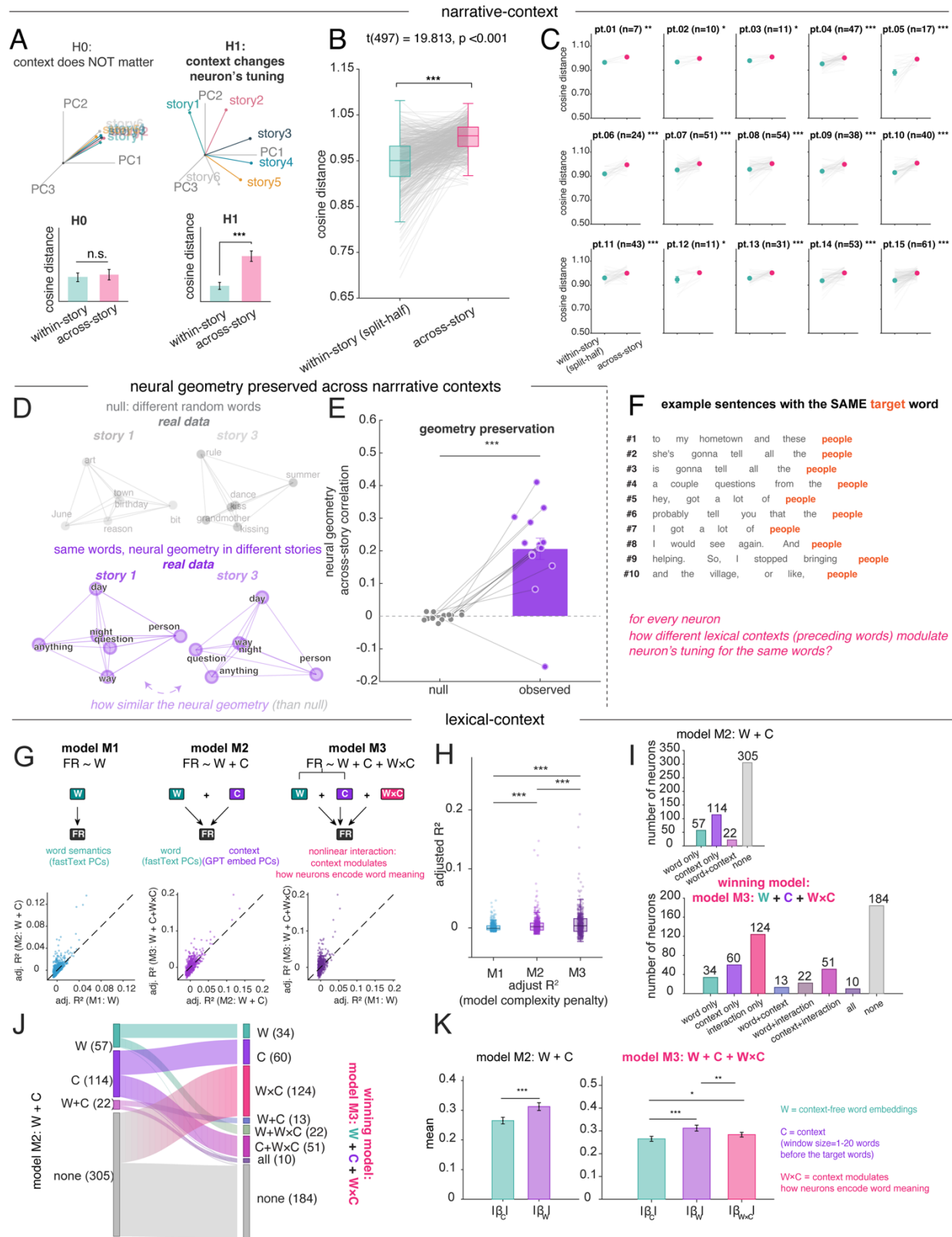


Figure 4. Story context and lexical context modulate semantic representations.

- (A) Static vs. contextualized hypotheses for semantic tuning across stories. Control analysis in **Figure S6**.
- (B) Semantic tuning is less stable across stories than within stories (story context), indicating context-dependent tuning profiles.
- (C) The story context modulating semantic tuning effect replicates across all 15 patients individually.
- (D) Neural geometry of shared nouns preserved across stories. For each story pair, RDMS of shared nouns are computed separately and compared via Spearman correlation. Null uses non-matching words (Methods).
- (E) RSA correlation for shared words (purple) exceeds the null (gray) across all 15 story pairs, demonstrating preserved geometric relationships regardless of story context.
- (F) The same target word (“people”) appears in 10 different lexical contexts across the podcast.
- (G) Model comparison framework. M1: firing rate $\sim$ word embeddings (W). M2: $\sim$ W + context embeddings (C). M3: $\sim$ W + C + W $\times$ C interaction. Scatter plots show neuron-by-neuron adjusted R^2 comparisons.
- (H) M3 is the winning model, confirming that both context and the word-by-context interaction contribute explanatory variance.
- (I) Neuron classification by significant predictors. M2: 57 neurons with significant word effects only, 114 with significant context effects only, 22 with both significant, 305 with neither. M3: 124 interaction-specific neurons emerge from the previously unclassified group; 207 total (41.6%) show significant interaction effects.
- (J) Sankey plot showing that interaction-only neurons in M3 derive entirely from unclassified neurons in M2.
- (K) Magnitude hierarchy: $|\beta_C| > |\beta_{W \times C}| \geq |\beta_W|$, stable across context window sizes (**Figure S7**).

Notes: Panels G, H, K show results for context window size = 20 words. Other window sizes in Figure S7.

Polysemanticity codes enable pattern separation and pattern completion

Pattern separation is the process by which the hippocampus transforms overlapping inputs into distinct, non-interfering neural representations (Yassa & Stark, 2011). We next asked whether and how hippocampus maps *identical* lexical inputs onto *distinct* neural representations when they occur in different semantic contexts (**Figure 5A-B**). Identical word inputs should evoke distinct population responses in different contexts, and the magnitude of this shift should scale with the distance between those contexts (Ethayarajh, 2019). For each noun appearing at least five times across the podcast, we computed pairwise distances among its occurrences in lexical context space (defined as the mean population firing rate vector across the 1-20 words immediately preceding each occurrence) and target word space (population response to the noun itself; **Figure 5C**). Across the population, observed context-repeated word geometry coupling (averaged across all repeated words with occurrence ≥ 5) significantly exceeded a random-context null baseline across all context window sizes (1 to 20 preceding words used to define the context vector; all $p < 0.05$, **Figure 5D; Methods**). This finding goes beyond showing that context shifts responses (**Figure 4D-E**): it demonstrates that the population maps contextual dissimilarities onto representational dissimilarity in a graded manner. This graded coupling between contextual distance and representational distance shows how vectorial semantic representations can implement semantic pattern separation.

Pattern completion is the computational process by which the hippocampus retrieves a holistic neural representation from a partial or degraded input (Rolls, 2013; Yassa & Stark, 2011). In language, the brain can map novel or incomplete inputs onto the appropriate semantic region defined by previously learned word relationships (**Figure 5E-F**). Concept cells are excellent at pattern completion, in the sense that they efficiently retrieve a holistic identity from degraded sensory inputs (Quiroga, 2012). However, they lack the high-dimensional capacity for **relational pattern completion**, in which the system can utilize the high-dimensional relationships between modes to infer a specific context. By leveraging the superposition of unrelated features, a single polysemantic neuron can participate in multiple attractor states, allowing the broader network to complete a complex narrative by collapsing onto the correct semantic manifold as contextual cues accumulate (Elhage et al., 2022; Whittington et al., 2020).

We reasoned that if the hippocampal population maintains a high-dimensional, structured semantic geometry that positions words along many semantic dimensions simultaneously, then

this geometric structure itself should support generalization. That is, a semantic boundary defined by one set of words should correctly classify words never used to define that boundary. We borrowed the framework of cross-condition generalization performance (CCGP, Bernardi et al., 2020; Courellis et al., 2024). For each of the first 100 principal components of GPT-2 Layer-36 embedding space, we split words into high versus low groups, trained a linear SVM on 70% of the words, and tested on the remaining 30% (Bernardi et al., 2020). Because the classifier never sees the test words during training, above-chance accuracy requires that the population geometry encodes the semantic dimension in an abstract format (**Figure 5E-F, Methods**). Generalization performance exceeded both chance and the shuffled baseline; 77 out of 100 semantic dimensions reached significance (all $p < 0.05$, **Figure 5G**), corresponding to a shattering dimensionality (SD) of 0.77 (**Figure 5H**).

This high-dimensional geometry simultaneously supports both pattern completion and pattern separation. That is, when a novel word is encountered, the neural ensemble can use many different semantic hyperplanes learned from the population responses of other words to correctly position the novel word. For example, when we hear “*dog*”, this word can be classified into the animacy plane representing “*animate*” or “*inanimate*”, and the affective plane representing “*warm*” or “*cool*” (exemplifying the process of completion). At the same time, the abundance of distinguishing dimensions ensures that words sharing one semantic feature can nonetheless be differentiated along others, allowing the same word to be interpreted differently depending on the context (exemplifying separation).

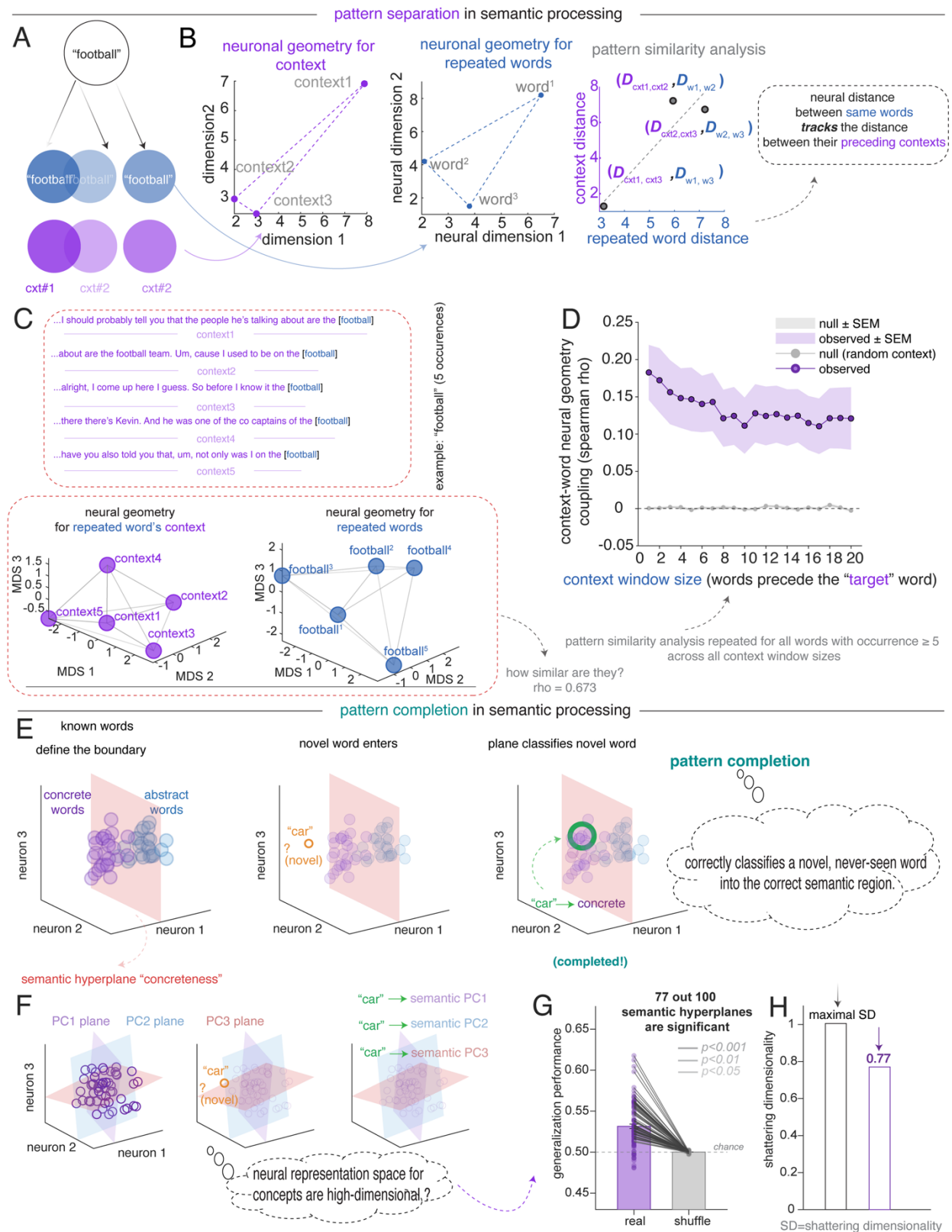


Figure 5. Hippocampal population supports semantic pattern completion and context-driven pattern separation.

(A) Schematic of context-driven pattern separation. The same word (“football”) appears in multiple lexical contexts; occurrences preceded by dissimilar contexts should elicit dissimilar population responses.

(B) Context-word geometry coupling schematic. Pairwise distances among occurrences of a repeated word are computed in lexical context space (left) and target word space (middle). Pattern separation predicts that representational distance tracks contextual distance (right).

(C) Example: “football” appears in 5 contexts. Cosine distances in lexical context space and target word space show correlated geometry (Spearman $\rho = 0.673$).

(D) Context-word coupling across windows of 1-20 preceding words. Observed coupling (purple) significantly exceeds random-context null (gray) at all window sizes (Methods).

(E) Pattern completion schematic. Population responses to known words define hyperplanes partitioning neural space. A novel word is classified by which side its response falls, without prior exposure.

(F) Multiple semantic hyperplanes partition neural space along different semantic contrasts, enabling multi-attribute pattern completion.

(G) Generalization performance across 100 semantic dimensions. Real performance (purple) significantly exceeds shuffle baseline (gray); 77 of 100 dimensions reach significance ($p < 0.05$, permutation test; Methods).

(H) Shattering dimensionality, computed as the proportion of semantic dimensions classifiable above chance. The observed value of 0.77 indicates that the population supports linear readout along three-quarters of tested dimensions, confirming a high-dimensional distributed code.

DISCUSSION

When processing natural speech, hippocampal neurons robustly encode semantic meaning; this semantic encoding is driven by selectivity for multiple distinct stimuli. The set of stimuli that most effectively drive a given neuron form an overdispersed, isotropic, and equidistant structure within the neural embedding space. This polysemanticity closely parallels the foundational architecture of LLMs, where networks utilize superposition to maximize information capacity within a finite-dimensional space, forcing individual units to represent multiple disparate features (Bricken et al., 2023; Elhage et al., 2022). Human language represents an ideal use case for superposition because the set of concepts available to language is both very high dimensional and sparse. Because these semantic features rarely co-occur, the hippocampus can project them into the same neural dimensions with low risk. By maintaining an isotropic and equidistant structure, the network ensures that these overlapping representations remain as close as possible to orthogonal, reducing the chance of catastrophic interference and allowing the broader neural population to unambiguously decode meaning (Olah et al., 2017, 2020).

Polysemantic tuning in hippocampal representations of meaning does not invalidate the theory of concept cells, although it suggests they may be part of a larger continuum of coding. Superposition theory predicts a phase transition governed by feature sparsity (Elhage et al., 2022). Classic concept cell protocols, which involve discrete concepts presented without context, naturally favor a dedicated-neuron regime (Quiroga, 2012), whereas naturalistic speech, with thousands of rapidly and transiently appearing concepts, favors polysemanticity. Both regimes may coexist within the same population, depending on the statistical structure of the input.

Most physiological studies report that conjunctive coding of stimulus and context is rare in human MTL, and that as a result pattern separation is limited (Bausch et al., 2026) or absent (Quiroga, 2020; Rey et al., 2025). In contrast, our results, obtained during language listening, revealed robust conjunctive coding (i.e., nonlinear mixing of context and semantics). This non-linear mixed selectivity is the key ingredient that allows for contextualization and efficient neural computation (Johnston et al., 2020) and its cousin, pattern separation (Barak et al., 2013; Bernardi et al., 2020; Fusi et al., 2016). It also increases coding capacity and fidelity (Johnston et al., 2020). Moreover, these mixed-selective patterns directly recapitulate those observed in hippocampal neural populations in a range of contexts (Bernardi et al., 2020; Boyle et al., 2024; Chericoni et al., 2026; Courellis et al., 2024; Tang et al., 2023). Given the strong need for

contextualization in language, such conjunctive coding is invaluable. This discrepancy, then, may reflect that hippocampal pattern separation is latent under discrete, static paradigms and actively recruited by the computational demands of continuous language.

