BACK TO THE FEATURE: TOWARD A FEATURE-CENTRIC ACCOUNT OF BRAIN–LM ALIGNMENT

Timna Wharton Kleinman
Department of Computer Science
The Hebrew University of Jerusalem, Israel
timna.kleinman@mail.huji.ac.il

Ariel Goldstein
Department of Cognitive and Brain Sciences & Business School
The Hebrew University of Jerusalem, Israel
Center for Human Inspired AI
University of Cambridge, UK
ariel.goldstein@mail.huji.ac.il

ABSTRACT

Large Language Models (LLMs) have emerged as powerful proxies for linguistic processing in the human brain, yet the standard practice of quantifying this alignment via a single scalar obscures its underlying drivers. This "scalar-centric" approach treats high-dimensional embeddings as atomic units, establishing that alignment exists while remaining agnostic to how it is achieved. Here, we study brain–LLM alignment during naturalistic language comprehension, using intracranial electrocorticography (ECoG) recordings from human participants listening to a spoken narrative. We propose a shift toward a feature-centric perspective that inspects the embedding dimensions contributing to neural alignment. We show that naive feature analyses suggest superficial homogeneity across cortical regions and time, but this reflects a methodological artifact of polysemantic representations. By combining representational disentanglement via Sparse Autoencoders (SAEs) with sparse encoding models (Lasso), we uncover distinct feature subsets that dissociate between cortical regions (superior temporal gyrus vs. inferior frontal gyrus) and temporal windows (pre- vs. post-onset). These results recast brain–LLM alignment not as aggregate similarity, but as a structural inquiry into which computational dimensions map onto distinct neural dynamics.

1 INTRODUCTION

While it is tempting to view words as static, atomic symbols, language unfolds across multiple levels of abstraction, timescales, and contexts, making simple definitions insufficient. In reality, a word is a complex, multidimensional entity whose interpretation depends not only on its local form, but on a dynamic integration of syntactic structure, discourse history, and broader world knowledge (Kintsch, 1988; Elman, 1990; Harris, 1954). This complexity places strong demands on any representational system, which must be capable of integrating these diverse information streams while preserving fine-grained distinctions in meaning. Large language models (LLMs) offer a solution by learning high-dimensional representations that jointly encode these elements (Sarzynska-Wawer et al., 2021; Vaswani et al., 2017; Brown et al., 2020; Team et al., 2024). Training on next-word (token) prediction, LLMs induce embeddings that reflect not only lexical identity but also syntactic expectations, semantic relations, and contextual constraints, providing a flexible representation of words as they appear in language (Hewitt & Manning, 2019; Tenney et al., 2019; Radford et al., 2019).

These learned representations have proven useful well beyond their original purpose (Wei et al., 2022; Bommasani et al., 2021; Doerig et al., 2025). Contextual word embeddings derived from LLMs have become the dominant representational tool for modeling linguistic meaning (Tao et al., 2024). These embeddings have proven particularly powerful as quantitative proxies for high-level linguistic representations in cognitive neuroscience, reliably accounting for neural responses to naturalistic language across recording techniques and experimental settings (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Tuckute et al., 2024). This alignment has been quantified using a family of representational comparison methods, including encoding models (Mitchell et al., 2008; Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2024a), representational similarity analysis (RSA) (Shoham et al., 2024; Kriegeskorte et al., 2008; Zhang et al.,

2025), centered kernel alignment (CKA) (Kornblith et al., 2019; Dapello et al., 2022; Han et al., 2023), and related approaches (Sucholutsky et al., 2023), which measure how well model-derived representations explain variance in neural activity. These metrics have been used to draw substantive conclusions about the human brain. For example, Goldstein et al. (2022) demonstrated that the brain and LLMs share fundamental computational principles, such as engaging in continuous next-word prediction before word onset and calculating post-onset surprise. Tuckute et al. (2024) moved beyond correlation to causation, demonstrating that these models can be used to non-invasively control neural activity by synthesizing sentences that precisely drive or suppress the human language network. Consequently, LLMs are increasingly treated as scientific instruments for probing the representational organization and semantics of language in the human cortex.

Despite their impact, these approaches, which measure similarity between LLM representations and brain signals, implicitly rely on a strong modeling assumption that is rarely examined. When an embedding predicts neural activity, it is not merely a methodological convenience; it constitutes a claim about how linguistic information is represented. In practice, however, most representational comparisons collapse this relationship into a single scalar - prediction accuracy, correlation, or other similarity scores, and proceed to draw neuroscientific conclusions from the magnitude or distribution of that score over time or space (e.g., Goldstein et al. (2022) use encoding models and interpret significant pre-onset correlations as evidence that the brain predicts the next word). These metrics are undoubtedly useful, but they collapse rich, high-dimensional interactions into summary statistics that can obscure the underlying drivers of alignment. Consequently, while these analyses establish that a given embedding space aligns with neural representations, they are largely agnostic to how these models achieve their apparent brain-likeness. We learn little about which aspects of the embedding space drive the observed similarity, or how different components contribute across brain regions or time. As a result, it remains unclear how model-derived representations map onto known functional distinctions in the language system. In effect, embeddings are treated as atomic objects rather than as high-dimensional models whose internal features, directions, and subspaces may carry distinct relationships to neural computation.

This lack of feature-level analysis becomes especially salient in light of the theoretical motivation for using linear, rather than deep or highly expressive, models in brain–language alignment. Linear models are preferred precisely because they are designed to capture explicit information in an embedding - information that is directly accessible in the representational geometry without requiring complex, nonlinear transformations to extract (Naselaris et al., 2011). This is in contrast to implicit information that may be present in a representation but requires a nonlinear decoder to access. A familiar example comes from vision: the identity of a person is embedded in the raw pixel (or retinal) array of an image, yet a linear classifier will struggle to extract it because the relevant structure is entangled with nuisance variation such as lighting, pose, and size. The use of linear methods in both neuroscience and machine learning (e.g., linear probes in interpretability research (Alain & Bengio, 2016; Park et al., 2023)) reflects a commitment to characterizing what information is explicitly represented, rather than what might be recovered through complex extraction (DiCarlo & Cox, 2007; Hewitt & Liang, 2019; Naselaris et al., 2011). If we adopt this principle seriously, then the standard practice of collapsing embeddings into scalar alignment scores is at odds with the very motivation for using linear methods - we should be examining which specific features in the embedding space contribute to brain alignment, not merely whether alignment exists in aggregate.

Prior work has taken important steps toward feature-level analyses (Singh et al., 2025). However, these approaches have typically relied on hand-crafted feature spaces, such as lexical, syntactic, or semantic annotations, chosen in advance (Wehbe et al., 2014; Binder et al., 2016; Brennan et al., 2016; Mitchell et al., 2008; Shimizu et al., 2025). While such features are readily interpretable, they necessarily reflect prior hypotheses about which aspects of language should matter for neural processing. Recent efforts have begun to relax this constraint by learning sparse, interpretable bases from data, yet these have been restricted to static, non-contextual embeddings (Zeng & Gallant, 2025). In contrast, we take a feature-based approach that works directly with the learned coordinates of contextual embeddings, treating them as an unsupervised set of candidate features. Rather than collapsing brain–model correspondence into a single score, we ask how alignment is distributed across embedding dimensions and which features drive prediction of neural activity. This lets us work in the model's representational space without committing to predefined linguistic categories.

