ALIGNING BRAINS INTO A SHARED SPACE IMPROVES THEIR ALIGNMENT TO LARGE LANGUAGE MODELS

Arnab Bhattacharjee[1], Zaid Zada[2], Haocheng Wang[2], Bobbi Aubrey[2], Werner Doyle[4], Patricia Dugan[4], Daniel Friedman[4], Orrin Devinsky[4], Adeen Flinker[4,6], Peter J. Ramadge[1], Uri Hasson[2], Ariel Goldstein[3,5*], and Samuel A. Nastase[2*]

[1]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ,
[2]Department of Psychology and the Neuroscience Institute, Princeton University, Princeton, NJ,
[3]Google Research,
[4]New York University Grossman School of Medicine, New York, NY,
[5]Business School, Data Science Department and Cognitive Department, Hebrew University, Jerusalem, Israel,
[6]New York University Tandon School of Engineering, Brooklyn, NY
[*]Equal contribution

ABSTRACT

Recent studies have shown that large language models (LLMs) can accurately predict neural activity measured using electrocorticography (ECoG) during natural language processing. To predict word-by-word neural activity, most prior work has estimated and evaluated encoding models within each electrode and subject—without evaluating how these models generalize across individual brains. In this paper, we analyze neural responses in 8 subjects while they listened to the same 30-minute podcast episode. We use a shared response model (SRM) to estimate a shared information space across subjects. We show that SRM significantly improves LLM-based encoding model performance. We also show that we can use this shared space to denoise the individual brain responses by projecting back into the individualized electrode space, and this process achieves a mean 38% improvement in encoding performance. The strongest improvement was observed for brain areas specialized for language comprehension, specifically in the superior temporal gyrus (STG) and inferior frontal gyrus (IFG). Critically, estimating a shared space allows us to construct encoding models that better generalize across individuals.

1 INTRODUCTION

Recent advances in the field of natural language processing have showcased the exceptional performance of large language models (LLMs) across various natural language tasks such as text generation, translation, summarization, and question-answering (Devlin et al., 2019; Brown et al., 2020; Manning et al., 2020). In parallel, recent studies in human neuroscience have begun positioning LLMs as computational models of human brain activity during context-rich, real-world language processing (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Kumar et al., 2022; Toneva et al., 2022). In these works, researchers use encoding models to estimate a linear mapping between internal representations—i.e. embeddings—extracted from an LLM and measurements of human brain activity, word-by-word during natural language comprehension. This simple approach of linearly "aligning" the LLM's internal feature space to human brain features has yielded remarkably good prediction performance in both functional magnetic resonance imaging (fMRI) and electrocorticography (ECoG). The high spatiotemporal resolution of invasive ECoG recordings, in particular, promises to provide finer-grained insights into shared representations and processes between LLMs and the brain (Goldstein et al., 2022; Cai et al., 2023; Goldstein, Wang, et al., 2023; Mischler et al., 2024; Zada et al., 2023; Goldstein et al., 2024).

When exposed to the same natural language stimulus, such as a spoken story, human neural activity converges on stimulus features ranging from basic acoustic attributes to more complex linguistic and narrative elements (Lerner et al., 2011; Honey et al., 2012; Hasson et al., 2015). However, while a coarse alignment exists across individual brains (Nastase et al., 2019; 2021), the finer cortical topographies for language representation exhibit significant idiosyncrasies among individuals (Fedorenko et al., 2010; Nieto-Castañón & Fedorenko, 2012; Braga et al., 2020; Lipkin et al., 2022). To address this, hyperalignment techniques have been developed in fMRI research to aggregate information across subjects into a unified information space while overcoming the misalignment of functional topographies across subjects (Haxby et al., 2011; Chen et al., 2015; Guntupalli et al., 2016; Haxby et al., 2020; Feilong et al., 2023). Unlike fMRI, where there is putative voxelwise correspondence during acquisition, ECoG presents a more difficult correspondence problem because each subject has a different number of electrodes in different locations (with placement guided by clinical considerations, not research goals). Thus, how to best aggregate electrodes across individuals is a matter of ongoing research (e.g., Owen et al., 2020). For this reason, encoding models are typically constructed separately at each electrode within individual subjects and are not assessed for their generalization to new subjects (e.g., Schrimpf et al., 2021; Goldstein et al., 2022).