Overall, our observations provide a bridge between semantic representation and classical models of hippocampal pattern separation (Bakker et al., 2008; Leutgeb et al., 2007; McClelland et al., 1995; Yassa & Stark, 2011). Traditional frameworks emphasizing sparse, amodal concept cells pose a computational paradox in this regard: their monosemantic rigidity lacks the high-dimensional capacity required to decorrelate closely overlapping lexical inputs (Quiroga, 2020). This limitation seemingly contradicts the amply demonstrated role of the hippocampus in contextualization and pattern separation (Suthana et al., 2021). Our results suggest that such sparse identity-coding is complemented by a polysemantic architecture—a distributed system that naturally supports the geometric requirements of pattern separation through superposition (Cayco-Gajic & Silver, 2019; Rigotti et al., 2013). Indeed, our findings in hippocampus suggest a conceptual unification between the psychology of pattern separation, which has typically been applied to relational memory, and language.

The polysemantic neuronal tuning we observe recalls the dispersed sampling characteristic of grid cells (Moser et al., 2008). Grid cells are a canonical example of neurons whose firing fields tile an environment with multiple, widely separated fields. Crucially, the canonical hexagonal periodicity of grid cells degrades in three-dimensional or topologically complex environments, resulting in fragmented, locally ordered firing fields (Ginosar et al., 2021, 2023; Grieves et al., 2021; Krupic et al., 2015). This fact suggests that grid cells, at least in spaces of more than two dimensions, exhibit a form of spatial polysemanticity, driven by the need to represent multiple independent spatial contexts within a finite neural population. The neurons in our dataset, which code locations in a high dimensional semantic embedding space, show a similar pattern: isotropic overdispersion without hexagonality. We hypothesize that these overdispersed response sets may reflect the same computational principle that may motivate 3-D grid coding, maximizing coverage of a representational space while minimizing redundancy and interference. In spatial navigation, distributed firing fields support efficient generalization across locations (Bellmund et al., 2018; Whittington et al., 2020). More speculatively, an overdispersed sampling of conceptual space could allow hippocampal neurons to link distant experiences while

preserving separability among neighboring memories, thereby facilitating relational binding, and flexible inference, and, potentially, creativity.

More broadly, our results underscore the value of adopting the population doctrine perspective (Ebitz & Hayden, 2021; Saxena & Cunningham, 2019; Urai et al., 2022; Vyas et al., 2020). Such an approach has proven fruitful in domains ranging from motor control to decision making (Cao et al., 2025; Churchland et al., 2012; Cunningham & Yu, 2014; Elsayed et al., 2016; Jazayeri & Ostojic, 2021; Tang et al., 2023; Yan et al., 2025). Specifically, conceptualizing language representations as geometric structures provides a natural bridge between neural data and computational models of semantics, which similarly represent meaning in continuous vector spaces (Franch et al., 2025; N. Kriegeskorte & Douglas, 2018; Piantadosi et al., 2024).

METHODS

Ethics approval

The experiment's procedures were conducted in accordance with the standards set by the Declaration of Helsinki and approved by the Institutional Review Board for Baylor College of Medicine, Houston, TX (H-50885 and H-18112). Participants provided written informed consent after the experimental procedure had been fully explained and were reminded of their right to withdraw at any time during the study.

Patient recruitment

Fifteen English-speaking epilepsy patients participated in the study and were aware that participation was voluntary and would not affect their clinical course. Patients' age ranged from 18-48 years old (mean $\pm$ s.e.m = 30.07 $\pm$ 9.15 years), with 9 female patients and 6 male patients.

Experimental stimuli

Participants listened to a set of naturalistic spoken narratives selected to be engaging and linguistically diverse (Franch et al., 2025). Specifically, the stimulus set comprised six episodes from The Moth Radio Hour (each approximately 5-13 minutes long), totaling 47 minutes and 25 seconds of audio (7346 words). The stories were: "Life Flight," "The One Club," "The Tiniest Bouquet," "My Father's Hands," "Wild Women and Dancing Queens," and "Juggling and Jesus." Each excerpt featured a single speaker delivering an autobiographical story to a live audience. Audio was presented continuously through the built-in speakers of the participant's hospital television. To synchronize stimulus timing with neural recordings, the audio output from the playback computer was routed as an analog input directly into the Neural Signal Processor (sampled at 30 kHz), ensuring precise temporal alignment between the stimulus waveform and recorded neural activity.

Microelectrodes recording in the hippocampus

None of the analyzed data included seizures. Single neuron data were recorded from stereo-electroencephalography (sEEG) electrodes with microwires extending from the tip (AdTech Medical Behnke-Fried). Each patient had an average of 3 probes terminating in the left and right hippocampus. Electrode locations were verified by co-registered pre-operative MRI and post-operative CT scans. Each probe includes 8 microwires, each with 1 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 (Chaure et al., 2018) and then manually evaluated. Noise was removed and each signal was classified as multi or single unit using several criteria: consistent spike waveforms, waveform shape (slope, amplitude, trough-to-peak), and exponentially decaying ISI histogram with no ISI shorter than the refractory period (1 ms).

The initial number of neurons was simply the total number of neurons that we collected. All preprocessing was done by a lab member blind to the goals of the study. In preprocessing (spike sorting), we only kept the isolated neurons. We analyzed all collected neurons and did not restrict analyses to task-responsive neurons.

Firing rate responses to words

We computed firing rates during a time window starting 80 ms after the onset of each word (to account for the approximate response latency) and lasting the duration of the word plus 40 ms, accounting for transmission delays to hippocampus (Charest et al., 2009; Chavez et al., 2025; Zhu et al., 2026). Spike counts were divided by window duration and multiplied by 1000 to yield spikes per second.

Control analysis: firing rate responses to words in narrower time windows (Figure S9)

To control for the possibility that early responses to the next word may cause spurious coding density and polysemanticity, we computed firing rates during a narrower time window (starting 80 ms after the onset of each word and lasting the duration of the word). Spike counts were divided by window duration and multiplied by 1000 to yield spikes per second.

Audio transcription

After experiments, the audio.wav file was automatically transcribed using Assembly AI, a state-of-the-art AI model trained to transcribe speech. Speaker diarization was achieved with Assembly AI's internal speaker diarization model. The transcribed words, corresponding timestamps, and speaker labels output from these models were converted to a TextGrid and then loaded into Praat (Boersma, 2001), a well-established software for speech analysis (<https://www.fon.hum.uva.nl/praat/>). The original .wav file was also loaded into Praat. Trained lab members manually inspected the spectrograms, timestamps, and speaker labels, correcting each word to ensure that onset and offset times were precise and assigned to the correct speaker (this process took about 4-5 hours per minute of speech). The TextGrid output of corrected words and timestamps from Praat was converted to an Excel file and loaded into Python (version 3.11) for further analysis.

Electrode visualization for Figure 1A

Electrode visualization was performed using MNE-Python (version 0.24.0) with FreeSurfer's fsaverage template brain. The visualization pipeline created an ultra-transparent glass brain (3% opacity) with highlighted hippocampal structures to clearly display electrode positions within deep brain regions. For each patient, DICOM images of the preoperative T1 anatomical MRI and the postoperative Stealth CT scans were acquired and converted to NIfTI format (Li et al., 2016). The CT was aligned to MRI space using FSL (Jenkinson & Smith, 2001; Jenkinson et al., 2002). The resulting co-registered CT was loaded into BioImage Suite (version 3.5β1; Joshi et al., 2011), and the electrode contacts were manually localized. Electrodes' coordinates were converted to the patient's native space using iELVis MATLAB functions (Groppe et al., 2017) and plotted on the FreeSurfer (version 7.4.1) reconstructed brain surface (Dale et al., 1999). Microelectrode coordinates are taken from the first (deepest) macro contact on the Ad-Tech Behnke Fried depth electrodes. RAVE (Magnotti et al., 2020) was used to transform each patient's brain and electrode coordinates into MNI152 average space. The coordinates were plotted together on a glass brain with the hippocampus segmentation.

Semantic embedding extraction from language models

fastText embeddings. We used the pre-trained *fastText* model in MATLAB to extract word embeddings for all words in our dataset (Joulin et al., 2016; Mikolov et al., 2013). This pre-trained model provides 300-dimensional word embedding vectors, trained on millions of words text, to capture semantic relationships between words. Any surname words, such as "Harwood" or proper nouns like "Applebee's" that did not have word embeddings were discarded from the analysis (these were rare in our sample).

GPT-2 embeddings. In an excel file of our transcribed words, each row corresponded to an individual token, including punctuation and sentence terminators such as “.”, “?”, and “!”. We fed each stimulus sentence by iterating through the file and appending words until encountering a recognized sentence delimiter. This process yielded multi-word segments that more accurately captured natural linguistic boundaries. Any remaining words following the last delimiter were grouped into a final sentence. Notably, punctuation tokens were preserved to maintain contextual fidelity but were tracked carefully to avoid introducing sub-word alignment errors during tokenization. To extract GPT-2 embeddings, we employed the “gpt2-large” model (Radford et al., 2019) from the Hugging Face Transformers library (Wolf et al., 2020). This version of GPT-2 features a total of 37 layers (1 initial embedding layer plus 36 transformer layers), each producing 1280-dimensional hidden states, with a maximum context length of 1024 tokens. Although GPT-2 supports a 1024-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 (e.g., “.”, “!”, “?”) 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 (BPE) 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 sub-word tokens and averaged their corresponding vectors to obtain a single word-level embedding. Because GPT-2 often splits words into multiple sub-word fragments (otherwise known as “tokens,” e.g., “computa” + “tion”), averaging those fragments preserves consistency with the transcript’s original word boundaries. The GPT-2 outputs resulted in 1280-dimensional embeddings for each layer.

Ridge regression analysis for Figure 1D

We used ridge regression to learn a linear mapping from semantic embeddings to single-neuron firing rates. The predictor matrix consisted of the first 100 PCs (from PCA, explain 61.52% variance) of GPT-2 Layer-36 embeddings (Radford et al., 2019), augmented with an intercept column. A separate ridge regression model was fit for each neuron. The ridge penalty was applied to all predictor dimensions except the intercept term. We selected the regularization parameter λ via 5-fold cross-validation over words. For each candidate λ (25 values logarithmically spaced from 10^{-3} to 10^5), the total cross-validated Gaussian log-likelihood was computed across all neurons and folds, then we selected the λ which maximizes this quantity. Specifically, for each fold, the model was trained on the training partition and evaluated on the held-out test partition. Encoding model quality was quantified as the improvement in cross-validated Gaussian log-likelihood relative to an intercept-only null model ($\Delta LL = LL_{\text{model}} - LL_{\text{null}}$). The null model predicted each neuron’s firing rate as the training-set mean, with its own variance estimated from training-set residuals. ΔLL was summed across neurons and test observations within each cross-validation fold and then summed across folds.

Control analyses for Figure 1D (results see Figure S1A-B)

To test whether the distributed scaling pattern reflects a genuine property of the neural code or an artifact of the encoding pipeline, we compared two encoding pipelines with opposing architectural biases. The original pipeline used PCA to reduce GPT-2 Layer-36 embeddings to 100 components, yielding a dense, signed basis, followed by ridge regression, which penalizes large weights and favors smooth, distributed solutions. The control pipeline replaced PCA with non-negative matrix factorization (NMF),

which decomposes embeddings into non-negative, parts-based components that naturally permit sparse word loadings, and replaced ridge regression with elastic net (L1 ratio = 0.9, the estimator strongly favors sparse solutions), which permits exactly zero weights and can reveal sparse pattern if present in the data. Because NMF requires non-negative input, GPT-2 Layer 36 embeddings were shifted to non-negative values by subtracting the column-wise minimum from each feature dimension prior to decomposition. The shifted matrix was then factorized into 100 components using the multiplicative update algorithm (Lee & Seung, 1999) with 10 random initializations, retaining the factorization with lowest reconstruction error. The resulting word-by-component score matrix (W) was z-scored across words to place components on comparable scales for regression. Elastic net regularization paths were fit per neuron with 5-fold cross-validation for lambda selection. Both pipelines were evaluated using the same cross-validated Δ LL metric, and the analysis was repeated for three embedding models (fastText, GPT-2 Layer-24, and GPT-2 Layer-36) to ensure generality.

Encoding-based decoding (Figure 1E-F)

To evaluate whether word identity could be recovered from population activity, we implemented an encoding-based decoding procedure (Tang et al., 2023). Both neural firing rates and GPT-2 Layer-36 embeddings were z-scored across words. The embedding matrix was further reduced to 100 principal components via PCA.

Per-neuron lambda tuning. A separate ridge regression encoding model was fit for each neuron. The regularization parameter λ was tuned independently per neuron using 5-fold cross-validation over words, selecting the λ (from 25 values logarithmically spaced between 10^{-3} to 10^5) that minimized mean squared prediction error.

Leave-one-out decoding. Decoding was performed using leave-one-out cross-validation (LOOCV) across words. On each iteration, one word was held out as the test item and the encoding model was trained on the remaining words using the per-neuron λ values determined above. The trained model was then used to generate predicted neural population vectors for all candidate words from their embeddings. The observed neural population vector for the held-out word was compared to each candidate's predicted vector using Pearson correlation, and candidates were ranked by similarity. The decoding rank indicates the position of the true word among all candidates (rank 1 = the true word has the highest similarity score; rank 2 = the true word has the second highest similarity score; rank = (W+1)/2, where W is the number of candidate words, represents the chance level).

Statistical testing. Decoding significance was assessed using a permutation test (500 permutations). On each permutation, the pairing between embeddings and neural responses was shuffled within the training set only, preserving the true test observation. The full LOOCV decoding procedure was re-run under this null, and top-1 accuracy was recorded. The p-value was computed as the proportion of permutation accuracies equal to or exceeding the observed accuracy.

Construction of Representational Dissimilarity Matrices (RDM) and representation similarity analyses (RSA) (Figure 1G)

For both neural and semantic spaces, we computed *Representational Dissimilarity Matrices* (RDMs; Diedrichsen & Kriegeskorte, 2017; Nikolaus Kriegeskorte & Kievit, 2013; Nili et al., 2014) using cosine distance as the dissimilarity metric. The cosine distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$d_{\text{cosine}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

For a dataset with n words, this yields a condensed distance vector of length $n*(n-1)/2$ containing all unique pairwise distances. The neural RDM for language l was computed as:

$$D_{\text{neural}}^{(l)} = \{d_{\text{cosine}}(\mathbf{n}_i^{(l)}, \mathbf{n}_j^{(l)}) : 1 \leq i < j \leq n\}$$

Where $\mathbf{n}_i^{(l)}$ represents the neural population vector for word i in language l . Similarly, the semantic RDM was computed as:

$$D_{\text{semantic}}^{(l)} = \{d_{\text{cosine}}(\mathbf{s}_i^{(l)}, \mathbf{s}_j^{(l)}) : 1 \leq i < j \leq n\}$$

Where $\mathbf{s}_i^{(l)}$ represents the GPT2-Layer-36 embedding vector for word i in language l .

Neural-semantic correspondence quantification

The correspondence between neural and semantic representational geometries was quantified by computing Spearman correlation between the vectorized RDMs (i.e., $D_{\text{semantic}}^{(l)}$ and $D_{\text{neural}}^{(l)}$). This correlation coefficient quantifies the degree to which neural dissimilarity patterns mirror semantic dissimilarity patterns. A high positive correlation indicates that words with similar meanings (small semantic distance) tend to evoke similar neural responses (small neural distance).

Representation similarity analyses and decimation scaling analyses for Figure 1H-I

To assess how the correspondence between neural and LLM semantic geometry scales with population size, we performed a decimation analysis on the RSA. For each population size $N \in \{5, 10, 20, 40, 80, 160, 320, 498\}$, we randomly sampled N neurons from the full population and computed the Spearman rank correlation (ρ) between pairwise cosine distance vectors in neural space and in GPT-2 Layer-36 embedding space (we also did analyses in GPT-2 Layer 24 and fastText embeddings). This procedure was repeated 200 times at each population size, each time with a different random draw of neurons. A shuffle control was performed by randomly permuting the word order of the neural firing rate matrix before computing pairwise distances, while the GPT embedding distances were held fixed (50 repetitions per population size). For visualization in Figure 1H, the plotted quantity is ρ above chance (real mean minus shuffle mean at each N). Shaded bands denote 95% confidence intervals (2.5th-97.5th percentiles) of the neuron-resampling distribution, shifted by the shuffle mean. That is, the lower and upper bounds correspond to the 2.5th and 97.5th percentiles of the real ρ distribution minus the shuffle mean at each N , treating the shuffle mean as a fixed baseline correction rather than propagating shuffle variance through error combination. At $N = 498$, the band collapses to zero width because all neurons are included in every iteration, leaving no variability across resampled subsets. Figure 1I reports the effect size (z-score) at each population size, computed as the mean ρ above chance divided by the standard error of the difference. Because z-scores are scalar values at each population size rather than distributions, they are plotted as discrete points with horizontal reference lines indicating significance thresholds ($p < 0.05$ and $p < 0.001$).