We begin with a minimal descriptive analysis, assessing each embedding feature independently by testing its correlation to neural activity across electrodes and time. This quantifies how alignment is

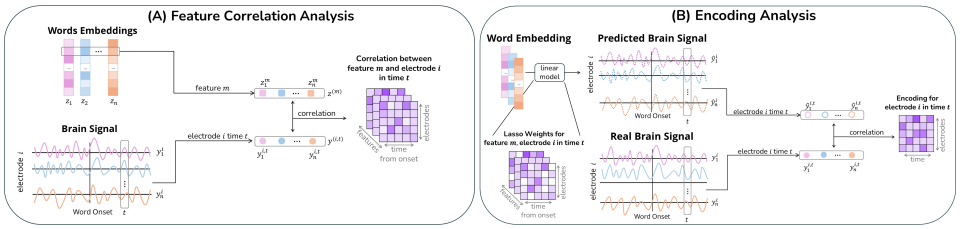

Figure 1: **Feature analysis**. **(A) Schematic of the feature correlation analysis.** Contextual embeddings are extracted for n words (top left) and aligned with ECoG neural responses relative to word onset (bottom left). For a given embedding dimension (feature) m the feature values are concatenated across words and correlated with the corresponding vector of neural responses for each electrode-time pair (i, t). Pearson correlations quantify how variation along individual embedding features covaries with neural activity, yielding a three-dimensional correlation tensor (purple right tensor) spanning features, electrodes, and time relative to word onset. **(B) Schematic of the encoding analysis.** A linear model maps high-dimensional word embeddings to the neural response at a specific electrode-time lag pair (i, t). The model generates a predicted brain signal $(\hat{y}^{i,t})$ for held-out data, which is correlated with the actual recorded neural activity $(y^{i,t})$ to quantify encoding accuracy (right purple heatmap). Lasso enforces sparsity, yielding a set of non-zero weights (purple bottom left) identifying a feature subset that drives the prediction.

distributed across features, and we show that the number of brain-correlated features closely tracks established encoding performance, linking predictive accuracy to the breadth of contributing dimensions. We then ask whether features that drive alignment differ across known functional distinctions in the language system, focusing on cortical regions (STG vs. IFG) and time relative to word onset.

While basic analyses suggest limited feature-level differentiation across these conditions,we hypothesize that this apparent homogeneity reflects limitations of the original coordinate system and analysis tools rather than true neural equivalence. We therefore adopt tools that (a) produce improved candidate embeddings, and (b) select a small subset of features that can explain neural activity. We use Sparse Autoencoders (SAEs) (Lieberum et al., 2024; Bricken et al., 2023) to address the first goal as they encourage sparse and disentangled representations of the original raw embeddings. Complementarily, we use L1-regularized regression (Lasso) (Tibshirani, 1996) to address the second goal by enforcing sparsity in the encoders weights, thereby isolating a small set of informative dimensions. With these, we observe systematic feature-level differentiation across brain regions and word-onset times that are not apparent in aggregate scores. Finally, we suggest a combined approach of SAEs and Lasso, providing us with principled way to ask which features matter for predicting neural responses, and whether those features change across conditions.

Crucially, our approach is composable with the broader literature on brain-LLM alignment: it can be layered on top of standard encoding pipelines and applied uniformly across architectures, layers, and context lengths. By allowing features to emerge in an unsupervised manner, it also enables the discovery of representational distinctions not presupposed by existing linguistic or neuroscientific theories, complementing, rather than replacing, hand-curated analyses.

2 Results

2.1 Experimental Setting

To examine how brain–model alignment is structured at the level of individual features, we focus on feature contributions in two complementary use-cases: (i) *spatial cortical regions,* comparing the superior temporal gyrus (STG) and the inferior frontal gyrus (IFG), and (ii) *temporal windows,* specifically pre- and post-onset intervals. This choice reflects established functional distinctions in the language system. STG and IFG are known to support partially dissociable stages of speech and language processing, with temporal regions more closely tied to acoustic–phonological analysis and frontal regions implicated in higher-order integration and control (Hickok & Poeppel, 2007; Friederici, 2011). Similarly, pre- and post-onset neural activity is associated with different computa-

tional demands, including anticipatory, predictive processes before onset versus stimulus-driven and integrative responses after it (Goldstein et al., 2022; Auksztulewicz et al., 2025). These contrasts provide natural test cases where brain–model alignment is expected to be present, but not necessarily identical. By asking whether features that contribute to alignment differ across spatial and temporal axes, we can assess whether embeddings capture distinctions beyond aggregate similarity and reflect known structure in the language network.

All analyses used intracranial electrocorticography (ECoG) recordings collected from neurosurgical patients listening to a 30-minute long spoken narrative (Podcast ECoG dataset; (Zada et al., 2025)). Neural activity was recorded from subdural electrode grids spanning large portions of the cortex, providing high temporal resolution and broad spatial coverage of language-responsive regions. Following prior work, we focused on a subset of 160 electrodes that exhibited significant word-level encoding using static GloVe embeddings, encompassing auditory, temporal, and frontal language areas. Neural signals were aligned to word onsets and analyzed across time. This design enabled comparisons across regions (STG vs. IFG) and temporal windows to assess how feature-level brain–model alignment varies during speech comprehension. Unless stated otherwise, we used all electrodes with significant encoding for GloVe embeddings ($n = 160$, (Goldstein et al., 2022)).

2.2 Motivation and Naive Approach

We begin with a naive approach, characterizing the relationship between individual embedding features and neural responses using a feature-wise correlation analysis. Here, we correlate each feature of gemma2-9b (Team et al., 2024) across all words with the brain signal of a specific time and electrode (Fig. 1A and Appendix). This results in a tensor of size electrodes × times × features, containing the correlation and significance of each feature. Results were corrected for multiple comparisons using false discovery rate (FDR, (Benjamini & Hochberg, 1995)) across all comparisons. Significance was assessed using q-values (FDR-adjusted p-values).

Number of Correlated Features. As a first step, we quantified the proportion of embedding features that showed a significant correlation ($q < .05$) with neural activity. Specifically, we traversed the correlation tensor and, for each electrode and time point, counted the number of features exhibiting a significant correlation. We observed a systematic temporal profile in the number of significant correlations (Fig. 2A; $r = .99$, $p < .001$), with counts increasing towards word onset, peaking shortly thereafter, and subsequently declining. This temporal profile closely tracked encoding performance and was consistent across distinct cortical regions, tracking encoding dynamics in both the inferior frontal gyrus (IFG; $r = .99$, $p < .001$) and the superior temporal gyrus (STG; $r = .97$, $p < .001$) (Fig. 2B). Moreover, across electrodes and time points, the number of correlated features was strongly associated with encoding accuracy (Fig. 2C; $r = .94$, $p < .001$). We further confirmed that this pattern was robust across different language models (Fig. S2; LLaMa: $r = .86$, $p < .001$, GPT2: $r = .90$, $p < .001$, GloVe: $r = .87$, $p < .001$).

This result suggests that stronger alignment between embeddings and neural activity (higher encoding) is associated with more embedding features showing significant correlations with the brain signal, motivating a closer look at feature-level structure.

Feature Identity. We next move beyond quantifying how many features are involved and ask whether the identities of those features differ across conditions. Specifically, we ask whether the same embedding features are implicated across brain regions (IFG vs. STG) and across temporal windows (pre-onset vs. post-onset), or whether distinct subsets of features contribute in each case.

We start by adapting a binary, set-membership definition of feature identity. For each comparison (brain area and time), we identify the set of features that exhibit significant correlations ($q < .05$) in each condition and ask which features appear exclusively in one condition versus the other. We visualize this overlap using Venn diagrams (Fig. 3A) and quantify it using Jaccard index (Jaccard, 1901; 1912) $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, i.e., Intersection-over-Union (IoU). Under this definition, we observe complete overlap in both comparisons ($J(\text{IFG, STG}) = 1$, $J(\text{pre-onset}, \text{post-onset}) = 1$). Meaning, the exact same set of features is implicated in IFG and STG, as well as before and after word onset, i.e., no feature appears uniquely in only one region or temporal window.