In this paper, we measured the neural responses of eight ECoG subjects implanted with invasive intracranial electrodes while they listened to a natural language stimulus. We develop a shared response model (SRM; Chen et al., 2015) to aggregate neural activity and isolate a stimulus-driven shared feature space that is shared across individuals. In parallel, we use LLMs to extract contextual embeddings for each word of the podcast. We then build encoding models to estimate a linear mapping from the contextual embeddings to the shared neural features (Van Uden et al., 2018; Nastase et al., 2020). We show that the SRM yields significantly higher encoding performance than the original individual-specific electrodes. Moreover, we show that we can use this shared space to "denoise" individual subject responses by projecting from the shared space back into the individual electrode space. We find that the SRM-reconstructed data yields the largest improvement in brain areas specialized for language comprehension. Finally, we demonstrate that the SRM allows us to construct encoding models that better generalize across subjects.

2 Materials and Method

2.1 Data Collection and Processing

We recorded the neural activity of eight participants (4 reported female, 20–48 years) using ECoG. Participants were presented with a 30-minute audio podcast "So a Monkey and a Horse Walk Into a Bar, Act One: Monkey in the Middle" from the *This American Life* podcast. We manually transcribed the story and aligned it to the audio by labeling the onset and offset of each word. An independent listener manually evaluated the alignment. There were a total of 5,013 words in the podcast. Using the Hugging Face environment (Wolf et al., 2019), we supplied the transcript to the large language model GPT-2 XL (Radford et al., 2019). For each word, a 1600-dimensional contextual embedding was extracted from the final layer of the model. The meaning of the embedding for each word (excluding the first word) was contextualized by the preceding words in the podcast stimulus. These embeddings were reduced to 50 dimensions using principal component analysis (PCA) for the core encoding analyses, based on Goldstein et al. (2022).

For ECoG data collection, 917 electrodes were placed on the left hemisphere and 233 on the right hemisphere. The ECoG data were sampled at 512 Hz. Line noise harmonics were excluded. We used a band-pass filter to extract activity in the high gamma range of 70-200 Hz.

2.2 Encoding Models

We use contextual word embeddings to predict held-out neural data for individual electrodes or SRM features (see below). First, we extracted gamma power in 200 ms windows at 161 lags ranging from -2000 ms to +2000 ms in 25 ms increments for epochs indexed to each word's onset. For each lag at a given electrode, we then estimated electrode-wise encoding models using ordinary least-squares multiple linear regression: this yields a linear mapping to predict word-by-word neural activity from the associated contextual embeddings (Naselaris et al., 2011; Huth et al., 2016). We employ 10-fold cross-validation to assess the performance of these models in predicting neural responses for

held-out, temporally-contiguous segments of the stimulus. We evaluated out-of-sample prediction by computing the Pearson correlation between the predicted and the actual signal for each held-out test set.

In comparing the encoding performance alignment methods, we performed paired t-tests between the two correlation scores across folds for each lag. To correct for multiple tests across 161 lags, we control the false discovery rate (FDR) at .01 (Benjamini & Hochberg, 1995). Lags with FDR less than .01 are considered to be significant.

2.3 Electrode Selection

To select a subset electrodes involved in language processing, following Goldstein et al. (2022), we first estimated encoding performance using non-contextual GloVe embeddings (Pennington et al., 2014) at 161 lags ranging from -2000 ms to +2000 ms in 25 ms increments for epochs indexed to each word's onset. To evaluate the statistical significance of GloVe-based encoding performance, we performed a randomization test. For each of the electrodes and lags, we randomized the phase of the signal so as to disrupt the temporal alignment while preserving the autocorrelation, then re-estimated the GloVe based encoding models. We repeated this procedure for 5,000 phase randomizations to construct a null distribution from the maximum encoding performance across lags for each electrode. We calculated p-values for each electrode as the percentile of the actual encoding performance relative to 5,000 phase-randomized samples from the null distribution. We controlled the false discovery rate (FDR; Benjamini & Hochberg, 1995) at $q = .01$ across electrodes. If the q-value of the electrode was less than .01, it was selected for further analysis. This yielded 184 electrodes (150 in the left hemisphere, 34 in the right hemisphere) across subjects (see Table S1 for a detailed description of electrode coverage).