Control analyses for Figure 1H-I (results see Figure SIC)

To rule out the possibility that scaling improvements reflect mere redundancy (i.e., averaging out noise across copies of the same signal), we implemented a copy-plus-noise control. We identified the single neuron with the highest cross-validated encoding performance (ΔLL relative to an intercept-only null), fit a ridge regression to decompose its activity into a predicted (signal) component and a residual (noise) component, then generated a synthetic population of up to 498 units by adding independent Gaussian noise (matched in variance to the empirical residuals) to copies of the signal. We then ran the

identical RSA decimation procedure on this synthetic population. If the real data exhibits a qualitatively different scaling profile from the copy-plus-noise control, particularly continued gains at large N where redundant copies would plateau, this constitutes evidence that additional neurons carry non-redundant semantic information, consistent with a distributed code.

Multidimensional Scaling (MDS) network visualization (Figure 1J-M, Figure 3H, Figure 4D and Figure 5C)

In Figure 1J-M, Figure 3H, Figure 4D and Figure 5C, we employed classical Multidimensional Scaling (MDS) (Mead, 1992) combined with network visualization techniques to visualize the geometric organization of neural representations for words. This approach reveals the topological structure of semantic representations in neural space by projecting high-dimensional neural patterns onto interpretable low-dimensional manifolds while preserving pairwise dissimilarities. The resulting 2D coordinates were visualized as network graphs where nodes represent words and edges encode distance between words.

Symmetric non-negative matrix factorization (NMF) of neural semantic tuning (Figure 1N-O)

To identify shared semantic modes across the neural population, we decomposed the structure of neuron-level semantic tuning using symmetric NMF.

Tuning vectors. For each neuron, a tuning vector was estimated by fitting a ridge regression model (method see above).

Neuron similarity matrix. Tuning vectors were rectified (negative values set to zero) and L2-normalized. The neuron-by-neuron similarity matrix S was computed as the dot product of the normalized tuning vectors, yielding a non-negative symmetric matrix with diagonal entries set to 1.

Symmetric NMF. The similarity matrix was decomposed as $S \approx WW^T$, where W ($N \times K$) is a non-negative weight matrix and K is the number of semantic modes. The factorization was performed using multiplicative update rules: at each iteration, W was updated element-wise as $W \leftarrow W \odot (SW) / (W(W^TW) + \epsilon)$, followed by column-sum normalization. The algorithm was run for 500 iterations per replicate, with 50 random initializations, and the solution with the lowest relative Frobenius reconstruction error was retained.

Selection of K . The optimal number of modes was selected using an elbow method. Reconstruction error was computed for $K = 2$ to 15, with 10 random restarts at each K . The second derivative of the reconstruction error curve was computed, and the K at which the second derivative was maximized (i.e., the point of greatest curvature change) was selected as optimal ($K = 13$).

Validation. We confirmed that each NMF solution captured meaningful structure by comparing its reconstruction error against a null distribution. For each of 500 permutations, we generated a random weight matrix of the same dimensions with entries drawn from a uniform distribution, column-normalized it, and scaled it to match the column sums of the real W . We computed the p-value as the proportion of random reconstruction errors less than or equal to the real reconstruction error.

Mode characterization. For each mode k , we computed a prototype vector in semantic embedding space as the weighted average of all neurons' tuning vectors (prior to rectification), with weights given by the k -th column of W , normalized to sum to 1. We then L2-normalized each prototype and projected the embedding matrix onto each prototype to obtain word scores, which quantify how strongly each word aligns with the mode's semantic direction. We reported the top 8 words per mode (in Figure 1O) after labeling repeated words with numeric suffixes (e.g., "golf¹", "golf²" in mode#1) to distinguish multiple occurrences.

Definitions of word categories

To identify the natural semantic categories present in our word data (Figure 1O), all unique words were clustered into groups based on related meanings using a word embedding approach (Joulin et al., 2016; Mikolov et al., 2013). We used the pre-trained *fastText Word2Vec* model in MATLAB to extract word embeddings for all words in our dataset (Joulin et al., 2016; Mikolov et al., 2013). This pre-trained model provides 300-dimensional word embedding vectors, trained on millions of words text, to capture semantic relationships between words. Any surname words, such as “*Harwood*” or proper nouns like “*Applebee’s*” that did not have word embeddings were discarded from the analysis. To compute and visualize semantic clusters, we used a t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm (Maaten & Hinton, 2008) on word embedding values to reduce the dimensionality of each unique word based on their cosine distance to all other words, thus reflecting their semantic similarity. Words with similar meanings now have similar 2D coordinates. We then applied the K-means clustering algorithm to these 2D word representations and visualized clustered words on a 2D word map ($k = 12$ clusters, determined from the silhouette score; Rousseeuw, 1987). We then manually inspected and assigned a distinct semantic label to each cluster and adjusted categories for accuracy. For example, words bordering the edges of categories might get mis-grouped and were manually corrected. This method is comparable to that used in other studies of semantics (Huth et al., 2016; Jamali et al., 2024).

Simulation of concept cells (sparse semantic coding) under the Waydo et al. theoretical model (Figure S2)

To enable a direct comparison between classic concept-cell results from image presentation paradigms and neural responses during continuous naturalistic language listening, we developed a simulation framework that generates concept-cell-like sparse coding under controlled target sparseness levels. The simulation follows the logic of the Waydo-style sparse coding model and published concept cell papers (Quiroga et al., 2008, 2007, 2005; Waydo et al., 2006): only a small fraction of neuron-stimulus pairs evoke detectable responses, and “responsive neurons” and “evocative stimuli” emerge from that sparse response matrix. The simulation was calibrated to reflect both (i) published concept-cell benchmarks and (ii) the timing constraints of naturalistic listening.

We simulated a population of $N = 498$ neurons and $W = 516$ stimuli (unique nouns), matching the scale of our empirical dataset used for comparison. To preserve the temporal structure of the language stimulus set, each simulated stimulus was assigned a real word duration taken from our noun stimulus list. Word durations were converted to seconds and used as the stimulus presentation duration for the simulated “response window.” In addition, each trial included a fixed baseline window (0.1 s) to mimic the baseline period used in image-based concept-cell paradigms.

For each simulated sparseness condition, we specified a target probability that a neuron responds to a stimulus (the target “sparseness” level). For each neuron-stimulus pair, a latent binary variable determined whether the response window was drawn from a baseline firing regime or an elevated firing regime. Baseline and response firing rates were specified by parameters r_0 (baseline rate) and r_1 (response rate). We set the r_0 to 0.02Hz and $r_1 = 5$ Hz, based on empirical results (Quiroga et al., 2005). Crucially, because stimulus durations vary across words, we generated expected spike counts by combining the firing rate regime with the word-specific duration.

To capture realistic trial-to-trial variability beyond Poisson spiking, spike counts were generated using a Gamma-Poisson (negative binomial) model governed by a Fano factor parameter ($F = 2.66$;

Quiroga et al., 2005). This creates overdispersed spike counts consistent with observed neural variability. A lower bound on the Gamma shape parameter was imposed ($\kappa_{\text{floor}} = 10$) to prevent unstable behavior for very small expected counts.

Waydo-style detection criterion and thresholding

To classify a neuron-stimulus pair as a “response,” we implemented a detection rule analogous to concept-cell screening criteria: (1) For each neuron, baseline spike counts were used to compute a neuron-specific baseline mean and standard deviation. (2) A neuron-specific detection threshold was set as mean + $kSD \times SD$, with $kSD = 5$, matching the commonly used stringent criterion in concept-cell studies. (3) A response was counted only if the response-window spike count exceeded the threshold and also exceeded a minimum spike-count criterion ($nSpkMin = 2$). This produced a binary response matrix of size $W \times N$, indicating whether each neuron responded to each stimulus.

Calibration of latent response probability

For each target sparseness level, the simulation calibrated the latent response probability (π) using a bisection search so that the empirically observed response probability (the fraction of ones in the response matrix) matched the target sparseness (α) as closely as possible. This ensured that different sparseness conditions could be compared on a common scale. We simulated 10 target sparseness levels spanning 0.2% to 1.0% (inclusive). For each sparseness condition, we computed the standard Waydo-style population statistics from the binary response matrix: (1) Number of responsive neurons (N_r): neurons that respond to at least one stimulus; (2) Stimuli per responsive neuron: average number of stimuli eliciting a response among responsive neurons; (3) Neurons per evocative stimulus: average number of neurons activated among evocative stimuli. We additionally computed the fraction N_r/N as a “responsive neuron fraction.” These quantities were compared to published benchmarks (**Figure S2E**).

Validation of the Waydo-style lifetime sparseness model

We repeat the simulation process above, but using the exact number of neurons and stimulus numbers to make sure the simulation results can be compared with the published results. Specifically, we set the time for baseline and response as 0.7s (Quiroga et al., 2008: “*The response to a picture at a specific duration was defined as the median number of spikes across trials between 300 and 1,000 ms after stimulus onset.*” Also see Quiroga et al., 2007, 2005). For Quiroga et al., 2008, neuron number=440, Stimulus=98. For Quiroga et al., 2007, number=1547, Stimulus=88. For Quiroga et al., 2005, neuron number=993, Stimulus=88.

Population sparseness (Figure 2A-B)

To quantify the fraction of the neural population activated by each word, we computed population sparseness (Rolls & Treves, 1997) for each word. Whereas lifetime sparseness characterizes the selectivity of a single neuron across stimuli, population sparseness characterizes the selectivity of a single stimulus across the population. For a given word with firing rates $r = [r_1, r_2, \dots, r_N]$ across N neurons:

$$S_{pop} = (1 - (\text{mean}(r_n)^2 / \text{mean}(r_n))) / (1 - 1/N)$$

Values of S_{pop} range from 0 (all neurons respond equally, maximally dense) to 1 (only a single neuron responds, maximally sparse). Words for which fewer than two neurons had valid (non-zero) firing rates, or for which all firing rates were zero, were excluded from the population sparseness computation. We

compared the observed population sparseness distribution to simulated populations across 10 levels of target sparseness (α).

Lifetime sparseness (Figure 2C-D)

To quantify the selectivity of individual neurons, we computed lifetime sparseness (Rolls & Treves, 1997) for each neuron. For a neuron with firing rates $r = [r_1, r_2, \dots, r_W]$ across W words, lifetime sparseness was defined as:

$$S_{life} = (1 - (\text{mean}(r_W)^2 / \text{mean}(r_W^2))) / (1 - 1/W)$$

where $\text{mean}(r_W)$ is the mean firing rate and $\text{mean}(r_W^2)$ is the mean squared firing rate across all words.

Values of S_{life} range from 0 (uniform response across all words, maximally dense) to 1 (response to only a single word, maximally sparse). Neurons with zero total firing or fewer than two valid observations were excluded. To contextualize the observed sparseness distribution, we compared real neurons against simulated populations with known sparse coding properties. Simulated populations were generated across 10 levels of target sparseness (parameterized by α), spanning a range from highly sparse to moderately dense coding. To avoid inflating selectivity estimates through stimulus repetition, all sparseness metrics were computed on unique words only; repeated words of the same word were averaged prior to analysis.

Normalized entropy (Figure 2E)

To quantify the uniformity of each neuron's response distribution across words, we computed normalized entropy (Lehky et al., 2005). For each neuron with firing rates $r = [r_1, r_2, \dots, r_W]$ across W words, firing rates were converted to a probability distribution by normalizing to unit sum:

$$p_w = r_w / \sum_{w=1}^W r_w$$

Shannon entropy was then computed as:

$$H = - \sum_{w:p_w > 0} p_w \cdot \log_2(p_w)$$

where the sum is restricted to words with $p_w > 0$. Entropy was normalized by the maximum possible entropy for W words:

$$H_{norm} = \frac{H}{\log_2(W)}$$

yielding values between 0 (neuron responds to only one word, maximally selective) and 1 (neuron responds uniformly to all words, maximally distributed). Neurons with zero total firing or fewer than two valid observations were excluded.

The distribution of normalized entropy for real neurons was compared against simulated populations across 10 levels of simulated sparseness, computed identically. Radar plots were additionally used to visualize the shape of the entropy distribution: for each condition, the histogram of normalized entropy values (20 bins spanning $[0, 1]$) was plotted in polar coordinates, with each bin mapped to an angular position and the bin probability mapped to radius. A uniform reference circle indicates the expected shape under a distributed code; deviations from circularity indicate sparse code. Computations were restricted to unique words, as described above.

Information per word (Figure 2F)

To quantify how much information each neuron carries about word identity, we computed information per word (bits/word) using the framework (Skaggs, McNaughton, et al., 1992) adapted to the lexical domain. For each neuron with firing rates $r = [r_1, r_2, \dots, r_W]$ across W words, each with prior probability $p(w)$ and duration $T(w)$:

$$I_{word} = \sum_{\substack{w: r_w > 0 \\ W}} p(w) \cdot T(w) \cdot r_w \cdot \log_2 \left(\frac{r_w}{\bar{r}} \right)$$

where $\bar{r} = \sum_{w=1}^W p(w) \cdot r_w$ is the overall mean firing rate computed across all words. Because each unique noun was represented once, $p(w) = 1/W$ for all words. $T(w)$ is the duration of each word in the continuous speech stimulus, which varies across words. The information sum is restricted to words with non-zero firing rates ($r_w > 0$), as the $\log_2$ term is undefined at zero. Intuitively, this metric accumulates information from every word for which a neuron's firing rate deviates from its mean: even neurons with high entropy (broadly distributed firing) carry substantial information if their rate modulations, though small, are systematic across words.

Dissemination RSA for simulated concept cells (Figure S3)

To test whether the monotonic scaling of representational similarity with population size (Figure 1H) could arise under a sparse concept-cell code, we repeated the RSA decimation analysis on simulated concept cell populations. For each of the 10 sparseness levels (response probability ranging from 0.20% to 1.00% of words; see Figure S2 and method Simulation of concept cells (sparse semantic coding) under the Waydo et al. theoretical model above), we constructed a population of 498 simulated neurons and computed pairwise cosine distances between all words in neural firing rate space. We then correlated the resulting neural distance matrix with the GPT-2 Layer 36 embedding distance matrix using Spearman rank correlation, following the same procedure as the real-neuron analysis. At each population size ($N = 5, 10, 20, 40, 80, 160, 320, 498$), we randomly subsampled N neurons and computed the RSA correlation, repeating this procedure 200 times to obtain means and 95% confidence intervals. Unlike the real hippocampal population, which shows a monotonic increase in RSA correlation with population size, simulated concept cell populations produced near-zero correlations at all population sizes and sparseness levels (Spearman rho range across all simulated sparseness: -0.014 to 0.030 at $N = 498$), confirming that a sparse code in which each neuron responds to very few words cannot reconstruct the continuous semantic geometry captured by the language model.

Single-neuron tuning landscapes in semantic space (Figure 3A-B)

To visualize the tuning profile of individual neurons across semantic space, we extracted GPT-2 embeddings (Layer 36) for 50,000 nouns from an independent large text corpus, z-scored them, and projected them into a two-dimensional semantic map using t-SNE (MATLAB's `tsne` function, Barnes-Hut algorithm, perplexity = 30).

Predicted firing rates. For each neuron, we used the ridge regression encoding model fit on the experimental data (see Methods for Figure 1D) to predict firing rates across the 50,000 nouns. The z-scored GPT-2 embeddings were projected onto the first 100 principal components (computed from the experimental embedding matrix), and predicted firing rates were computed as $FR = X \cdot \beta + \beta_0$, where X is the $50,000 \times 100$ PC score matrix, β are the regression weights, and β_0 is the intercept. Predicted rates

were passed through a softplus nonlinearity ($f(x) = \log(1 + e^x)$, implemented in numerically stable form) to enforce non-negativity, then z-scored across nouns.