Because binary intersection is strictly conservative, we relaxed our definition to capture graded differences in engagement. Rather than evaluating whether a feature appears exclusively in one

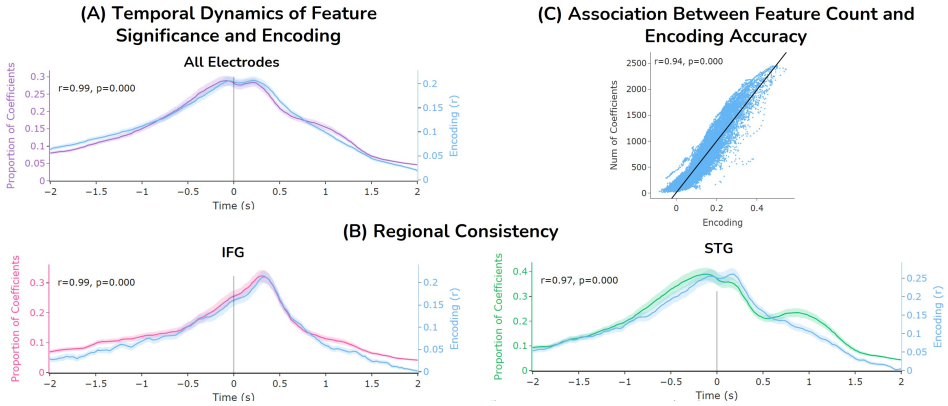

Figure 2: **Feature engagement tracks neural encoding.** Proportion of embedding features with significant correlation ($q < .05$) to neural activity. **(A) Whole-cortex temporal coupling.** Temporal trajectories of the average proportion of significant embedding features (purple, left axis) and average encoding accuracy (blue, right axis) across all electrodes ($n = 160$) relative to word onset. The density of significant features closely tracks the model's predictive performance over time. Shaded areas indicate standard error. **(B) Regional consistency.** This coupling is robust across distinct cortical regions. The proportion of significant features in the IFG (left, pink, $n = 53$) and the STG (right, green, $n = 47$) tracks the encoding accuracy (blue, left and right) in each respective area. Curves represent the average across all electrodes within each anatomical region. Shaded areas indicate standard error. **(C) Overall coupling.** Relation between the number of significant features and encoding accuracy across all electrode-time pairs. Stronger alignment is associated with a broader recruitment of embedding features rather than a sparse set of predictors.

condition, we quantified its frequency of appearance. Formally, for feature m and condition C (e.g., STG or pre-onset), we define engagement frequency $\nu_{m,c}$ as the proportion of electrode-time pairs showing significant ($q < .05$) relationship to neural activity:

$$\nu_{m,c} = \frac{1}{|\mathcal{P}_C|} \sum_{(i,t)\in\mathcal{P}_C} \mathbb{I}(q_{m,i,t} < 0.05)$$

where $\mathcal{P}_C$ denotes the set of all valid electrode-time pairs (i, t) in condition C, $|\mathcal{P}_C|$ is the cardinality of that set, and $\mathbb{I}(\cdot)$ is the indicator function. This yields a coordinate pair $(\nu_{m,c_1}, \nu_{m,c_2})$ for each feature, allowing for a direct comparison of utilization rates across experimental conditions.

To summarize comparisons with a single condition-selectivity score, we compute the 1D Canberra distance $d(x, y) = \frac{|x-y|}{|x|+|y|}$ (Lance & Williams, 1967) which quantifies the feature's bias toward one condition, or the deviation from the identity line ($\nu_{m,c_1} = \nu_{m,c_2}$) in the frequency space (Fig. S5).

We choose the Canberra distance as it accounts for the higher variance inherent in larger proportions by weighting differences at the lower end of the scale more heavily (e.g., the distance between 0.10 and 0.13 yields a higher score than the distance between 0.9 and 0.93). Moreover, the resulting score lies in $[0, 1]$, simplifying interpretation and comparison. A value of 0 indicates equal frequencies across conditions (no preference), while 1 corresponds to maximal bias toward a single condition. Notably, features appearing exclusively in one condition, i.e., corresponding to a unique Venn region, naturally receive the maximal score of 1.

Applying this frequency-based analysis yields a more nuanced conclusion than the strict categorical comparison (Figs. 4 and S6). We find that the distance scores are significantly greater than zero for both time-based contrasts ($M = 0.17$, $SD = 0.14$, 95% CI $[0.16, 0.17]$, $t(3583) = 71.08$, $p < .001$; Fig. 4A) and brain-region contrasts ($M = 0.23$, $SD = 0.12$, 95% CI $[0.22, 0.23]$, $t(3583) = 113.77$, $p < .001$; Fig. 4B). This indicates that even in the basic embedding space, feature-frequency patterns vary across conditions. However, while these deviations are statistically robust, their magnitude remains modest, suggesting that subtle condition-specific preferences coexist with a dominant pattern of broad, shared recruitment across the language network and times.

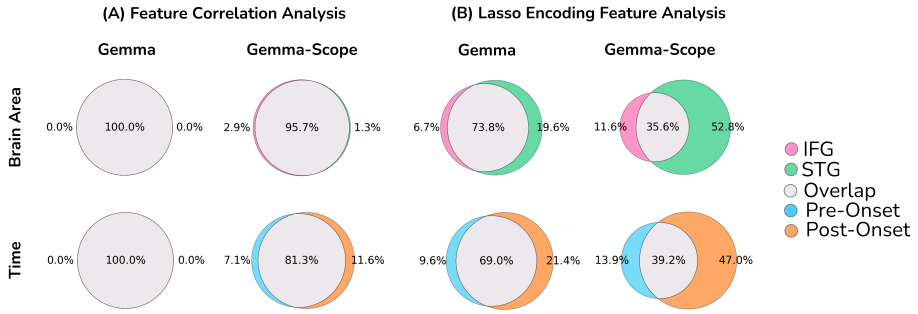

Figure 3: **Disentangling feature identity across space and time.** Overlap of embedding features between cortical regions (top row; IFG vs. STG) and temporal windows (bottom row; pre- vs. post-onset). Features absent from both conditions are omitted (total proportions in Fig. S4). **(A) Feature Correlation Analysis.** A feature is included if it shows a significant correlation ($q < .05$, FDR-corrected) with neural activity in at least one electrode-time pair, and survives a permutation test (see Appendix). While the base model (Gemma, left) shows complete overlap, SAE (Gemma-Scope, right) introduces some separation. **(B) Lasso Encoding Analysis.** A feature is included if Lasso assigned it a non-zero coefficient. While the base model (Gemma, left) retains substantial overlap, Lasso-enforced sparsity on disentangled features (Gemma-Scope, right) reveals more distinct feature sets which drive encoding in different brain areas and time windows.

2.3 Enhancing Feature Separation

Although our preliminary analyses suggest the conditions are difficult to distinguish at the feature level, we hypothesize that this reflects limitations of the tools used rather than genuine equivalence. We therefore introduce two complementary approaches: (i) sparse autoencoders (SAEs) to disentangle latent structure within embeddings, (ii) sparse encoding models using Lasso regression to isolate informative features. These techniques uncover previously hidden distinctions, with a combined approach yielding the highest degree of discriminability.

Feature Sparsity with Sparse Autoencoders (SAEs). The extensive feature overlap observed in the preceding analyses, where the same embedding dimensions appear to relate to neural activity across brain regions and temporal windows, can be explained by properties of the embedding space itself rather than true neural equivalence. Individual dimensions in modern LLM embeddings are widely understood to be polysemantic, often encoding multiple, unrelated concepts simultaneously (Elhage et al., 2022; Arora et al., 2018). Consequently, these models rely on distributed representations, where meaningful linguistic variables are spread across feature populations rather than localized to single units (Park et al., 2023; Rumelhart et al., 1986). Under these conditions, analyses that operate directly on the native embedding coordinates may obscure present but entangled distinctions.

To address this limitation, we use sparse autoencoders (SAEs) to reparameterize the embedding space into a more disentangled feature basis. SAEs learn a transformation of the original representations (here, embeddings from a fixed model layer), under a sparsity constraint that encourages latent units to be selectively active (Elhage et al., 2022; Lieberum et al., 2024; Templeton et al., 2024). This encourages individual features to capture more specific, approximately monosemantic factors of variation, reducing polysemantic mixing across coordinates and yielding a feature basis that can make it easier to isolate which directions are most relevant for predicting neural responses.

To ensure a controlled comparison, all SAE analyses use Gemma-Scope-9B-16k-residual (Lieberum et al., 2024) representations, derived from the same layer and context length as the original Gemma2-9B embeddings, with the SAE applied directly on top of the residual stream.

We first verify that SAE-derived features retain a meaningful relationship to neural activity. Encoding models built on SAE representations achieve reliably above-zero prediction performance within the $[-0.5, 0.5]$ time range (Fig. 5; one-sided one-sample t-test against 0: $M = 0.13$, $SD = 0.02$, 95% CI $[0.12, 0.14]$, $n = 41$, $t(40) = 46.84$, $p < .001$, Cohen's $d = 7.32$). At the same time, and consistent with the expected cost of sparsity and dimensional reparameterization (Lieberum et al., 2024; Karvonen et al., 2025), SAE-based ridge models exhibit a significant reduction relative to

ridge models trained on the original embeddings (paired one-sided t-test across times - $n = 161$, $t(160) = 31.95$, $p < .001$, Cohen's $d = 2.52$). Crucially, however, the persistence of robust above-zero encoding indicates that SAE features preserve brain-relevant information, justifying their use as an alternative representational basis for feature-level analysis.

Second, as in the original embedding space, the number of SAE features that exhibit significant correlations with neural activity closely tracks encoding performance over time ($r = .98$, $p < .001$), and over all time-electrode pairs ($r = .84$, $p < .001$; Fig. S3). This mirrors the relationship observed in the raw embeddings, suggesting that the coupling between predictive accuracy and the breadth of contributing features is kept under the SAE transformation.

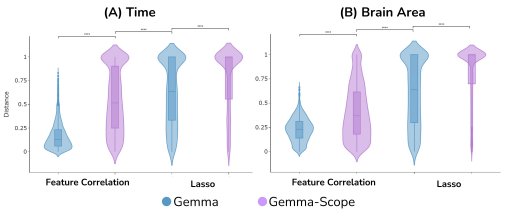

Figure 4: **Feature-level bias across space and time.** Distribution of bias scores (Canberra distance between engagement frequencies) for each feature when contrasting **(A) Temporal** (Pre- vs. Post-onset) and **(B) Cortical** (IFG vs. STG) Separation. Bias scores are shown for two feature-selection approaches - feature correlation (left columns) and Lasso encoding (right columns), and for two representational bases - Gemma (blue) and SAE-derived Gemma-Scope (purple). Higher scores indicate stronger condition-selectivity (0 = equal engagement; 1 = complete exclusivity). Boxplots show median and IQR. Asterisks indicate Bonferroni-corrected significance from one-sided Welch's t-tests (**** $p < .0001$).

Critically, analyses of feature identity in the SAE space reveal a structure that was absent in the original basis. Venn diagram comparisons demonstrate partial separation between features associated with IFG versus STG ($J(\text{IFG, STG}) = 0.957$), as well as between pre- and post-onset temporal windows ($J(\text{pre-onset, post-onset}) = 0.813$; Fig. 3). Unlike the raw embeddings, where feature sets overlapped completely across conditions, the SAE space contains features that are selectively associated with specific brain regions or temporal regimes. This provides direct support for the hypothesis that latent representational structure differs across space and time, but is obscured in the original coordinate system.

Frequency-based analyses further reinforce this dissociation. Many SAE features, while present in both conditions, exhibit systematic biases toward one temporal window ($M = 0.55$, $SD = 0.34$) or brain region ($M = 0.41$, $SD = 0.28$). Crucially, these condition-specific preferences were significantly more pronounced than those observed in the native embedding space (Time: Welch's $t = 76.13$; Brain: Welch's $t = 44.65$; all $p < .001$, Bonferroni corrected), reflected by a higher distance from equal engagement across conditions in frequency space (Fig. 4). This pattern indicates that SAE features capture graded, condition-specific preferences rather than merely introducing binary distinctions.

SAEs are often motivated not only by disentanglement but also by the prospect of automated feature interpretation via activation-based analyses. In principle, such interpretability could provide valuable insight into the linguistic or cognitive content of individual features. In practice, however, out-of-the-box feature interpretations did not yield linguistically or neuroscientifically meaningful structure in our setting. For example, one of features with the highest distance score in the STG was interpreted as "mentions of saints—especially the words and abbreviations for 'Saint'—in names of people, places, or institutions," with an interpretation score of 100. This is consistent with known limitations of SAE interpretability, which depend strongly on the choice of training data (Binder et al., 2016; Templeton et al., 2024). We attempted to improve interpretability by generating feature explanations using data thematically matched to the stimulus material (more "This American Life" podcast (Mao et al., 2020) episodes), but these efforts did not yield clear or stable semantic labels.

Taken together, these results highlight an important distinction between two related aspects of SAEs: disentanglement and interpretability. Although interpretability would enrich the neuroscientific conclusions, we show that disentanglement alone already provides significant value. By reparameterizing the embedding space into a sparser and more factorized basis, SAEs expose condition-specific structure that is otherwise inaccessible, even in the absence of human-interpretable labels.

Sparse Encoding using Lasso Regression. While SAEs address limitations arising from the representational basis itself, complementary insights can be gained by imposing sparsity directly at the level of the encoding model. To this end, we next turn to encoding using Lasso regression (Fig. 1B; Tibshirani (1996)), which adds an L_1 regularization to the encoding model. Unlike ridge regression, which distributes explanatory power across many features, Lasso yields encoding models in which only a subset of features is assigned non-zero weights, that jointly explain neural variance. This makes it a natural tool for identifying candidate features that are most informative for predicting brain responses. Here, a chosen feature is defined not via a significance test on correlation, but by whether a feature is actively selected by the encoding model, that is, whether its weight is non-zero.

We first verified that enforcing sparsity does not come at the cost of predictive validity. Indeed, encoding models trained with Lasso achieve performance reliably above-zero within the $[-0.5, 0.5]$ time range (Fig. 5; one-sided one-sample t-test against 0: $M = 0.17$, $SD = 0.02$, 95% CI $[0.16, 0.17]$, $n = 41$, $t(40) = 63.99$, $p < .001$, Cohen's $d = 9.99$). Critically, it demonstrates that much of the predictive signal can be captured by relatively few dimensions, motivating closer examination of the selected weights themselves. Notably, the Lasso time course also shows a sharper return to baseline: encoding drops to approximately zero by 2s post-onset, in contrast to ridge models that tend to retain a small positive tail. This temporal specificity aligns with the expectation that word-locked information should not persist indefinitely, and it reduces concern that sustained above-zero accuracy reflects slow drifts or incidental correlations rather than stimulus-related signal. Together, these observations reinforce our shift from maximizing prediction scores toward a modeling-oriented analysis that interrogates the learned weight structure — i.e., which features are selected and when they contribute.

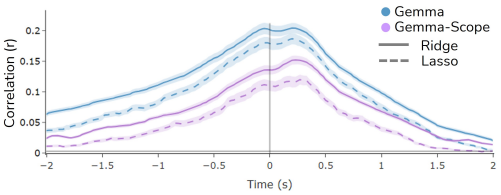

Figure 5: **Comparing encoding.** Average encoding accuracy across all electrodes relative to word onset ($t = 0$). Encoding performances are compared between original embeddings (Gemma, blue) versus disentangled features (Gemma-Scope, purple), and dense (Ridge, solid) versus sparse (Lasso, dashed) regression models. Although enforcing sparsity results in reduced predictive magnitude, temporal dynamics remain consistent. Shaded areas indicate standard error.

A known limitation of Lasso is its potential instability with correlated predictors, where selection among highly correlated features can be arbitrary. We therefore assessed the degree of correlation among embedding features in our data. Feature correlations were low (mean absolute correlation 0.011 for Gemma features, see Fig. S7), reducing concern that sparsity patterns are driven by collinearity rather than signal. To further ensure robustness, we adopted a conservative cross-validation strategy: Lasso models were trained across 10 folds, and only features that were selected consistently (i.e., assigned non-zero weights in all folds) were retained. This procedure prioritizes stability over completeness, yielding a core set of reliably informative features.

Applying this sparse encoding framework brings out patterns that are not readily apparent in naive feature-wise correlation analyses. Feature sets selected by Lasso differ systematically across brain regions (IFG vs. STG; $J(\text{IFG, STG}) = 0.738$) and across temporal windows (pre- vs. post-onset; $J(\text{pre-onset, post-onset}) = 0.69$), with reduced overlap relative to the complete overlap observed previously (Fig. 3B). Beyond strict feature uniqueness, frequency-based analyses show a shift away from condition balance (i.e., equal-frequency) both in the temporal setting ($M = 0.62$, $SD = 0.34$) and in the spatial setting ($M = 0.61$, $SD = 0.34$), indicating that selected features tend to be recruited more frequently in one condition than in the other (Fig. 4). This separation is significantly better (closer to 1) compared to the feature-correlation analysis using gemma (Time: Welch's $t = 70.04$; Brain: Welch's $t = 62.12$; all $p < .001$, Bonferroni corrected). This increased separation in frequency space demonstrates that sparse encoding not only improves categorical separability but also amplifies graded condition-specific biases that were previously obscured.