Visualization. For each neuron, the predicted firing rate (z-scored) was plotted as a color-coded scatter on the t-SNE coordinates. To reduce file size, only nouns with raw predicted firing rate > 0 were plotted. Points were colored using a diverging colormap (white for values ≤ 0 , warm/inferno gradient for positive values, clipped at ± 2.5 z-score). Marginal distributions along each t-SNE axis were computed by binning coordinates into 100 equally spaced bins and computing the mean absolute z-scored firing rate within each bin. These marginals were displayed as gray bar plots along the top (t-SNE dimension 1) and right (t-SNE dimension 2) edges of the scatter plot, providing a summary of the distribution of the neuron's tuning across semantic space. Axis limits were held constant across neurons using the full range of the 50,000-noun t-SNE coordinates.

Grid-cell-like feature analysis of semantic lobes

To assess whether neurons with multiple semantic lobes exhibit spatial organization analogous to grid cells, we quantified four canonical grid-cell properties: regularity, equidistance, isotropy, and hexagonal symmetry. Statistical significance was established via permutation testing (500 permutations for regularity, equidistance, isotropy, and gridness score). Details of each analysis are described below.

Lobe identification and centroid computation.

For each neuron, predicted firing rates across 50,000 nouns were softplus-transformed and z-scored. Words exceeding the 98th percentile of this distribution were designated as “high-firing words.” These high-firing words were then clustered using DBSCAN ($\epsilon = 5$, $\text{minPts} = 5$) applied to their coordinates in the t-SNE semantic plane, where each resulting cluster defines one semantic lobe, which is a spatially contiguous hotspot of elevated firing in semantic space. For each lobe, we computed a centroid as the firing-rate-weighted average coordinate of its constituent words, such that words eliciting stronger responses exert greater influence on the centroid's position. For example, if a lobe contains the words “dog,” “cat,” “lion,” “fish,” and “bird,” and the neuron fires most strongly to “dog,” the centroid will be pulled toward the position of “dog” in the embedding space rather than lying at the unweighted geometric center of all five words. Centroids were computed in both the two-dimensional t-SNE plane (for visualization) and the original high-dimensional embedding space (for regularity, equidistance, and isotropy analysis). Neurons with fewer than two lobes were excluded from regularity and equidistance analyses; neurons with fewer than three lobes were excluded from isotropy analyses.

Rationale for CV-based metrics in high-dimensional embedding space.

Classical grid cell analyses quantify regularity, equidistance, and isotropy using two-dimensional spatial autocorrelation maps (Hafting et al., 2005; Stensola et al., 2012). These methods are inherently tied to 2D geometry and may not be directly applied to semantic embeddings, which reside in a 1280-dimensional space. We therefore adopted the coefficient of variation (CV) of pairwise distances as our primary metric, for several reasons. First, pairwise distances (here, cosine distances) are well-defined in arbitrary dimensions and do not require dimensionality reduction, avoiding distortions introduced by nonlinear projections such as t-SNE. Second, CV is a scale-free measure of dispersion that directly captures the intuition behind regularity: if all inter-lobe distances are similar, CV is low regardless of the absolute scale or dimensionality. Third, a critical concern in high-dimensional spaces is distance concentration, whereby pairwise distances between random points converge to similar values as

dimensionality increases, potentially inflating apparent regularity. Our permutation null model controls for this by shuffling word assignments among the same high-firing words while preserving lobe sizes, so that any distance concentration effect applies equally to real and null distributions. Significance is therefore assessed relative to the regularity expected from random partitions within the same high-dimensional geometry, not relative to an absolute standard.

Regularity (Figure 3C)

For each neuron with ≥ 2 lobes, we computed pairwise cosine distances in original embedding space, then measured the coefficient of variation ($CV = \sigma/\mu$) of these distances. Lower CV indicates more regular inter-lobe spacing. To construct a null distribution, we randomly partitioned the set of non-noise high-firing words into groups matching the original lobe sizes, recomputed centroids from shuffled assignments, and recalculated the pairwise distance CV. This was repeated 500 times per neuron. The z-score was defined as $z = (CV_{\text{real}} - \mu_{\text{null}}) / \sigma_{\text{null}}$, where negative values indicate greater regularity than expected by chance. The permutation p-value was computed as the fraction of null CV values less than or equal to the observed CV.

Equidistance (Figure 3D)

Using the same lobe centroids and null model as above, we assessed whether each lobe's nearest neighbor ($k = 1$) was at a consistent distance across lobes. We extracted the nearest-neighbor distance for every lobe and computed the CV across lobes. The same random-partition permutation procedure provided null distributions. Negative z-scores indicate that nearest-neighbor distances are more uniform than expected by chance. The permutation p-value was computed as the fraction of null CV values less than or equal to the observed CV.

Isotropy in embedding space (Figure 3E).

For each neuron with ≥ 3 lobes, we tested whether lobe centroids were distributed uniformly in direction around the grand centroid in the original embedding space. We computed the grand centroid as the arithmetic mean of all lobe centroids in embedding space, then obtained direction vectors from the grand centroid to each lobe centroid, normalized to unit length. We computed all pairwise cosine similarities among these unit direction vectors and measured the CV of the resulting similarity values. For centroids distributed uniformly on a high-dimensional hypersphere, pairwise cosine similarities concentrate near zero with low variance; thus, lower CV indicates more isotropic arrangement. To establish a null distribution, we generated 500 sets of random unit vectors on the hypersphere (by drawing from an isotropic Gaussian and normalizing), matched in number to the observed lobes, and computed the pairwise cosine similarity CV for each. The z-score was defined as $z = (CV_{\text{real}} - \mu_{\text{null}}) / \sigma_{\text{null}}$, and the p-value as the fraction of null CV values less than or equal to the observed CV.

Hexagonal symmetry (Figure 3F).

To test whether the spatial organization of semantic lobes exhibits six-fold rotational symmetry analogous to entorhinal grid cells, we computed a gridness score for each neuron following established procedures from the spatial navigation literature, adapted for the semantic domain (Hafting et al., 2005; Langston et al., 2010). For each neuron, we constructed a two-dimensional firing rate map by binning all 50,000 nouns onto the t-SNE semantic plane using a uniform grid (bin width = 1 t-SNE coordinate unit) and averaging the z-scored predicted firing rates (after softplus transformation) within each occupied bin.

Rate maps were smoothed with a Gaussian kernel ($\sigma = 3$ bins). We then computed the two-dimensional spatial autocorrelation of the smoothed rate map. To avoid edge effects and unvisited locations, bins with zero occupancy were excluded prior to smoothing, and the rate map was mean-centered using only occupied bins before computing the autocorrelation.

The central peak of the autocorrelation (arising from the trivial zero-lag self-correlation) was excluded using a data-driven procedure following (Langston et al., 2010). We computed the radial autocorrelation profile by averaging autocorrelation values within concentric annular rings (width = 1 bin) at increasing distances from the center. The central exclusion radius was defined as the first local minimum or the first zero crossing (positive-to-negative transition) in this radial profile, whichever occurred at a shorter distance. This approach avoids arbitrary parameter choices and adapts to the autocorrelation structure of each individual neuron.

Outside the central exclusion zone, we identified the six nearest peaks (local maxima with positive autocorrelation values) sorted by distance from the autocorrelation center. A minimum of three detectable peaks was required for a neuron to be included in the analysis. The annular region for rotational symmetry analysis was defined with inner radius equal to the distance of the nearest peak and outer radius equal to the distance of the farthest peak among the six selected, with the central exclusion zone removed. This data-driven annulus definition ensures that the analysis targets the spatial frequency band containing the first-order periodic structure, without relying on arbitrary scaling factors.

To quantify hexagonal symmetry, we extracted autocorrelation values within the annulus and computed the Pearson correlation between these values and those obtained after rotating the autocorrelation map by 30° , 60° , 90° , 120° , and 150° (using bilinear interpolation). For a pattern with perfect six-fold symmetry, rotations by 60° and 120° map the pattern onto itself (yielding high correlations), whereas rotations by 30° , 90° , and 150° produce maximal mismatch (yielding low correlations). The gridness score was defined as the mean correlation at hexagonal-consistent angles (60° , 120°) minus the mean correlation at hexagonal-inconsistent angles (30° , 90° , 150°), such that positive values indicate hexagonal periodicity.

To assess statistical significance, we generated a null distribution for each neuron by randomly permuting predicted firing rates across the 50,000 nouns (500 iterations), recomputing the full rate map, autocorrelation, and gridness score for each permutation while holding the annulus geometry fixed to that of the real data. This word-identity shuffle is analogous to the spike-train time-shift procedure used in rodent grid cell studies (Langston et al., 2010), destroying any relationship between firing rate and position in semantic space while preserving the marginal distributions of both. The z-score was defined as $z = (\text{gridness}_{\text{real}} - \mu_{\text{null}}) / \sigma_{\text{null}}$, and the permutation p-value as the fraction of null gridness scores greater than or equal to the observed value.

Semantic dispersion of preferred stimuli (Figure 3G-K)

To assess whether each neuron's most-preferred words are semantically clustered (as expected under a concept-cell model) or dispersed across semantic space (as expected under a distributed code), we analyzed the pairwise semantic distances among each neuron's top-ranked words. For each neuron, words with non-zero firing rates were ranked in descending order. The top N words ($N \in \{10, 20, 30, \dots, 100\}$) were identified, and the mean pairwise cosine distance in GPT-2-Layer-36 embedding space was computed among all unique pairs within this set. This observed distance was compared against a null distribution constructed by randomly sampling N words from the remaining words and computing the same mean pairwise distance. The null procedure was repeated 500 times per neuron. Then for each

neuron, a z-score was computed as the difference between the observed mean distance and the null mean, divided by the null standard deviation. Two one-sided p-values were computed: p_{cluster} (proportion of null distances $\leq$ observed, testing for semantic clustering) and p_{disperse} (proportion of null distances $\geq$ observed, testing for semantic dispersion). This analysis was repeated across all values of N to assess the robustness of the effect across different thresholds of stimulus preference. We further used these two p-values to decide which neurons showed significant semantic clustering or semantic dispersion ($p < 0.05$).

Control analyses for top- N words dispersion (Figure S5)

To address the concern that semantic overdispersion among each neuron's preferred words might be driven by noise in the tail of the firing rate distribution rather than by the neuron's strongest responses, we performed a firing-rate-weighted control analysis. For each neuron, we identified the top N words by firing rate ($N = 10$ to 100) and computed a weighted mean pairwise cosine distance in GPT-2 Layer-36 embedding space, where each word pair was weighted by the product of their firing rates:

$$d_{\text{weighted}} = \frac{\sum_{i < j} F R_i \cdot F R_j \cdot d_{\cos}(\mathbf{w}_i, \mathbf{w}_j)}{\sum_{i < j} F R_i \cdot F R_j}$$

where $F R_i$ is the firing rate of the neuron to word i and $d_{\cos}(\mathbf{w}_i, \mathbf{w}_j)$ is the cosine distance between the GPT-2 Layer-36 embedding vectors of words i and j . This weighting ensures that the highest-firing words dominate the distance metric, effectively downweighting low-firing words near the selection threshold that might reflect noise rather than genuine tuning.

Neuron tuning across different story-contexts (Figure 4A-C)

To assess whether semantic tuning is stable or context-dependent across different story contexts, we performed the following analyses.

Story-specific encoding. The podcast stimulus comprised 6 distinct stories. For each story, we independently fit a ridge regression encoding model (the same ridge regression fitting procedure as Figure 1D) for every neuron, yielding a story-specific tuning vector per neuron.

Split-half within-story reliability. To quantify the stability of semantic tuning within a single story, we performed a split-half analysis. For each story, PCA was computed once on all nouns, and the resulting PC scores were then randomly split into two halves. Then we fitted separate ridge regression models (the same ridge regression fitting procedure as Figure 1D) on each half, yielding two tuning vectors per neuron per story. The cosine distance between the two tuning vectors was computed as $1 - \cos(\beta_{\text{half1}}, \beta_{\text{half2}})$. This split-half procedure was repeated 50 times with different random splits, and the mean cosine distance across repeats was taken as the within-story tuning distance for each neuron-story pair. The final within-story distance for each neuron was averaged across all 6 stories.

Across-story tuning distance. Across-story tuning variability was quantified by computing pairwise cosine distances between story-specific tuning vectors (averaged across folds) for each neuron across all pairs of stories.

This across-story distance was compared to the within-story split-half distance: if tuning is context-independent, the two should be comparable; if tuning is modulated by story context, across-story tuning vector distances should exceed within-story tuning vector distances.

Cross-story representational similarity analysis (Figure 4D-E).

To test whether the population-level geometric representation of shared words is preserved across different story contexts, we performed cross-story RSA. The podcast stimulus comprised 6 stories, and many nouns appeared in multiple stories, allowing us to ask whether the same word elicits a similar pattern of population activity regardless of the story in which it occurs. We performed RSA separately for each of the 15 unique story pairs (stories 1-2, 1-3, 1-4, 1-5, 1-6, 2-3, 2-4, 2-5, 2-6, 3-4, 3-5, 3-6, 4-5, 4-6, 5-6) rather than restricting the analysis to words shared across three or more stories. This pairwise approach maximizes statistical power: while only 3 nouns appeared in all 6 stories and 8 appeared in 5 or more, the number of shared nouns per story pair ranged from 8 to 32 (mean $\pm$ SD: 19.7 ± 6.5), yielding substantially more pairwise distances for RSA computation. Moreover, restricting to words shared across many stories would bias the sample toward high-frequency, semantically general words (e.g., “thing,” “way,”), undermining the goal of testing whether semantic representations generalize across contexts for a diverse vocabulary.

For each story pair, we identified all nouns that appeared in both stories. For each shared word, we computed a population firing rate vector separately for each story by averaging firing rates across all occurrences of that word within that story for each neuron. We then computed representational dissimilarity matrices (RDMs) for each story using pairwise cosine distances among the shared words’ population vectors and quantified the correspondence between the two story-specific RDMs as the Spearman rank correlation (ρ) between the upper-triangular distance vectors. We excluded story pairs with fewer than 3 valid shared words, as fewer than 3 words yield too few pairwise distances (≤ 2) for a meaningful correlation.

Null distribution. The null hypothesis was that population geometry reflects general neural state differences between stories rather than word-specific representational structure. To test this, we constructed a null distribution by randomly selecting the same number of (different) words from each story: we drew N words without replacement from one story and N different words from another story, where N equals the number of valid shared words for that pair. Because the two RDMs in each null iteration describe different sets of words, any correlation reflects only incidental similarity in population geometry rather than word-identity-driven structure. We computed RDMs for these mismatched word sets and recorded their Spearman correlation. We repeated this procedure 500 times per story pair.

Statistical testing. For each story pair, we computed a permutation p-value as the proportion of null correlations equal to or exceeding the observed correlation. At the group level, we tested whether observed RSA values systematically exceeded null means across all 15 story pairs using the Wilcoxon signed-rank test.

Lexical context modulation (Figure 4F-K)

To test whether hippocampal semantic tuning is modulated by the local lexical context, we performed the following analyses.

Context embedding extraction. For each noun in the stimulus, we computed a context embedding as the mean of the GPT-2 Layer-36 embeddings of the preceding words within a context window, using the original word order in the 7346-word corpus. We defined context window sizes from 1 to 20 preceding words to avoid any bias from choosing a specific window. The context window excluded the target word itself. Words at the beginning of the stimulus with fewer than 20 preceding words used all available preceding context. For the model comparison analysis, we reduced context embeddings to 3 PCs via PCA computed on noun embeddings only; we then projected all words (including non-nouns in the context window) onto these noun-defined PCs to ensure a shared coordinate system.

Model comparison (word, context, and interaction effects). To decompose the contributions of word identity, local context, and their interaction to neural firing rates, three nested linear models were compared for each neuron using:

- M1 (word only): $FR \sim W$, where W is the static fastText embeddings (reduced to low-dimensional PCs, we used the first three PCs)
 - M2 (word + context): $FR \sim W + C$, where C is the contextual GPT-2 embedding (3 noun-defined PCs of the preceding context window)
 - M3 (word + context + interaction): $FR \sim W + C + W \times C$
- Where FR is firing rate of each neuron.

We z-scored and mean-centered all predictors prior to model fitting. We orthogonalized interaction terms ($W \times C$, all pairwise products of W and C dimensions) with respect to the main effects by regressing each interaction column on $[1, W, C]$ and retaining the residuals, ensuring that interaction variance was not confounded with main effects. We quantified model quality using adjusted R^2 , which penalizes model complexity.

Standardized beta magnitudes. To compare the relative contribution of word, context, and interaction terms, the mean absolute standardized regression coefficient was computed for each predictor group ($|\beta_W|$, $|\beta_C|$, $|\beta_{W \times C}|$) in M3, as well as ($|\beta_W|$, $|\beta_C|$) in M2 for each neuron. Pairwise comparisons were performed using paired t-tests across neurons.

Context-driven pattern separation (Figure 5A-D)

We tested whether the hippocampus performs context-driven *pattern separation* by asking whether repeated occurrences of the same word evoke neural population patterns that reflect differences in their preceding lexical context. If the hippocampus performs context-driven pattern separation, then occurrences of the same word (e.g., “people” appearing five times) should produce more dissimilar population responses when their preceding contexts are more dissimilar. In contrast, if word representations are strictly context-invariant (as predicted by a concept-cell model), repeated occurrences should evoke similar neural patterns regardless of context.

Context firing rate extraction. For each word at position t in the original 7346-word corpus, we computed a context firing rate vector as the mean population firing rate across the preceding w words

$$\text{contextFR}(t) = \frac{1}{w} \sum_{i=t-w}^{t-1} \text{FR}(i)$$

where $\text{FR}(i) \in \mathbb{R}^{498}$ is the population firing rate vector at position i and w is the context window size. The target word itself was excluded. For words near the beginning of the corpus with fewer than w -size preceding words, all available preceding words were used. Context extraction operated on the original corpus ordering (prior to noun selection), ensuring that non-noun words contributed to the context representation. Words with no preceding context (position = 1) were assigned a zero vector. The context firing rate vectors were stored separately for each occurrence of each repeated word type, preserving the word-level variability needed for the context-driven pattern separation analysis. This procedure was repeated independently for context window sizes $w = 1$ to 20.

Repeated words and their context geometry coupling. For each noun type with ≥ 5 occurrences, we extracted two matrices across the k occurrences of that word: a context matrix $\mathbf{C}$ ($k \times 498$ neurons), where each row is the mean population firing rate vector of the preceding context window (see *Context firing rate extraction* above), and a target word matrix $\mathbf{W}$ ($k \times 498$ neurons), where each row is the population firing rate evoked by that occurrence. We computed pairwise cosine distances separately

within **C** and **W**, yielding a context RDM and a target word RDM each of size $k*(k-1)/2$. We then calculated the Spearman rank correlation between these two distance vectors to quantify the degree of context-driven pattern separation: a positive correlation indicates that occurrences preceded by distinct contexts also elicited distinct neural responses.

Constructing baseline with random context. To confirm that observed coupling reflects specific semantic context rather than generic neural state fluctuations, we replaced actual preceding context firing rates with firing rates from randomly sampled words elsewhere in the recording. On each of 500 null iterations per word type, we drew k words without replacement from the full noun pool (1221 nouns $\times$ 498 neurons), assembled their firing rates into a surrogate context matrix, and recomputed the coupling between repeated words and random context geometry as the null baseline. At the group level, we tested the difference between observed and random-context coupling using a Wilcoxon signed-rank test. We repeated both analyses for context window sizes from 1 to 20 preceding words to avoid any bias from choosing a specific window

Generalization performance across semantic dimensions (Figure 5E-H)

To test whether hippocampal population activity encodes semantic dimensions in an abstract, generalizable format, we computed generalization performance for each principal component (100 PCs in total) of the GPT-2 embedding space, borrowing the framework of cross-condition generalization performance (Bernardi et al., 2020).

Semantic dimension definition. GPT-2 Layer-36 embeddings for all noun words were z-scored and reduced via PCA. For each of the first 100 PCs, words were binarized into “high” (top 30th percentile of PC scores) and “low” (bottom 30th percentile) groups.

Generalization performance computation. For each PC, we trained a linear Support Vector Machine (SVM, MATLAB “fitsvm”, linear kernel) to classify words as high versus low using their population firing rate vectors. On each of 20 repetitions, 70% of words from each group were randomly assigned to training and 30% to testing (minimum 3 test words per group). Generalization performance was defined as the mean classification accuracy across repetitions. The critical feature of this approach is that the classifier must generalize across different word exemplars: because training and test sets contain non-overlapping words, the SVM cannot succeed by memorizing the firing rate pattern of any specific word. Instead, above-chance classification requires that the population geometry encodes the semantic dimension in a linearly separable format that holds across arbitrary subsets of the vocabulary. This is a more stringent test than standard decoding, which can exploit stimulus-specific patterns. If a semantic dimension is encoded through a distributed, abstract representation, where the population vector systematically varies along that dimension regardless of which particular words are sampled, then a linear hyperplane learned from one subset of words will correctly classify novel words. Conversely, if the representation is entangled with word-specific features, a linear classifier trained on one subset will fail to generalize to held-out exemplars. Generalization performance thus directly tests whether semantic dimensions are represented in the low-dimensional, abstractly formatted geometry predicted by recent theoretical frameworks for flexible generalization (Bernardi et al., 2020).

Shuffle baseline. To establish a null distribution, we repeated the full 20-repetition procedure 50 times with randomly permuted group assignments: on each shuffle, the same number of words were assigned to “high” and “low” groups irrespective of their true PC scores. The p-value for each PC was computed as the proportion of shuffle values greater than or equal to the real generalization performance. PCs with $p < 0.05$ were considered to have significantly generalizable representations.

Shattering dimensionality. To quantify the overall representational capacity of the population, we computed shattering dimensionality (Bernardi et al., 2020), defined as the proportion of tested semantic dimensions for which generalization performance significantly exceeded the shuffle baseline ($p < 0.05$). A shattering dimensionality of 1 would indicate that all 100 semantic dimensions are abstractly and linearly decodable; the observed value provides an estimate of how many independent semantic dimensions are represented in a generalizable format in the population code.

Generalization performance significantly above the shuffle baseline indicates that the hippocampal population represents semantics in a high-dimensional and structured format that generalizes across different word exemplars, a hallmark of abstract, linearly decodable coding geometry.

REFERENCE

- Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Database Theory — ICDT 2001* (pp. 420–434). Springer Berlin Heidelberg.
- Babadi, B., & Sompolinsky, H. (2014). Sparseness and expansion in sensory representations. *Neuron*, 83(5), 1213–1226.
- Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. L. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science (New York, N.Y.)*, 319(5870), 1640–1642.
- Barak, O., Rigotti, M., & Fusi, S. (2013). The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 33(9), 3844–3856.
- Bausch, M., Niediek, J., Reber, T. P., Mackay, S., Boström, J., Elger, C. E., & Mormann, F. (2026). Distinct neuronal populations in the human brain combine content and context. *Nature*, 650(8102), 690–700.
- Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science (New York, N.Y.)*, 362(6415), eaat6766.
- Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4), 954–967.e21.
- Blanchard, T. C., Wolfe, L. S., Vlaev, I., Winston, J. S., & Hayden, B. Y. (2014). Biases in preferences for sequences of outcomes in monkeys. *Cognition*, 130(3), 289–299.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International*, 5, 341–345. - References - Scientific Research Publishing. (n.d.). Retrieved October 31, 2025, from <https://www.scirp.org/reference/referencespapers?referenceid=913785>
- Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A., & Fusi, S. (2024). Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron*, 112(8), 1358–1371.e9.
- Bricken, T., Templeton, A., Batson, J., Chen, B., & Jermyn, A. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. <https://scholar.google.com/citations?user=CP6aLusAAAAJ&hl=en&oi=sra>
- Cao, R., Wang, J., Lin, C., De Falco, E., Peter, A., Rey, H. G., Brunner, P., Willie, J. T., DiCarlo, J. J., Todorov, A., Rutishauser, U., Li, X., Brandmeir, N. J., & Wang, S. (2025). Feature-based encoding of face identity by single neurons in the human amygdala and hippocampus. *Nature Human Behaviour*, 9(9), 1959–1974.
- Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 134.
- Cayco-Gajic, N. A., & Silver, R. A. (2019). Re-evaluating circuit mechanisms underlying pattern separation. *Neuron*, 101(4), 584–602.
- Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.-P., & Belin, P. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, 10(1), 127.
- Chauve, F. J., Rey, H. G., & Quiñones, R. (2018). A novel and fully automatic spike-sorting implementation with variable number of features. *Journal of Neurophysiology*, 120(4),

- 1859–1871.
- Chavez, A. G., Franch, M., Mickiewicz, E. A., Baltazar, W., Belanger, J. L., Devara, D., Etta, M., Hamre, T., Ismail, T., Joiner, B., Kim, Y., Kona, A., Mansourian, K., Nangia, A., Pluenneke, M., Soubra, S., Venkateswaran, T., Venkudusamy, K., Chericoni, A., ... Hayden, B. Y. (2025). Mirror manifolds: partially overlapping neural subspaces for speaking and listening. In *bioRxiv.org* (p. 2025.09. 20.677504). <https://doi.org/10.1101/2025.09.20.677504>
- Chericoni, A., Diao, C., Yan, X., Ismail, T., Mickiewicz, E. A., Franch, M., Chavez, A. G., Paulo, D., Bartoli, E., Provenza, N. R., Yoo, S. B. M., Hennig, J., Jacobs, J., Hayden, B. Y., & Sheth, S. A. (2026). Neural geometry in the human hippocampus enables generalization across spatial position and gaze. In *arXiv [q-bio.NC]*. arXiv. <https://doi.org/10.48550/arXiv.2603.04747>
- Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405), 51–56.
- Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science (New York, N.Y.)*, 352(6292), 1464–1468.
- Courellis, H. S., Minxha, J., Cardenas, A. R., Kimmel, D. L., Reed, C. M., Valiante, T. A., Salzman, C. D., Mamelak, A. N., Fusi, S., & Rutishauser, U. (2024). Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632(8026), 841–849.
- Cunningham, J. P., & Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11), 1500–1509.
- Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9(2), 179–194.
- Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, 13(4), e1005508.
- Dijksterhuis, D. E., Self, M. W., Possel, J. K., Peters, J. C., van Straaten, E. C. W., Idema, S., Baaijen, J. C., van der Salm, S. M. A., Aarnoutse, E. J., van Klink, N. C. E., van Eijsden, P., Hanslmayr, S., Chelvarajah, R., Roux, F., Kolibius, L. D., Sawlani, V., Rollings, D. T., Dehaene, S., & Roelfsema, P. R. (2024). Pronouns reactivate conceptual representations in human hippocampal neurons. *Science (New York, N.Y.)*, 385(6716), 1478–1484.
- Eisenreich, B. R., Hayden, B. Y., & Zimmermann, J. (2019). Macaques are risk-averse in a freely moving foraging task. *Scientific reports*, 9(1), 15091.
- Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W., & Hayden, B. Y. (2019). Tonic exploration governs both flexibility and lapses. *PLoS computational biology*, 15(11), e1007475.
- Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, 109(19), 3055–3068.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition. In *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/2209.10652>

- Eliav, T., Maimon, S. R., Aljadeff, J., Tsodyks, M., Ginosar, G., Las, L., & Ulanovsky, N. (2021). Multiscale representation of very large environments in the hippocampus of flying bats. *Science (New York, N.Y.)*, 372(6545), eabg4020.
- Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., & Cunningham, J. P. (2016). Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications*, 7(1), 13239.
- Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *arXiv [cs.CL]*. arXiv. <https://doi.org/10.48550/arXiv.1909.00512>
- Fenton, A. A., Kao, H. Y., Neymotin, S. A., Olypher, A., Vayntrub, Y., Lytton, W. W., & Ludvig, N. (2008). Unmasking the CA1 ensemble place code by exposures to small and large environments: more place cells and multiple, irregularly-arranged, and expanded place fields in the larger space. *J Neurosci*, 28, 11250–1126.
- Franch, M., Mickiewicz, E. A., Belanger, J. L., Chericoni, A., Chavez, A. G., Katlowitz, K. A., Mathura, R., Paulo, D., Bartoli, E., Kemmer, S., Piantadosi, S. T., Provenza, N. R., Watrous, A. J., Sheth, S. A., & Hayden, B. Y. (2025). A vectorial code for semantics in human hippocampus. In *bioRxiv.org*. <https://doi.org/10.1101/2025.02.21.639601>
- Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37, 66–74.
- Ginosar, G., Aljadeff, J., Burak, Y., Sompolinsky, H., Las, L., & Ulanovsky, N. (2021). Locally ordered representation of 3D space in the entorhinal cortex. *Nature*, 596(7872), 404–409.
- Ginosar, G., Aljadeff, J., Las, L., Derdikman, D., & Ulanovsky, N. (2023). Are grid cells used for navigation? On local metrics, subjective spaces, and black holes. *Neuron*, 111(12), 1858–1875.
- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., ... Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369–380.
- Grieves, R. M., Jedidi-Ayoub, S., Mishchanchuk, K., Liu, A., Renaudineau, S., Duvelle, É., & Jeffery, K. J. (2021). Irregular distribution of grid cell firing fields in rats exploring a 3D volumetric space. *Nature Neuroscience*, 24(11), 1567–1573.
- Groppe, D. M., Bickel, S., Dykstra, A. R., Wang, X., Mégevand, P., Mercier, M. R., Lado, F. A., Mehta, A. D., & Honey, C. J. (2017). iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *Journal of Neuroscience Methods*, 281, 40–48.
- Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052), 801–806.
- Hayden, B. Y., Heilbronner, S. R., & Yoo, S. B. M. (2026). Rethinking the centrality of brain areas in understanding functional organization. *Nature Neuroscience*, 29(2), 267–278.
- Hayden, B. Y. (2019). Why has evolution not selected for perfect self-control?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1766).
- Hayman, R., Verriotes, M. A., Jovalekic, A., Fenton, A. A., & Jeffery, K. J. (2011). Anisotropic encoding of three-dimensional space by place cells and grid cells. *Nature Neuroscience*, 14(9), 1182–1188.
- Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural

- speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), 453–458.
- Jamali, M., Grannan, B., Cai, J., Khanna, A. R., Muñoz, W., Caprara, I., Paulk, A. C., Cash, S. S., Fedorenko, E., & Williams, Z. M. (2024). Semantic encoding during language comprehension at single-cell resolution. *Nature*, 631(8021), 610–616.
- Jazayeri, M., & Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. In *arXiv [q-bio.NC]*. arXiv. <http://arxiv.org/abs/2107.04084>
- Jenkinson, M., & Smith, S. M. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2), 143–156.
- Jenkinson, Mark, Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2), 825–841.
- Johnston, W. J., Fine, J. M., Yoo, S. B. M., Ebitz, R. B., & Hayden, B. Y. (2024). Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. *Nature Neuroscience*, 27(11), 2218–2230.
- Johnston, W. J., Palmer, S. E., & Freedman, D. J. (2020). Nonlinear mixed selectivity supports reliable neural computation. *PLoS Computational Biology*, 16(2), e1007544.
- Joshi, A., Scheinost, D., Okuda, H., Belhachemi, D., Murphy, I., Staib, L. H., & Papademetris, X. (2011). Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics*, 9(1), 69–84.
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/1607.01759>
- Kafashan, M., Jaffe, A. W., Chettih, S. N., Nogueira, R., Arandia-Romero, I., Harvey, C. D., Moreno-Bote, R., & Drugowitsch, J. (2021). Scaling of sensory information in large neural populations shows signatures of information-limiting correlations. *Nature Communications*, 12(1), 473.
- Karkowski, K., Kehl, M. S., Qin, Y., Darcher, A., Karkowski, P., Borger, V., Surges, R., Hebart, M. N., Reber, T. P., Dehaene, S., Lakretz, Y., & Mormann, F. (2025). Semantic Tuning of Single Neurons in the Human Medial Temporal Lobe. In *Neuroscience* (biorxiv;2025.10.21.682935v1). bioRxiv. <https://www.biorxiv.org/content/10.1101/2025.10.21.682935v1.full.pdf>
- Katlowitz, K. A., Cole, E. R., Mickiewicz, E. A., Shah, S., Franch, M. C., Adkinson, J., Belanger, J. L., Mathura, R. K., Meszéna, D., McGinley, M., Muñoz, W., Banks, G. P., Cash, S. S., Hsu, C.-W., Paulk, A. C., Provenza, N. R., Watrous, A., Williams, Z., Goldman, A. M., ... Sheth, S. A. (2025). Plasticity and Language in the Anesthetized Human Hippocampus. In *Neuroscience* (biorxiv;2025.04.09.648012v2). bioRxiv. <https://www.biorxiv.org/content/10.1101/2025.04.09.648012v2>
- Knudsen, E. B., & Wallis, J. D. (2021). Hippocampal neurons construct a map of an abstract value space. *Cell*, 184(18), 4640–4650.e10.
- Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9), 1148–1160.
- Kriegeskorte, Nikolaus, & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8), 401–412.
- Kriegeskorte, Nikolaus, Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2,