Taken together, these results show that enforcing sparsity at the level of the encoding model exposes condition-specific structure that is invisible to univariate correlation-based analyses. By isolating a stable subset of informative features, Lasso reveals meaningful differences in how embedding features relate to neural activity across cortical regions and over time, without sacrificing predictive

power. These findings reinforce the interpretation that the apparent feature homogeneity observed earlier arises from methodological smoothing rather than true representational equivalence, and they establish sparse encoding as a principled tool for feature-level analysis in brain–language alignment.

Taking the Best of Both Worlds. The preceding analyses demonstrate that SAEs and Lasso address complementary limitations. SAEs reparameterize the representational basis, reducing polysemantic entanglement among features, while Lasso enforces sparsity at the level of the encoding model, isolating the subset of features that are actually used to predict neural activity. We therefore examine their combination, applying sparse encoding to SAE-derived features.

Here, too, we begin by verifying the existence of significant encoding to validate the use of the proposed combination. Indeed, we get a significant above-zero encoding within the [-0.5, 0.5] time range (Fig. 5; one-sided one-sample t-test against 0: $M = 0.10$, $SD = 0.02$, 95% CI $[0.10, 0.11]$, $n = 41$, $t(40) = 40.91$, $p < .001$, Cohen's $d = 6.39$).

This combined approach yields the strongest separation across brain regions ($J(\text{IFG}, \text{STG}) = 0.356$) and temporal windows ($J(\text{pre-onset}, \text{post-onset}) = 0.392$; Fig. 3B), relative to using either method alone. Frequency-based analyses show the largest distance between features in different conditions (Time: $M = 0.78$, $SD = 0.33$; Brain Region: $M = 0.82$, $SD = 0.29$), receiving significantly higher distances compared to using Lasso with Gemma (Time: Welch's $t = 18.06$; Brain Region: Welch's $t = 26.25$; all $p < .001$, Bonferroni corrected; Fig. 4). In total, this approach yields both sharper categorical distinctions and stronger graded biases. These results suggest that disentangling the feature space and enforcing sparsity at the encoding level act synergistically to expose structure that remains partially obscured when only one constraint is applied.

Importantly, the stability assumptions underlying Lasso are well satisfied in the SAE space. Correlations between gemma-scope features are low (mean absolute correlation = 0.006, Fig. S7), further reducing concerns about correlated predictors and reinforcing the robustness of the selected features. As in our earlier SAE analyses, however, this increased structural clarity does not translate into straightforward semantic interpretability: even when generating automatic interpretations using the podcast dataset itself, we did not recover linguistically or neuroscientifically meaningful labels.

These results indicate that combining SAEs with sparse encoding provides the most effective lens for revealing condition-specific structure in brain–embedding alignment. By jointly leveraging representational reparameterization and feature selection, this approach exposes distinctions that are invisible to correlation-based analyses and difficult to recover with either method in isolation.

3 Discussion

In this work, we move beyond aggregate measures of brain-model alignment to examine how alignment is distributed across features. Rather than treating embeddings as monolithic objects whose relationship to neural activity can be summarized by a scalar, we adopt a feature-centric perspective to identify which components drive neural prediction across cortical regions and time. This reveals structural granularity in brain-language alignment often obscured by standard alignment metrics alone (Mitchell et al., 2008; Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2024a).

Our analyses demonstrate three main findings. First, brain-model alignment strength tracks feature breadth: stronger encoding is associated with a larger number of features correlating with neural activity. Second, in the model's native coordinates, feature identities appear largely invariant across regions and time, suggesting superficial homogeneity. Third, this homogeneity arises from limitations of the representational basis and analysis tools rather than the neural data (Arora et al., 2018; Elhage et al., 2022; Park et al., 2023; Rumelhart et al., 1986). By reparameterizing embeddings using sparse autoencoders (Lieberum et al., 2024; Templeton et al., 2024) and enforcing sparsity of the encoding model via Lasso regression (Tibshirani, 1996), we uncover systematic, condition-specific feature structure across cortical regions (STG vs. IFG) and temporal epochs (pre vs. post).

Importantly, the pre-post split exposes a qualitative shift that can't be explained as simply "more of the same" information once the word arrives. In particular, we observe pre-onset structure that is not straightforwardly recoverable from the post-onset feature set, implying that anticipatory or predictive activity is not merely a weaker version of stimulus-evoked next-word, but recruits representational ingredients that are at least partly distinct. This aligns with prior evidence for pre-

dictive processing in language (Lioubashevski et al., 2025) while sharpening the claim: predictive and stimulus-driven codes may differ in what they represent, not only in when they appear or how strongly they are expressed.

Our approach differs from prior feature-level analyses that rely on hand-crafted annotations or predefined features (Wehbe et al., 2014; Binder et al., 2016; Brennan et al., 2016; Mitchell et al., 2008; Shimizu et al., 2025). While interpretable, these methods are constrained by the researcher's hypotheses about which linguistic aspects matter. By operating directly on the learned coordinates of embeddings, our framework allows features to emerge unsupervised, enabling the discovery of distinctions not prespecified by theory. We emphasize that the goal is not to replace hand-curated analyses, but to complement them with tools probing representational structure at scale.

A related question is why analyze embedding features rather than focusing directly on differences in the neural signal itself. A key advantage of our framework is that it isolates variance in neural activity specifically related to linguistic representations, rather than reflecting broader differences in signal dynamics, anatomy, or recording conditions (Naselaris et al., 2011; Zada et al., 2023). By anchoring analyses in a shared space, we can compare how language-related information is encoded across regions and time in a way that is difficult to achieve from neural data alone.

Several limitations are worth noting. SAEs face challenges, including instability across training runs or datasets (Paulo & Belrose, 2025), and the fact that supervised tools (e.g., linear probes optimized for specific annotations) often outperform them on targeted benchmarks (Heindrich et al., 2025). These limitations matter for many use cases but are less central here. Our primary objective is not to recover a unique or canonical decomposition, nor to optimize performance on predefined labels; rather, it is to test whether some disentangled reparameterization reveals structure obscured in the original basis. From this perspective, instability is a practical drawback, but it does not undermine the core observation that even partial disentanglement can enable feature-level separation. Likewise, stronger supervised tools are largely orthogonal: our framework's value lies in surfacing distinctions unsupervised, without prespecifying dimensions along which representations should differ.

Lasso regression introduces its own trade-offs. As an L_1-regularized estimator, it can reduce encoding performance relative to denser models by discarding weak-but-consistent predictors. However, encoding remains significant, and for our objective — characterizing features implicated across regions and time - sparse solutions better match the inference problem (Naselaris et al., 2011). Lasso yields compact candidate sets comparable across conditions, allowing feature identities (and their overlap or separation) to become explicit objects of analysis rather than washed out in distributed weight patterns. We mitigated instability by verifying low feature correlations and retaining only features selected across all cross-validation folds. Nonetheless, this likely underestimates the full set of relevant features, and future work should explore complementary sparsity-inducing or Bayesian approaches that better represent uncertainty in feature selection while preserving interpretability.

Our work opens several avenues for future research. One extension is examining how feature structure differs for probable versus improbable words (Goldstein et al., 2024b), testing whether prediction-related features engage selectively under varying surprisal. More broadly, these analyses could be applied across tasks, e.g., reading versus listening, modalities (Doerig et al., 2025), or production versus comprehension, to assess the extent to which feature–brain mappings generalize. At a larger scale, encoding models themselves could be compared across individuals or tasks, shifting emphasis from prediction accuracy to the structure of learned weights. Finally, integrating this framework with advances in mechanistic interpretability, such as more stable or semantically grounded feature discovery methods, may eventually allow feature-level neuroscientific conclusions to be linked more directly to interpretable linguistic or cognitive constructs (similar to the disentangled semantic axes recovered from static embeddings Zeng & Gallant (2025)).