- 4.
- Krupic, J., Bauza, M., Burton, S., Barry, C., & O’Keefe, J. (2015). Grid cell symmetry is shaped by environmental geometry. *Nature*, 518(7538), 232–235.
- Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15(1), 5523.
- Langston, R. F., Ainge, J. A., Couey, J. J., Canto, C. B., Bjerknes, T. L., Witter, M. P., Moser, E. I., & Moser, M.-B. (2010). Development of the spatial representation system in the rat. *Science (New York, N.Y.)*, 328(5985), 1576–1580.
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791.
- Lehky, S. R., Sejnowski, T. J., & Desimone, R. (2005). Selectivity and sparseness in the responses of striate complex cells. *Vision Research*, 45(1), 57–73.
- Leutgeb, J. K., Leutgeb, S., Moser, M.-B., & Moser, E. I. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science (New York, N.Y.)*, 315(5814), 961–966.
- Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, 264, 47–56.
- Maaten, L., & Hinton, G. E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605.
- Magnotti, J. F., Wang, Z., & Beauchamp, M. S. (2020). RAVE: Comprehensive open-source software for reproducible analysis and visualization of intracranial EEG data. *NeuroImage*, 223(117341), 117341.
- Maisson, D. J. N., Cervera, R. L., Voloh, B., Conover, I., Zambre, M., Zimmermann, J., & Hayden, B. Y. (2023). Widespread coding of navigational variables in prefrontal cortex. *Current Biology*, 33(16), 3478–3488.
- Maisson, D. J. N., Cash-Padgett, T. V., Wang, M. Z., Hayden, B. Y., Heilbronner, S. R., & Zimmermann, J. (2021). Choice-relevant information transformation along a ventrodorsal axis in the medial prefrontal cortex. *Nature communications*, 12(1), 4830.
- McClelland, J. L., McNaughton, B. L., & O’Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 419–457.
- Mead, A. (1992). Review of the development of multidimensional scaling methods. *The Statistician: Journal of the Institute of Statisticians*, 41(1), 27.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/1301.3781>
- Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain’s spatial representation system. *Annual Review of Neuroscience*, 31(1), 69–89.
- Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4), e1003553.
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5(3). <https://doi.org/10.23915/distill.00024.001>

- Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, 2(11), e7.
- Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. *Nature Neuroscience*, 24(9), 1292–1301.
- Piantadosi, S. T., Muller, D. C. Y., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9), 844–856.
- Provenza, N. R., Reddy, S., Allam, A. K., Rajesh, S. V., Diab, N., Reyes, G., ... & Sheth, S. A. (2024). Disruption of neural periodicity predicts clinical response after deep brain stimulation for obsessive-compulsive disorder. *Nature Medicine*, 30(10), 3004–3014.
- Quian Quiroga, R. (2020). No pattern separation in the human hippocampus. *Trends in Cognitive Sciences*, 24(12), 994–1007.
- Quiroga, R. Quian, Mukamel, R., Isham, E. A., Malach, R., & Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), 3599–3604.
- Quiroga, R. Quian, Reddy, L., Koch, C., & Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of Neurophysiology*, 98(4), 1997–2007.
- Quiroga, R. Quian, Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045), 1102–1107.
- Quiroga, Rodrigo Quian. (2012). Concept cells: the building blocks of declarative memory functions. *Nature Reviews. Neuroscience*, 13(8), 587–597.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*.
- Rey, H. G., Panagiotaropoulos, T. I., Gutierrez, L., Chaure, F. J., Nasimbera, A., Cordisco, S., Nishida, F., Valentin, A., Alarcon, G., Richardson, M. P., Kochen, S., & Quian Quiroga, R. (2025). Lack of context modulation in human single neuron responses in the medial temporal lobe. *Cell Reports*, 44(1), 115218.
- Rich, P. D., Liaw, H.-P., & Lee, A. K. (2014). Place cells. Large environments reveal the statistical structure governing hippocampal representations. *Science (New York, N.Y.)*, 345(6198), 814–817.
- Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451), 585–590.
- Rolls, E. T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, 7, 74.
- Rolls, E., & Treves, A. (1997). *Neural networks and brain function*. Oxford University Press.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, 55, 103–111.
- Skaggs, W., McNaughton, B., & Gothard, K. (1992). An Information-Theoretic Approach to Deciphering the Hippocampal Code. In *Neural Information Processing Systems* (pp. 1030–1037).
- Skaggs, W., McNaughton, B., & Gothard, K. (1992). An information-theoretic approach to deciphering the hippocampal code. *Neural Information Processing Systems*, 1030–1037.
- Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M.-B., & Moser, E. I. (2012). The

- entorhinal grid map is discretized. *Nature*, 492(7427), 72–78.
- Suthana, N., Ekstrom, A. D., Yassa, M. A., & Stark, C. (2021). Pattern separation in the human hippocampus: Response to Quiroga. *Trends in Cognitive Sciences*, 25(6), 423–424.
- Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5), 858–866.
- Tang, W., Shin, J. D., & Jadhav, S. P. (2023). Geometric transformation of cognitive maps for generalization across hippocampal-prefrontal circuits. *Cell Reports*, 42(3), 112246.
- Tye, K. M., Miller, E. K., Taschbach, F. H., Benna, M. K., Rigotti, M., & Fusi, S. (2024). Mixed selectivity: Cellular computations for complexity. *Neuron*, 112(14), 2289–2303.
- Urai, A. E., Doiron, B., Leifer, A. M., & Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 25(1), 11–19.
- Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience*, 43(1), 249–275.
- Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 26(40), 10232–10234.
- Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5), 1249–1263.e23.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.
- Yacoub, E., Grier, M. D., Auerbach, E. J., Lagore, R. L., Harel, N., Adrian, G., ... & Zimmermann, J. (2020). Ultra-high field (10.5 T) resting state fMRI in the macaque. *Neuroimage*, 223, 117349.
- Yan, X., Krishna, A., Van Arsdell, K., Gautam, I., Kim, B., Shrivastava, A., Belanger, J., Chavez, A. G., Chericoni, A., Franch, M., Ismail, T., Katlowitz, K. A., Mickiewicz, E. A., Paulo, D., Zhu, H., Goldman, A. M., Krishnan, V., Maheshwari, A., Bartoli, E., ... Sheth, S. A. (2025). Shared neural geometries for bilingual semantic representations. In *bioRxiv.org* (p. 2025.11.16.688726). bioRxiv. <https://doi.org/10.1101/2025.11.16.688726>
- Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, 34(10), 515–525.
- Yoo, S. B. M., & Hayden, B. Y. (2020). The transition from evaluation to selection involves neural subspace reorganization in core reward regions. *Neuron*, 105(4), 712–724.
- Zada, Z., Nastase, S. A., Li, J., & Hasson, U. (2025). Brains and language models converge on a shared conceptual space across different languages. In *ArXiv*. arXiv. <https://www.ncbi.nlm.nih.gov/pubmed/41019213>
- Zhu, H., Franch, M., Mickiewicz, E. A., Belanger, J. L., Cowan, R. L., Katlowitz, K. A., Chavez, A. G., Chericoni, A., Paulo, D., Yan, X., Bartoli, E., Hennig, J. A., Provenza, N. R., Smith, E. H., Piantadosi, S. T., Sheth, S. A., & Hayden, B. Y. (2026). Semantic axes in the brain support analogical representations. In *bioRxiv.org*. bioRxiv. <https://doi.org/10.64898/2026.01.28.702241>

Supplementary Information

Polysemanticity in human hippocampal neurons

Xinyuan Yan¹, Ji-An Li², Melissa Franch¹, Hanlin Zhu¹, Rhiannon Cowan³, James Belanger¹, Ana G. Chavez¹, Assia Chericoni¹, Taha Ismail¹, Kalman A. Katlowitz¹, Luca D. Kolibius⁴, Elizabeth A. Mickiewicz¹, Danika Paulo¹, Eleonora Bartoli¹, Jay A. Hennig^{1,5,6,7}, Tomasz M. Frączek¹, Nicole R. Provenza^{1,5,6,7}, Shervin Rahimpour³, Ben Shofty³, Elliot Smith³, Joshua Jacobs⁴, Benjamin Y. Hayden^{*†1,5,6,7}, and Sameer A. Sheth^{*1,5,6,7}

1. Department of Neurosurgery, Baylor College of Medicine
2. Department of Psychology, New York University
3. Department of Neurosurgery, University of Utah
4. Department of Neurology, University of Chicago
5. Department of Electrical and Computer Engineering, Rice University
6. Department of Bioengineering, Rice University
7. Neuroengineering Initiative, Rice University

† **Correspondence:** Benjamin.Hayden@bcm.edu

* These authors contributed equally

Funding statement

This research was supported by the McNair Foundation and by NIH R01 MH129439, U01 NS121472, UE5NS070694-15, the SNS Allan Friedman RUNN Research Grant, the NLM Training Program in Biomedical Informatics & Data Science for Predoctoral & Postdoctoral Fellows, T15LM007093-33, Gordon and Mary Cain Pediatric Neurology Research Foundation

Competing interests

S.A.S has consulting agreements with Boston Scientific, Zimmer Biomet, Koh Young, Abbott, and Neuropace. SAS is Co-founder of Motif Neurotech.

Acknowledgements

We thank Joshua Adkinson, Justin Fine, Victoria Gates, Suzanne Kemmer, George Kokalas, and Steven Piantadosi for their assistance.

Supplementary Table

Table S1. NMF results in each patient.

patient	K(mode)	clustered	not clustered (dispersed or indistinguishable from chance)	% non- clustered
P01	5	0	5	100.0%
P02	7	1	6	85.7%
P03	9	0	9	100.0%
P04	14	0	14	100.0%
P05	11	0	11	100.0%
P06	14	0	14	100.0%
P07	4	2	2	50.0%
P08	8	1	7	87.5%
P09	11	0	11	100.0%
P10	12	0	12	100.0%
P11	13	0	13	100.0%
P12	9	1	8	88.9%
P13	12	0	12	100.0%
P14	11	0	11	100.0%
P15	13	2	11	84.6%

Table Notes. NMF results in each patient. For each patient, symmetric NMF was applied to the neuron-by-neuron tuning similarity matrix, and the optimal number of modes (K) was determined by the elbow method (see Methods). Each mode's top 100 words were tested for semantic clustering by comparing their mean pairwise distance against a null distribution of randomly sampled words (500 permutations, $p < 0.05$). Modes were classified as “clustered” (significantly smaller distance than chance) or “not clustered” (indistinguishable from or significantly greater than chance, i.e., dispersed). Across 15 patients, 10 showed no clustered modes (100% non-clustered), and the remaining 5 patients had at most 2 clustered modes out of their total.

Supplementary Figures

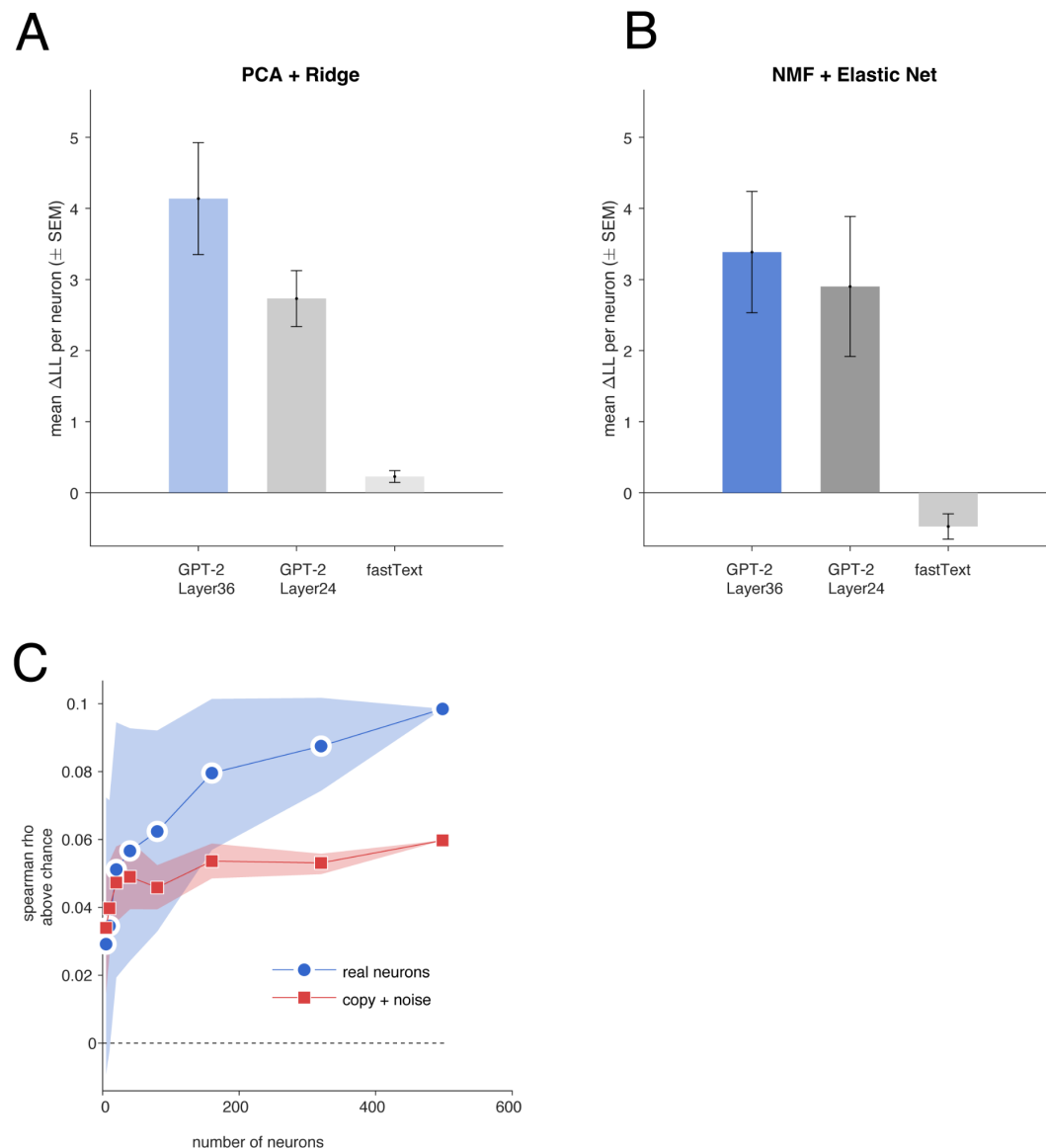


Figure S1. Control analyses for encoding pipeline and RSA scaling

(A, B) One potential limitation is that our encoding pipeline, PCA for dimensionality reduction paired with ridge regression, could favor distributed solutions by construction: PCA produces dense, signed components that spread variance across all dimensions, and ridge regression penalizes large weights, discouraging the sparse, high-magnitude tuning expected under a concept-cell scheme. To address this, we conducted control analyses using a sparse-friendly alternative: non-negative matrix factorization (NMF; Lee and Seung, 1999) for embedding decomposition, which yields parts-based, non-negative components that naturally accommodate sparse representations, paired with elastic net regression (L1 ratio = 0.9; see **Methods**), which drives irrelevant coefficients to exactly zero. Layer 36 again provided the strongest fit, and the distributed pipeline marginally outperformed the sparse alternative (paired t test, $t(497) = 1.81$, $p = 0.071$), indicating that the encoding results are not an artifact of pipeline choice but reflect the underlying structure of the neural code.