Taken together, our results underscore a gap in brain–language modeling: alignment is often summarized by a single scalar, masking the underlying linguistic drivers and their functional dynamics. Our feature-centric approach turns contextual embeddings into analyzable units - components that can be localized, compared across conditions, and tracked across electrode-time samples. This reframes "brain-likeness" as a set of mechanistic inquiries about which dimensions matter in which neural circuits and at which moments. Ultimately, this framework moves beyond aggregate similarity to offer a more fine-grained accounting of how representational structure maps onto cortical dynamics during language processing.

ACKNOWLEDGMENTS

We thank the Hebrew University Starting Grant for support. We also thank Eliahu Horwitz for invaluable assistance, and Daria Lioubashevski and Miriam Havin for helpful feedback.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Ryszard Auksztulewicz, Ozan Bahattin Ödül, Saskia Helbling, Ana Böke, Drew Cappotto, Dan Luo, Jan Schnupp, and Lucía Melloni. "what" and "when" predictions jointly modulate speech processing. *Journal of Neuroscience*, 45(20), 2025.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174, 2016.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Geary, Michael Ferguson, David D Cox, and James J DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. *BioRxiv*, pp. 2022–07, 2022.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8):1220–1234, 2025.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/toy_model/index.html`.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011.

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

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A Nastase, Zaid Zada, Eric Ham, Amir Feder, et al. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature communications*, 15(1):2768, 2024a.

Ariel Goldstein, Haocheng Wang, Tom Sheffer, Mariano Schain, Zaid Zada, Leonard Niekerken, Bobbi Aubrey, Samuel A Nastase, Harshvardhan Gazula, Colton Casto, et al. Information-making processes in the speaker's brain drive human conversations. *bioRxiv*, pp. 2024–08, 2024b.

Yena Han, Tomaso A Poggio, and Brian Cheung. System identification of neural systems: If we got it right, would we know? In *International conference on machine learning*, pp. 12430–12444. PMLR, 2023.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Lovis Heindrich, Philip Torr, Fazl Barez, and Veronika Thost. Do sparse autoencoders generalize? a case study of answerability. *arXiv preprint arXiv:2502.19964*, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

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

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint arXiv:2503.09532*, 2025.

Walter Kintsch. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2):163, 1988.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Godfrey N Lance and William T Williams. Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*, 1(1):15–20, 1967.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Daria Lioubashevski, Daniel Friedman, Adeen Flinker, and Ariel Goldstein. The road not taken: Unsaid word alternatives are represented in the brain. *bioRxiv*, pp. 2025–11, 2025.

Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison Cottrell. Speech recognition and multi-speaker diarization of long conversations. *arXiv preprint arXiv:2005.08072*, 2020.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

David E Rumelhart, James L McClelland, PDP Research Group, et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry research*, 304:114135, 2021.

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

Riki Shimizu, Richard J Antonello, Chandan Singh, and Nima Mesgarani. Interpretable embeddings of speech enhance and explain brain encoding performance of audio models. *arXiv preprint arXiv:2507.16080*, 2025.

Adva Shoham, Rotem Broday-Dvir, Itay Yaron, Galit Yovel, and Rafael Malach. Text-related functionality and dynamics of visual human pre-frontal activations revealed through neural network convergence. *bioRxiv*, pp. 2024–04, 2024.

Chandan Singh, Richard J. Antonello, Sihang Guo, Gavin Mischler, Jianfeng Gao, Nima Mesgarani, and Alexander G. Huth. Evaluating scientific theories as predictive models in language neuroscience. *bioRxiv: The Preprint Server for Biology*, pp. 2025.08.12.669958, August 2025. ISSN 2692-8205. doi: 10.1101/2025.08.12.669958.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. Llms are also effective embedding models: An in-depth overview. *arXiv preprint arXiv:2412.12591*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Zaid Zada, Ariel Goldstein, Sebastian Michelmann, Erez Simony, Amy Price, Liat Hasenfratz, Emily Barham, Asieh Zadbood, Werner Doyle, Daniel Friedman, et al. A shared linguistic space for transmitting our thoughts from brain to brain in natural conversations. *BioRxiv*, 2023.

Zaid Zada, Samuel A. Nastase, Bobbi Aubrey, Itamar Jalon, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Sasha Devore, Adeen Flinker, Orrin Devinsky, Ariel Goldstein, and Uri Hasson. The "podcast" ecog dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12 (1):1135, 2025.

Alicia Zeng and Jack Gallant. Disentangling superpositions: Interpretable brain encoding model with sparse concept atoms. *bioRxiv*, pp. 2025–11, 2025.

Zhejun Zhang, Shaoting Guo, Wenqing Zhou, Yingying Luo, Yingqi Zhu, Lin Zhang, and Lei Li. Brain-model neural similarity reveals abstractive summarization performance. *Scientific Reports*, 15(1):370, 2025.

A APPENDIX

A.1 CONTEXTUAL EMBEDDING EXTRACTION

For our analysis, we extracted contextualized word embeddings from Gemma 2 9B and from Gemma Scope 16k Residual, using Gemma 2 9B as the base model. For Gemma, we relied on the pretrained implementation available through the Hugging Face framework. Words from the raw transcript, including punctuation and capitalization, were first tokenized into whole-word or subword units. We then applied a sliding window of 32 tokens, advancing one token at a time, and extracted the embedding corresponding to the final word in each sequence.

The resulting tokens were encoded as integer indices and passed through the model, yielding hidden-state activations at each network layer. Although Gemma 2 9B comprises 48 layers, we restricted our analysis to the middle layer. While the target token was the final word in the sequence, we used the hidden state of the penultimate token, as this representation is the one used by the model to predict the final word.

After obtaining embeddings for all words in the raw transcript, we normalized them (per feature) and aligned them with the spoken-word transcript (see Fig. S1), which excludes punctuation and retains only complete words. We then trained the encoding model using 10-fold cross-validation, selecting the regularization strength via cross-validation within each training split (i.e., an inner cross-validation) for each electrode and time point.

For Gemma Scope, we followed an analogous pipeline, extracting activations from the same layer used for Gemma.

A.2 PATIENT METADATA AND ELECTRODE LOCATIONS

The Podcast ECoG dataset (Zada et al., 2025) comprises 9 participants with 1,330 electrodes recorded during clinical intracranial monitoring for epilepsy at NYU Langone Medical Center. Electrode placement was determined solely by clinical need, without regard to any research protocol, and electrode localization was performed by co-registering pre- and post-surgical T1-weighted MRI scans to the MNI152 template. Of the 1,330 total electrodes, we focused on the 160 that exhibited significant word-level encoding using static GloVe embeddings, spanning auditory, temporal, and frontal language areas. Of these, 53 were localized to the IFG (from 5 different patients) and 47 to the STG (from 6 different patients). Because electrode implantation was clinically driven, IFG and STG electrodes are not uniformly distributed across patients; some participants contribute electrodes to both regions, while others contribute to only one. In total, 4 patients have electrodes in both IFG and STG.

A.3 FEATURE CORRELATION ANALYSIS

In this analysis, for each word $k \in \{1, ..., n\}$ in the podcast, we extracted a contextual embedding $z_k \in \mathbb{R}^M$ from a language model, where M denotes the embedding dimensionality. In parallel, we aligned the ECoG signal to word onset and extracted neural responses, yielding a scalar of brain signal response $y_k^{i,t}$ for every word k, electrode i, and time t relative to word onset.

For a given feature (embedding dimension) m, we formed the vector $z^{(m)} = [z_1^m, ..., z_n^m]$ across words. Likewise, for each electrode–time pair (i, t) we formed the neural response vector $y^{(i,t)} = [y_1^{i,t}, ..., y_n^{i,t}]$. We then computed the Pearson correlation $r_{i,t,m} = corr(z^{(m)}, y^{(i,t)})$ quantifying the degree to which variation along feature m covaries with neural activity at electrode i and time t. Repeating this procedure for all electrodes, time points, and features yielded a three-dimensional correlation tensor of size electrodes×times×features.

Statistical significance was assessed independently for each element of this tensor. Correlation values were converted to p-values using a standard parametric test, and multiple comparisons were controlled using false discovery rate (FDR) correction across all electrode–time–feature combinations. Unless otherwise noted, we report only correlations that survived FDR correction at $q < 0.05$.

A.4 Venn Diagrams Permutation Test

For each comparison between conditions C_1 and C_2 (e.g., IFG vs. STG, or pre- vs. post-onset), the Venn analysis yields a set of presumably unique features - those that appear in exactly one condition. To validate that these features reflect a reliable condition difference rather than sampling noise, we apply a nonparametric permutation test to each feature m.

Let $\mathcal{P}_C$ denote the set of valid electrode-time pairs (i, t) in condition C. For a given feature m, we extract its per-sample correlation $r_{m,i,t}$ across all $(i, t) \in \mathcal{P}_{C_1} \cup \mathcal{P}_{C_2}$. We then compute the observed difference in mean feature values between conditions:

$$\Delta_m \;=\; \frac{1}{|\mathcal{P}_{C_1}|} \sum_{(i,t)\in\mathcal{P}_{C_1}} r_{m,i,t} \;-\; \frac{1}{|\mathcal{P}_{C_2}|} \sum_{(i,t)\in\mathcal{P}_{C_2}} r_{m,i,t} = \bar{r}_m^{C1} - \bar{r}_m^{C2} \tag{1}$$

To obtain a null distribution under the hypothesis that feature values are exchangeable across conditions, we pool all values $\{r_{m,i,t}\}$ across $\mathcal{P}_{C_1} \cup \mathcal{P}_{C_2}$, randomly permute them $10,000$ times, and each time split the permuted vector back into two groups of sizes $|\mathcal{P}_{C_1}|$ and $|\mathcal{P}_{C_2}|$. For each permutation b, we recompute $\Delta_m^{(b)}$, producing a null distribution $\{\Delta_m^{(b)}\}_{b=1}^{N}$. A two-tailed permutation p-value is computed as:

$$p_m \;=\; \frac{1}{N} \sum_{b=1}^{N} \mathbb{I}\Big(\Big|\Delta_m^{(b)}\Big| \geq |\Delta_m|\Big)\,. \tag{2}$$

Finally, we retain a feature m in the "unique-to-C" set only if it shows evidence of a condition difference under this test ($p < .05$); otherwise, it is removed from the corresponding unique set. This filtering step ensures that Venn-unique features reflect not only binary appearance under thresholding, but also a statistically supported separation in their average engagement/value across electrode-time samples in the two conditions.

A.5 Permutation Test for the Jaccard Index

To verify that the observed Jaccard indices reflect genuine condition differences rather than stochastic properties of the pipeline, we constructed a null distribution for J via condition-label permutation. For each comparison (brain area: IFG vs. STG; temporal window: pre- vs. post-onset), condition labels were permuted 10,000 times and J recomputed after each shuffle. Permutation was performed at the appropriate unit level: for the brain-area comparison, labels were shuffled across electrodes, preserving within-electrode coefficient structure; for the temporal comparison, labels were shuffled across time bins, preserving within-bin structure across electrodes. This ensures the null distribution captures aggregate pipeline behavior under label exchangeability, rather than merely disrupting low-level correlation structure. A one-tailed p-value was computed as the proportion of null J values falling at or below the observed J, testing whether conditions are more distinct than chance. For all the Lasso comparisons, the observed J fell well below the null distribution ($p < .001$), confirming that the reported feature separation exceeds what the pipeline produces under the null hypothesis and cannot be attributed to L1-induced selection instability or SAE stochasticity alone.

A.6 Statistical Assessment of Feature Bias

To quantify the degree of condition-specificity, we computed the bias score (Canberra distance) for each feature. To assess whether these scores significantly exceeded the null hypothesis of equal engagement (a score of zero), we performed one-sided one-sample t-tests. We estimated 95% confidence intervals (CI) for the mean bias scores using bias-corrected and accelerated (BCa) bootstrapping with 10,000 resamples (Efron, 1987). All statistical analyses were performed using the SciPy library in Python.

Comparisons between conditions (e.g., Correlation-Gemma vs. Correlation-Gemma-Scope) were performed using one-sided Welch's independent samples t-tests (Welch, 1947) to account for unequal variances between distributions. p-values for these pairwise comparisons were corrected for multiple comparisons using the Bonferroni method.

A.7 Statement on the Use of LLMs

Adhering to conference disclosure policies, we note the use of OpenAI ChatGPT 5.2, Google Gemini 3 Pro, and Claude Opus 4.5 in the preparation of this manuscript. Our use of these tools was strictly limited to grammatical correction and enhancing the text's readability. We confirm that the scientific substance, including research ideation, contribution development, experimental design, data analysis, and result interpretation, originated solely from the authors. We have manually verified all LLM-edited text to ensure accuracy and originality, and we retain full responsibility for the final content. No LLM is listed as an author.

A.8 FIGURES

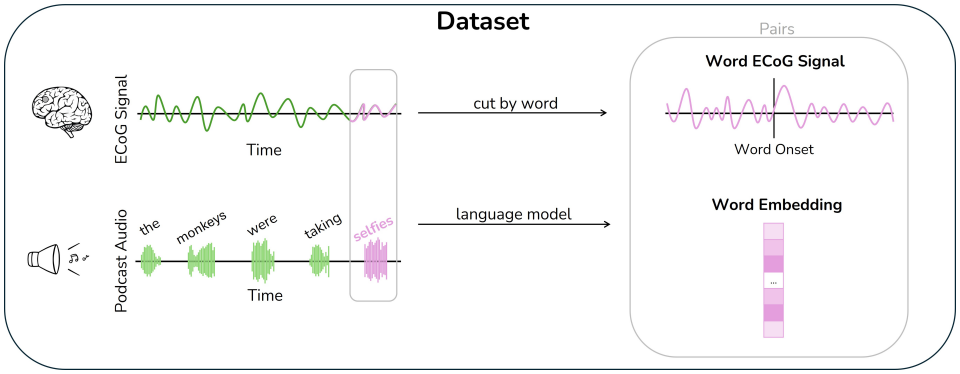

Figure S1: **Dataset construction.** Continuous ECoG recorded during podcast listening was aligned to the audio stream and segmented into word-locked neural responses (relative to word onset). Each word was also embedded with a language model, yielding paired samples of a word's ECoG segment and its contextual embedding for subsequent analyses.

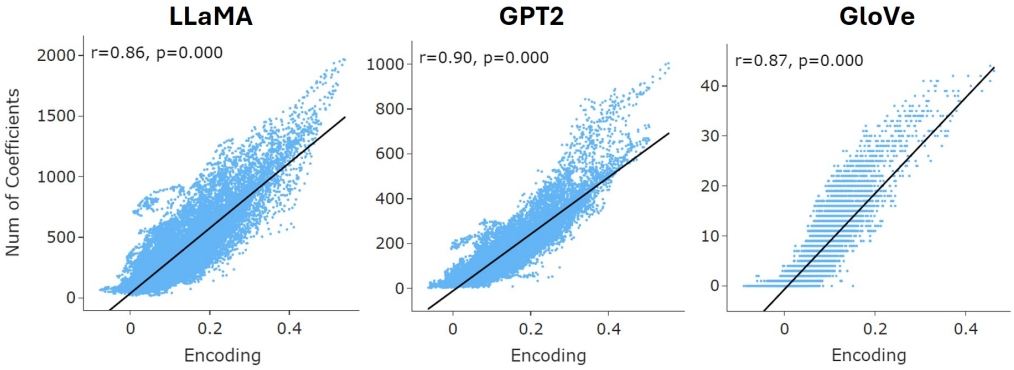

Figure S2: **Coefficient prevalence–encoding relationship generalizes across models.** For all models, the number of significantly correlated features (per electrode–time pair; $q < .05$) is positively associated with encoding accuracy. Each point denotes an electrode-time pair; black lines indicate linear fits, with Pearson's r and corresponding p-values reported in each panel.