(C) RSA decimation scaling: real neurons vs. copy-plus-noise control. Spearman rank correlation (ρ above chance) between hippocampal and GPT-2 Layer 36 representational dissimilarity matrices (cosine distance) as a function of the number of neurons included in the population. For each population size $N \in \{5, 10, 20, 40, 80, 160, 320, 498\}$, random subsets of N neurons were drawn without replacement from the full population of 498 units, and the pairwise cosine distance matrix over noun tokens was computed and compared to the corresponding GPT-2 RDM via Spearman correlation. This procedure was repeated 200 times per population size. Chance level was estimated by a shuffle control (50 iterations per N) in which the mapping between words and neural firing rates was randomly permuted; plotted values reflect the mean ρ minus the shuffle mean at each N . Shaded regions indicate 95% confidence intervals (2.5th-97.5th percentiles) of the neuron-resampling distribution. The copy-plus-noise control (red) was constructed by identifying the single neuron with the highest cross-validated encoding performance (ΔLL), decomposing its activity into a signal component (ridge-predicted) and residual noise, and generating 498 synthetic neurons by adding independent Gaussian noise (variance matched to the empirical residuals) to copies of the signal. The same decimation procedure was then applied to this synthetic population (100 iterations per N). Real neurons (blue) show monotonically increasing RSA with population size, whereas the copy-plus-noise control plateaus, confirming that additional neurons carry non-redundant semantic information rather than simply averaging out shared noise. At $N = 498$, the confidence interval collapses to zero width because all neurons are included in every iteration, leaving no variability across resampled subsets.

Related to Figure 1.

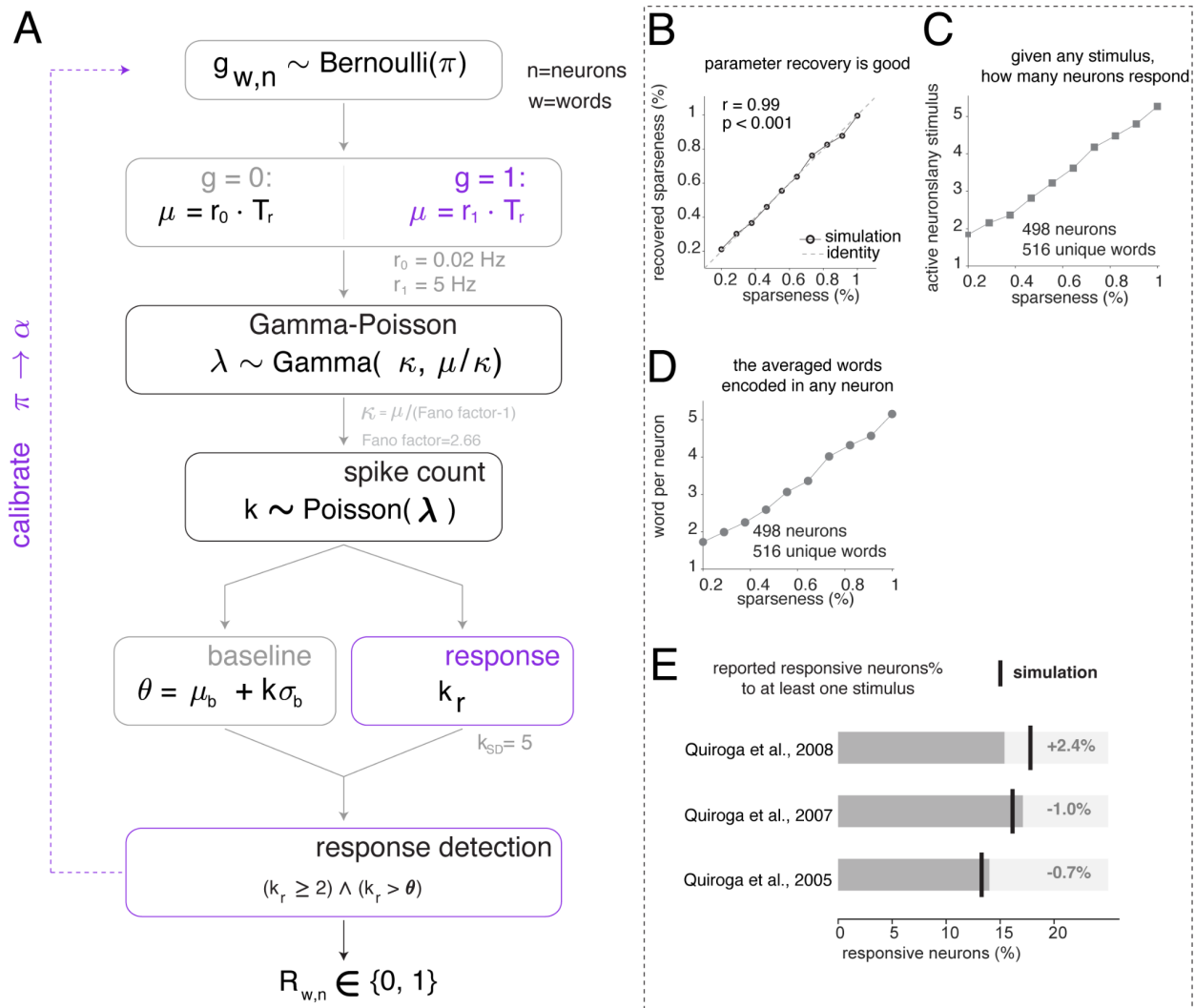


Figure S2. Concept cell simulation framework and validation results.

(A) Concept cell simulation framework (Method). To benchmark the observed sparseness against theoretical concept-cell predictions, we developed a generative model calibrated to published concept-cell parameters and naturalistic listening constraints. For each neuron-word pair, a Bernoulli gate (with probability π , the target sparseness) determines whether the neuron is “responsive” to that word. Responsive pairs fire at an elevated rate ($r_1 = 5$ Hz) and non-responsive pairs at baseline ($r_0 = 0.02$ Hz), with trial-to-trial variability generated by a Gamma-Poisson process (Fano factor = 2.66). Spike counts are converted to firing rates using the word’s duration, and a response is detected if the spike count exceeds a threshold defined as baseline mean plus k standard deviations ($k_{SD} = 5$). This yields a binary response matrix (words $\times$ neurons) from which sparseness metrics are computed.

(B) Simulation parameter recovery. To validate that the simulation accurately implements the intended sparseness level, recovered sparseness (measured from the simulated response matrix) is plotted against the input sparseness parameter (π) across a range of target values. The near-perfect correlation ($r = 0.99$, $p < 0.001$) confirms reliable parameter recovery.

(C) Number of responsive neurons per stimulus. To assess how population recruitment scales with coding sparseness, the average number of neurons responding to any given stimulus is plotted as a function of

target sparseness (498 neurons, 516 unique words). At the sparseness levels reported in the concept-cell literature, only 1-2 neurons respond per word, whereas real data show substantially broader recruitment. (D) Number of words encoded per neuron. To assess stimulus selectivity, the average number of words eliciting a response in any given neuron is plotted as a function of target sparseness. At concept-cell sparseness levels, each neuron responds to approximately 1 word, whereas real neurons respond to many more.

(E) Comparison with published concept-cell benchmarks. To anchor the simulation to the existing literature, the percentage of responsive neurons reported in three published concept-cell studies (horizontal bars) is shown alongside the simulation's predictions at corresponding sparseness levels. The vertical black line indicates the stimulus presentation threshold used in each study. The close agreement between published values and simulation output validates the model's calibration to the concept-cell literature.

Related to Figure 2.

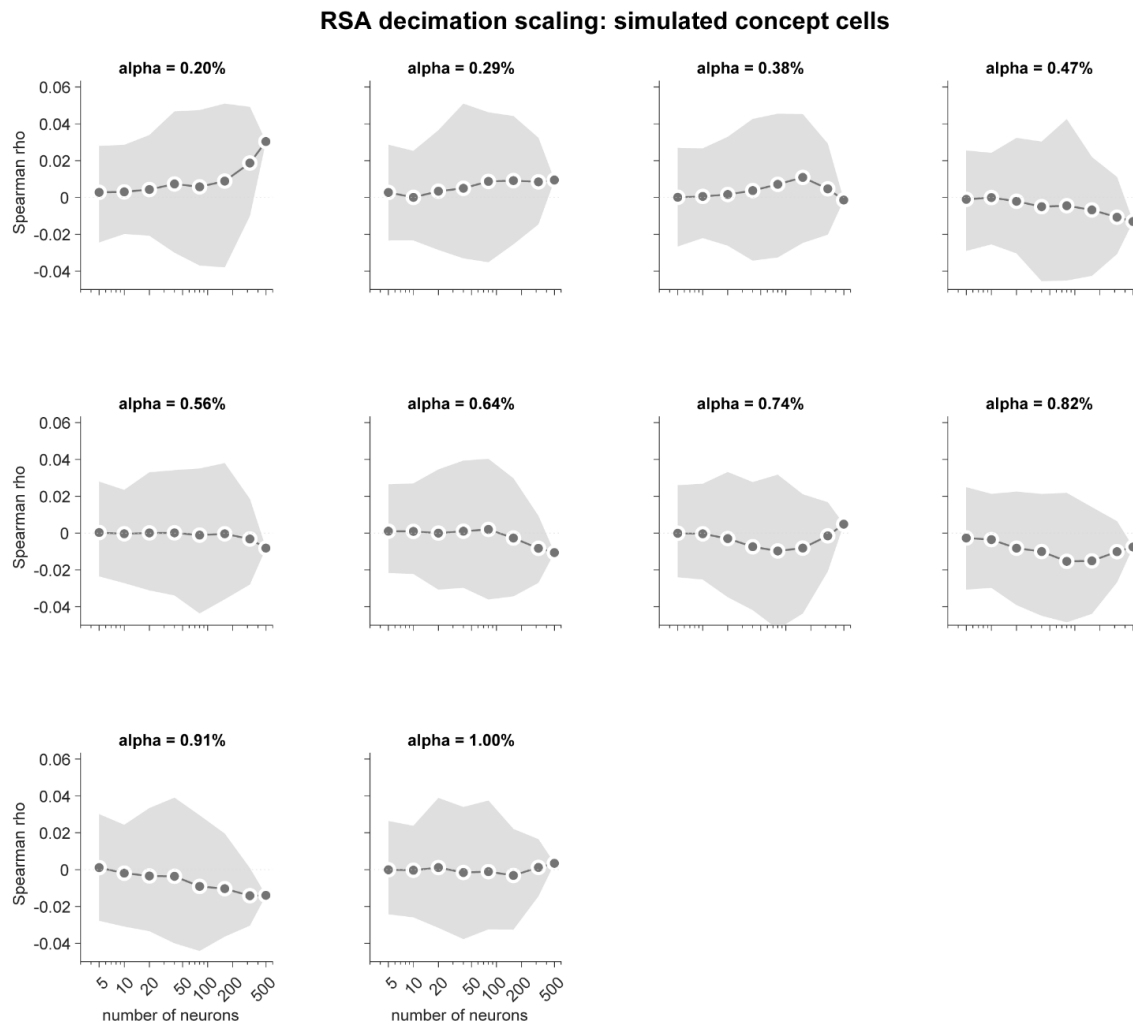


Figure S3. RSA decimation scaling for simulated concept cell populations. Each panel shows the Spearman rank correlation between the neural and GPT-2 Layer 36 pairwise cosine distance matrices as a function of the number of neurons included in the analysis, for simulated concept cell populations at a given sparseness level (response probability α (sparseness level) ranging from 0.20% to 1.00%). Grey lines and dots indicate mean correlation across 200 random subsamples; shaded regions indicate 95% confidence intervals. Across all sparseness levels and population sizes, simulated concept cells produce near-zero RSA correlations (rho range: -0.014 to 0.030 at $N = 498$), with no significant monotonically increasing trend as population size grows. This result indicates that sparse, concept-cell-like populations cannot reconstruct the continuous semantic geometry of the language model embedding space. Compare with the monotonically increasing RSA observed in the real hippocampal population (Figure 1H). Related to Figure 1.

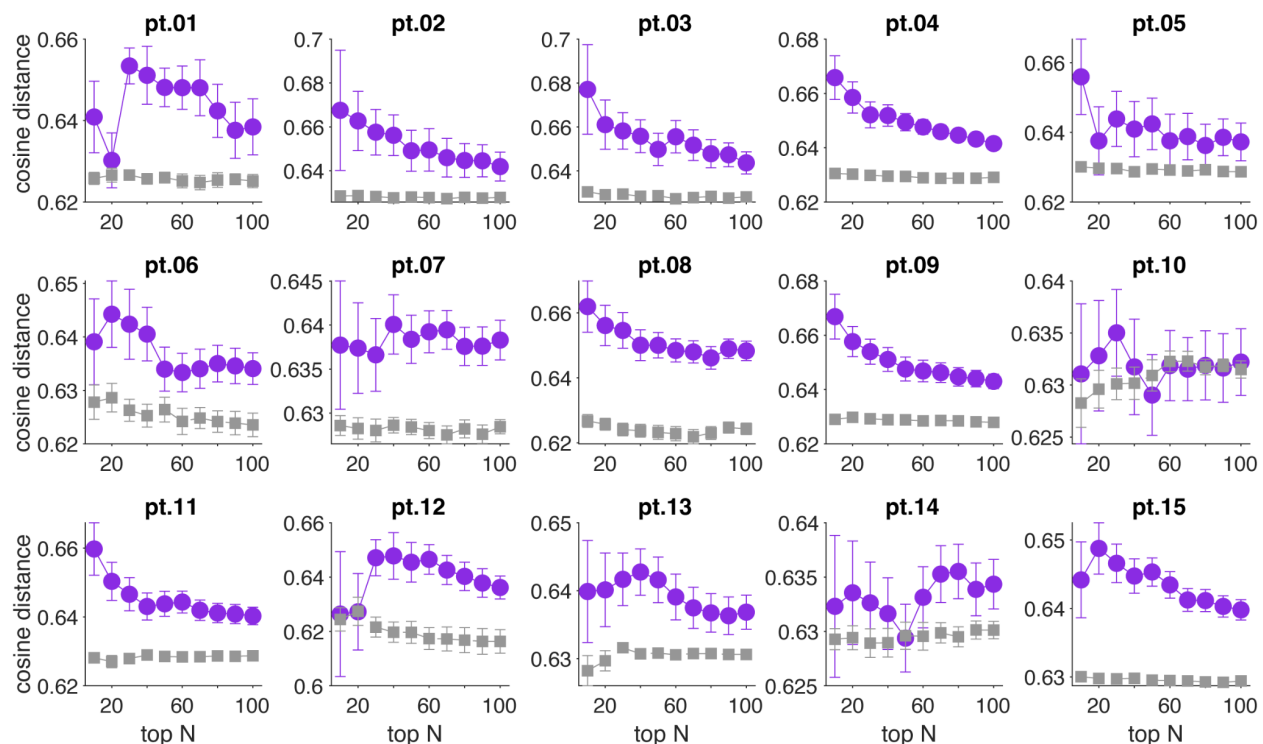


Figure S4. Mean cosine distance for top N words vs. random N words for individual patients

Per-patient replication of the group-level semantic dispersion analysis. For each patient, the mean pairwise cosine distance among each neuron's top N preferred words (purple) is plotted alongside the null expectation from random words (gray, mean of 500 permutations) across top N values (N = 10-100).

Error bars denote $\pm$ SEM across neurons. Results are broadly consistent with the group-level finding: in the majority of patients, top-preferred words exhibit greater semantic dispersion than random words, confirming that the preference for semantically diverse stimuli is not driven by a subset of patients but is a general property of hippocampal neurons across individuals.

Related to Figure 3.

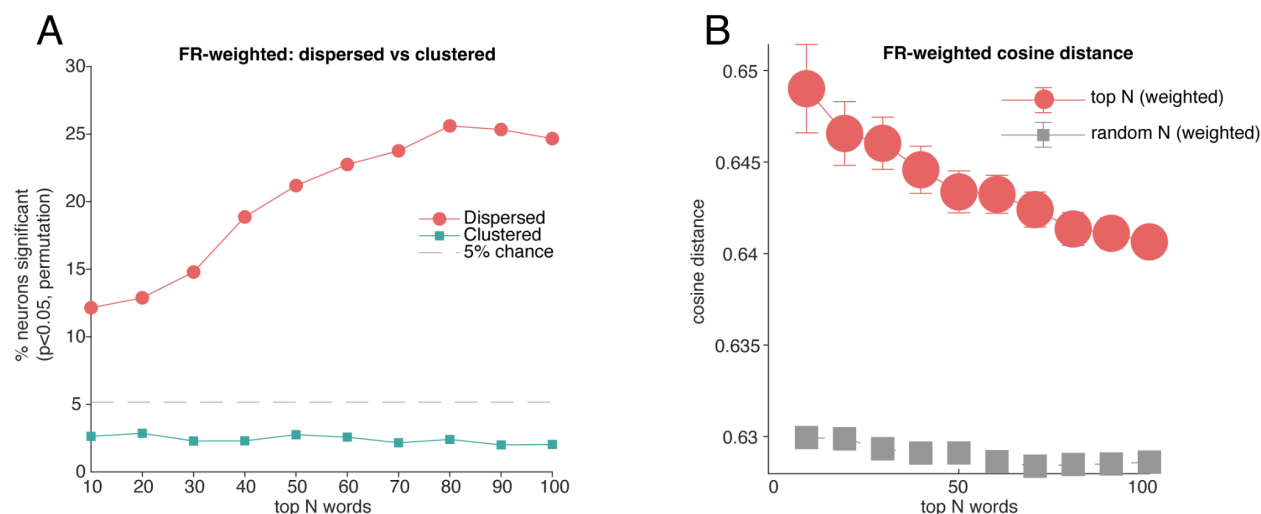


Figure S5. Results from the control analyses for top-N words dispersion.