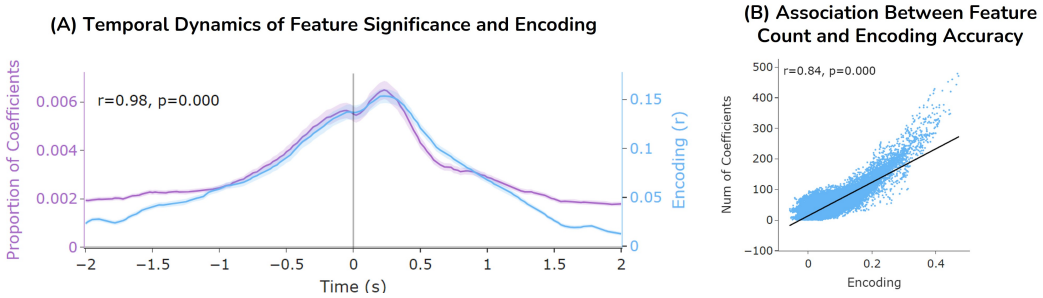

Figure S3: **SAE feature engagement tracks neural encoding.** Similar to Fig. 2, the proportion of SAE features that show a significant correlation with neural activity ($q < .05$) is quantified at each electrode and time point. **(A)** Over time, encoding accuracy (blue) closely tracks the proportion of significantly correlated SAE features (purple), peaking near word onset. **(B)** Across electrode–time pairs, higher encoding accuracy is associated with a larger number of significantly correlated SAE features (line shows linear fit).

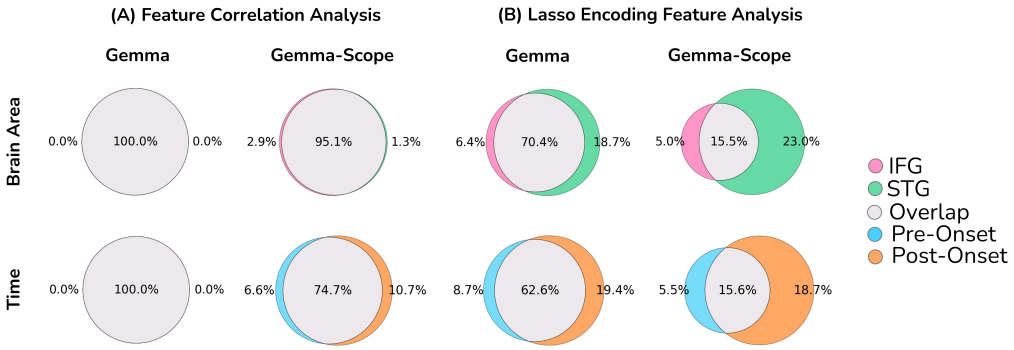

Figure S4: **Feature overlap relative to the full embedding dimensionality.** As in Fig. 3, overlap between conditions is quantified via set intersection/union.Here, counts are normalized for each model by the total number of embedding dimensions that were active during the podcast ($n_{\text{Gemma-Scope}} = 6064$; $n_{\text{Gemma}} = 3584$). Percentages therefore reflect absolute prevalence within the full embedding space, rather than only within the Venn-identified subset.

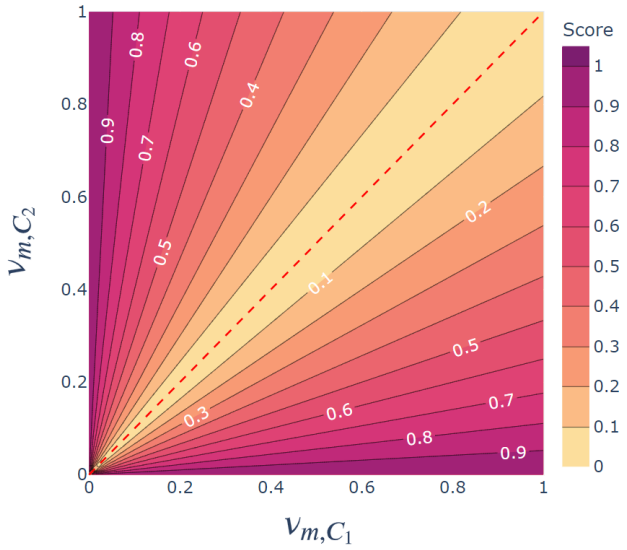

Figure S5: **Visualizing the bias score.** This contour map illustrates the bias score (Canberra distance) used to quantify a feature's preference for one condition over another. The axes represent feature engagement frequencies ($v_{m,C}$) in two competing conditions. The dashed diagonal indicates equal utilization (score 0), while the gradient rises as values diverge toward the axes, culminating in a score of 1 which represents complete exclusivity to a single condition.

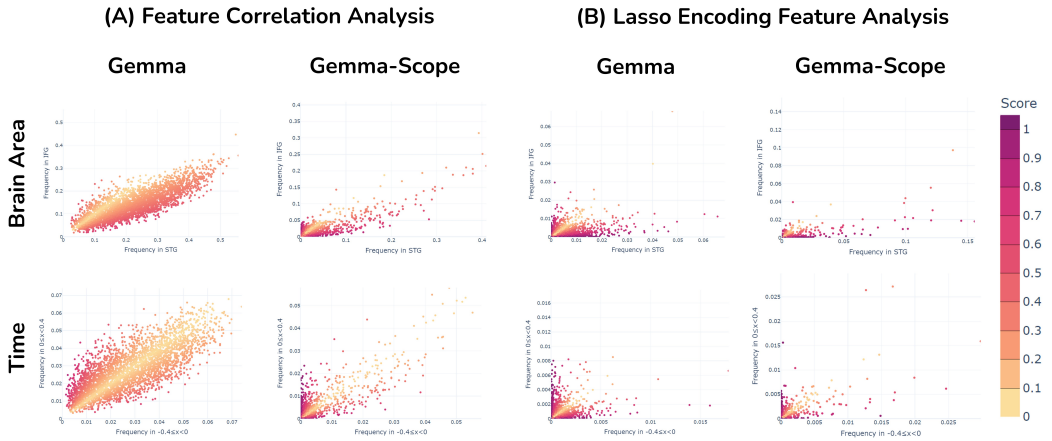

Figure S6: **Feature Engagement Frequencies Across Conditions.** Scatter plots visualize the engagement frequency ($v_{m,C}$) of individual features across opposing conditions: IFG versus STG (top) and pre- versus post-word onset intervals (bottom). The position of each point represents a feature's utilization rate in the two conditions. Features clustering along the diagonal ($v_{m,C_1} = v_{m,C_2}$) indicate equal recruitment, while dispersion toward the axes reveals condition-specific bias. Points are colored according to the bias score (Canberra distance, Fig. S5), where darker hues indicate a greater metric distance from the identity line and thus higher condition exclusivity.

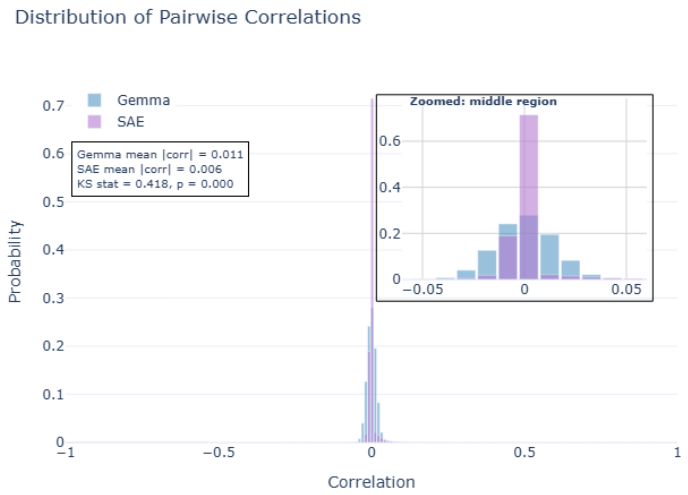

Figure S7: **Pairwise correlation structure for Gemma and SAE.** Distribution of pairwise feature–feature correlations computed across the podcast stimulus for Gemma features (orange) and SAE features (purple). Summary statistics (mean $|r|$) and a two-sample Kolmogorov-Smirnov test are reported in the panel.