The weighted analysis produced results nearly identical to the unweighted analysis.

(A) Under firing-rate weighting, the proportion of dispersed neurons (red line) substantially exceeds both the 5% chance level (gray line) and the proportion of clustered neurons at all values of N.

(B) At every value of N from 10 to 100, the FR-weighted mean pairwise distance among each neuron's top N words was significantly greater than the weighted null expectation (all $t > 7.87$, all $p < 0.001$).

Related to Figure 3.

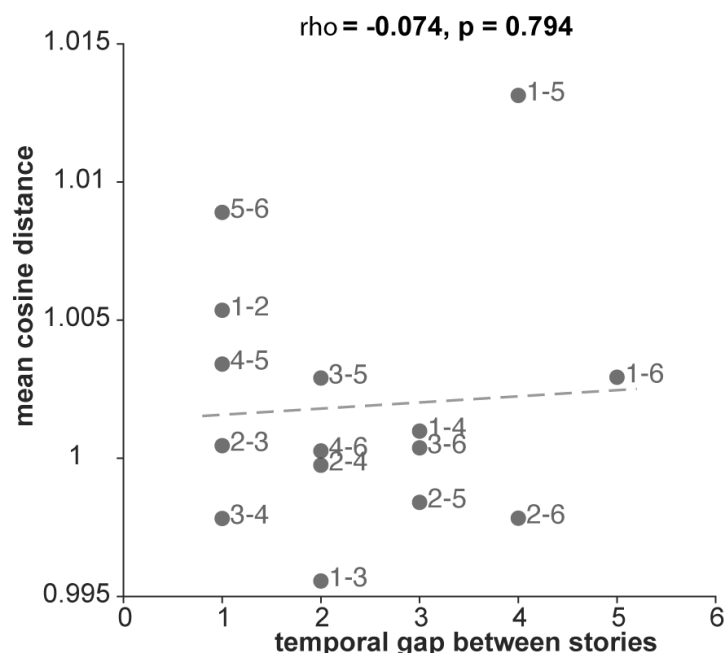


Figure S6. Across-story tuning distance is not explained by representational drift. Mean cosine distance between tuning vectors for each story pair ($n = 15$ pairs) plotted against the temporal gap between stories (ordinal distance). Tuning distance did not correlate with temporal separation (Spearman $\rho = -0.074$, $p = 0.794$), indicating that the increased across-story tuning distance (Figure 4A-C) reflects context-dependent semantic retuning rather than gradual drift in neural representations over time. Related to Figure 4.

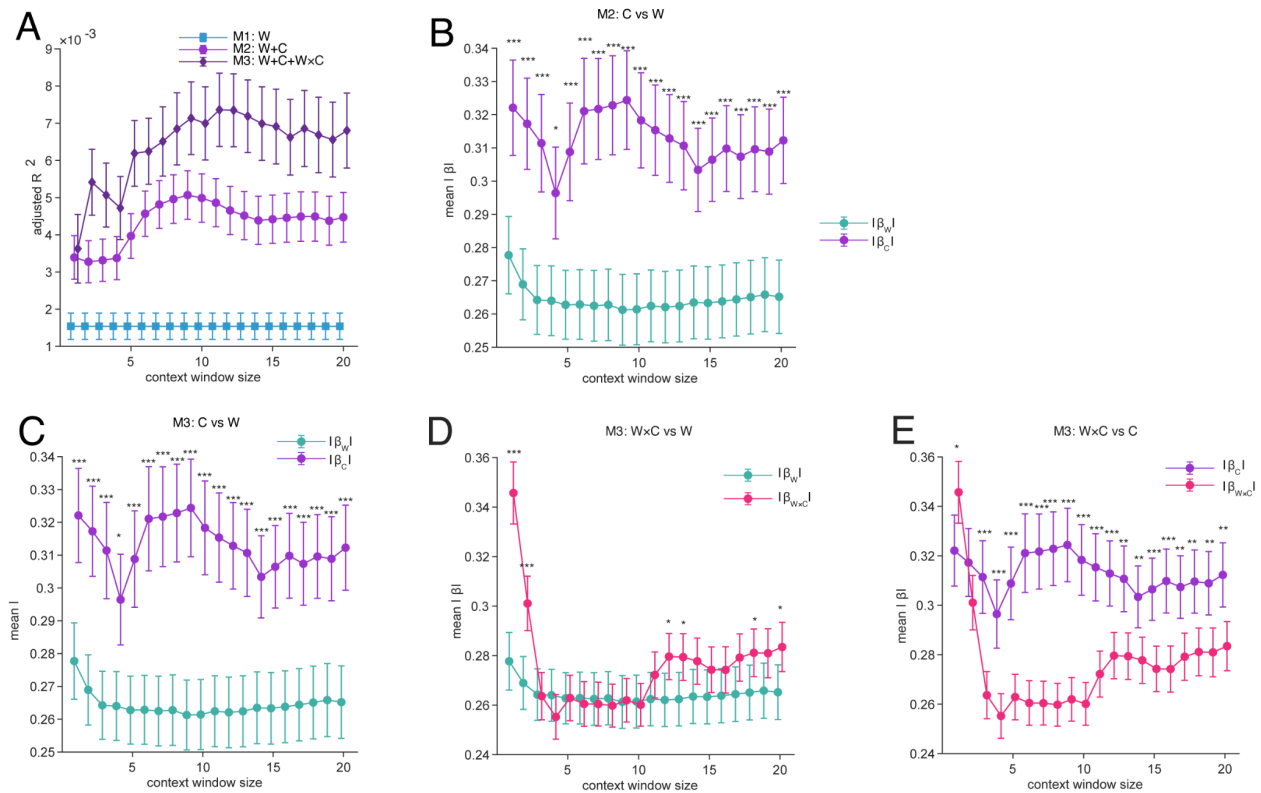


Figure S7. Lexical context modulation of semantic tuning across context window sizes.

(A) Mean adjusted R^2 ($\pm$ SEM) for three nested regression models across 498 hippocampal neurons as a function of context window size (1-20 preceding words). M1 (blue) includes word embeddings only; M2 (light purple) adds context embeddings; M3 (dark purple) further adds word $\times$ context interaction terms. Both M2 and M3 significantly outperform M1 at all window sizes (all $p < 0.001$), and M3 significantly outperforms M2 at 19 of 20 window sizes (all $p < 0.05$, except window size=1).

(B) Mean absolute standardized beta coefficients ($\pm$ SEM) for word ($|\beta_w|$, teal) and context ($|\beta_c|$, purple) predictors in M2. Context coefficients are significantly larger than word coefficients at all window sizes (all $p < 0.05$).

(C) Same comparison in M3. The pattern replicates: $|\beta_c| > |\beta_w|$ across all window sizes (all $p < 0.05$).

(D) Comparison of interaction ($|\beta_{w\times c}|$, pink) and word ($|\beta_w|$, teal) coefficients in M3. The two are comparable in magnitude at windows 3-11 (all $p > 0.05$), with interaction coefficients significantly exceeding word coefficients at larger windows (window size ≥ 19 and first 2 window sizes).

(E) Comparison of interaction ($|\beta_{w\times c}|$, pink) and context ($|\beta_c|$, purple) coefficients in M3. Context main-effect coefficients consistently exceed interaction coefficients at windows 3-20 (all $p < 0.05$). Together, these results establish a magnitude hierarchy of $|\beta_c| > |\beta_{w\times c}| \geq |\beta_w|$, indicating that local lexical context is the dominant driver of hippocampal firing rate variance, while its nonlinear modulation of word encoding constitutes a secondary but substantial influence that exceeds the contribution of static word identity alone across the big context window sizes.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (paired t-test, $n = 498$ neurons).

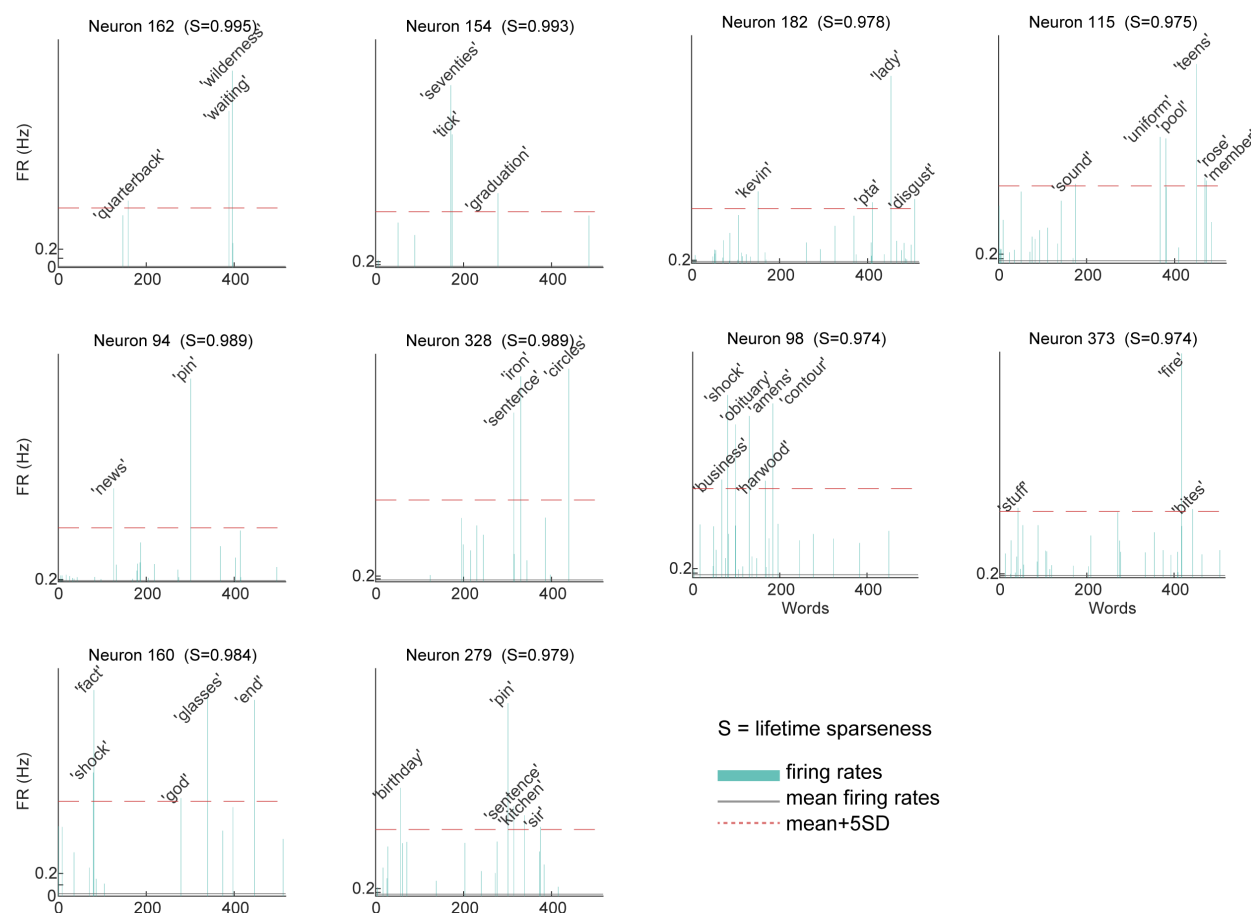


Figure S8. Concept-cell-like neurons identified by lifetime sparseness in language processing.

Firing rate profiles across all 516 unique nouns for the 10 neurons with the highest lifetime sparseness (top 2%, $S \geq 0.974$). Each panel shows one neuron's firing rate (teal bars) to each word in its original presentation order. The solid grey line indicates the mean firing rate and the dashed red line indicates the mean + 5 SD threshold. Words exceeding this threshold are labeled. These neurons exhibit sparseness values above 0.97, responding strongly to only 2-6 words out of 516. Notably, the words driving each neuron's responses span unrelated semantic categories (e.g., Neuron 162: "quarterback", "waiting", "wilderness"; Neuron 160: "shock", "fact", "god", "glasses", "end").

Related to Figure 2.

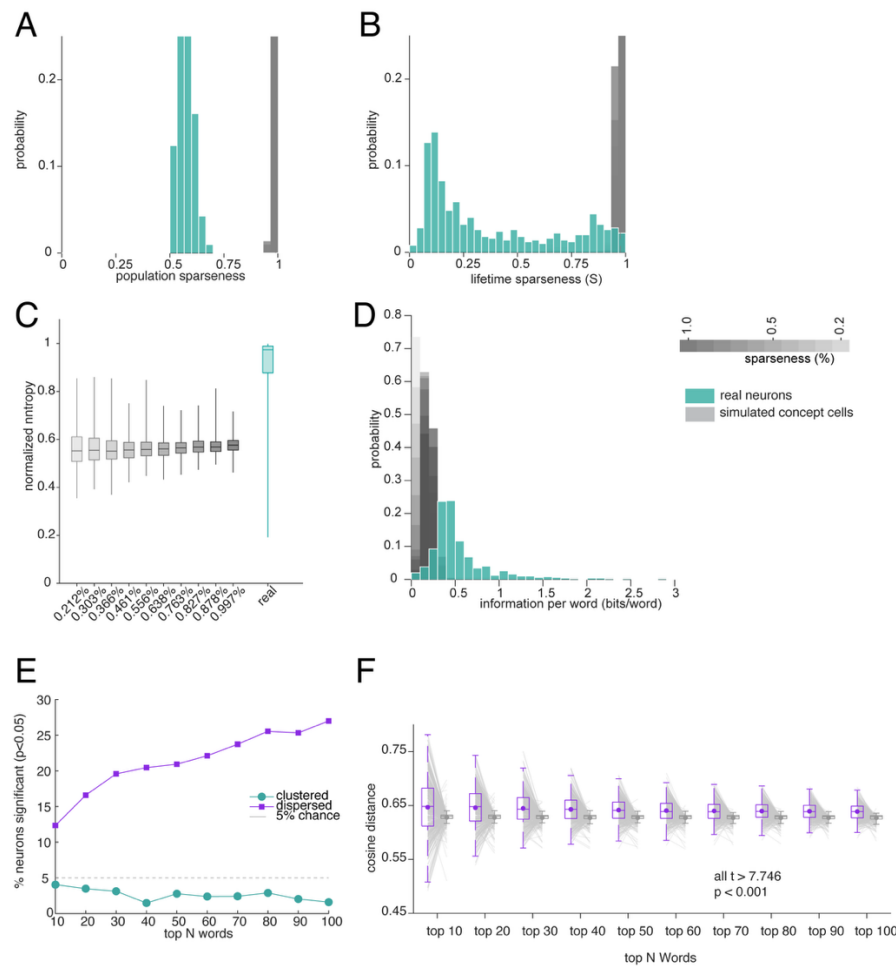


Figure S9. Dense coding and semantic overdispersion replicate with narrower response windows.

To control for the possibility that the 40 ms post-offset extension of our response window captures early responses to the subsequent word, thereby inflating apparent polysemanticity, we repeated key analyses to support polysemanticity using a narrower window (80 ms post-onset to word offset, with no post-offset extension).

(A) Population sparseness. Real neurons (teal) remain substantially denser than simulated concept cells (gray).

(B) Lifetime sparseness. Real neurons remain denser than simulated concept cells.

(C) Normalized entropy. Real neurons cluster at the high-entropy end, consistent with distributed coding.

(D) Information per word. Real neurons carry more information per word than simulated concept cells.

(E) Proportion of neurons showing significant dispersion vs. clustering across top N words.

Overdispersion dominates at all N values.

(F) Mean cosine distance for top N words vs. random N words. Top-preferred words remain significantly more dispersed than chance (all t > 7.746, p < 0.001).

Related to Figure 2 and Figure 3.