2.4 Shared Response Model

Although all the subjects listened to the same story, both the placement of their electrodes and the functional properties of similarly placed electrodes will tend to differ from individual to individual. We use a shared response model (SRM; Chen et al., 2015) to aggregate ECoG data across subjects into a common information space that accounts for different electrode placement and functional topographies across individuals. SRM learns subject-specific transformations that map from each subject's idiosyncratic functional space into a shared space based on a subset of training data, then uses these learned transformations to map a subset of test data into the shared space.

To clarify this, let $\{X_i \in \mathbb{R}^{e \times d}\}_{i=1}^{m}$ be the training data (e electrodes over d time points) for m subjects. We use this training dataset to learn subject-specific bases $W_i \in \mathbb{R}^{e \times k}$ (where k is a hyperparameter that corresponds to the number of components in the new, shared space) and a shared matrix $S \in \mathbb{R}^{k \times d}$, such that $X_i = W_i S + E_i$, where E_i is an error term corresponding to deviation from the subject's original brain activity. The bases W_i represent the individual functional topographies, while S represents latent features that capture components of the response that are shared across subjects. For the solution to be unique, W_i is subject to the constraint of linearly independent columns and W_i is assumed to have orthonormal columns, $W_i^T W_i = I_k$ (Chen et al., 2015). The following optimization problem is solved to estimate W_i and the shared response S:

$$\begin{aligned} &\min_{W_i, S} \sum_i \|X_i - W_i S\|_F^2 \\ &s.t.\ W_i^T W_i = I_k \end{aligned} \tag{1}$$

The S and W parameters of the SRM model are jointly estimated using a constrained EM algorithm. We can utilize the learned subject-specific bases to project data from shared space back into the individual shared response subspace (S_i) to reconstruct a "denoise" version of the data in the original electrode space X_i:

$$\begin{aligned} S_i &= W_i^T X_i \\ \hat{X}_i &= W_i S_i \\ \hat{X}_i &= W_i W_i^T X_i \end{aligned} \tag{2}$$

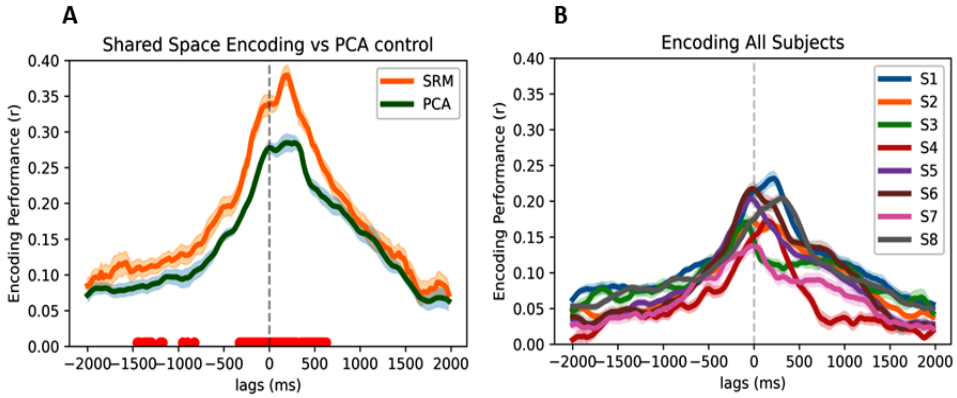

Figure 1: SRM improves model-based encoding performance. (**A**) Encoding model performance based on SRM (orange) and control analysis based on PCA (blue) with matched dimensionality (k = 5). As a control analysis, PCA aggregates neural signals across subjects with the same dimensionality reduction and the same orthogonality constraint, but does not align neural response trajectories across subjects. Encoding performance is averaged across features. The red dots at bottom indicate lags with a significant difference between SRM and PCA-based encoding model performance across folds (FDR controlled at .01). Error bands indicate standard error of the mean across cross-validation folds. (**B**) Encoding model performance based on the original neural activity in each subject (N = 8). Encoding performance is averaged across electrodes within each subject.

3 Results

3.1 Linguistic Encoding in Shared Space

We estimated a shared response model (SRM; Chen et al., 2015) on a training subset of the data (9 out of 10 segments of the story stimulus) across 8 subjects with 5 shared features (hyperparameter k = 5). We selected k = 5 shared features to maximize the number of eligible subjects, given that subject S4 had only eight electrodes survive the electrode selection procedure. This fitting procedure yields a shared space and the corresponding subject-specific weights (W_i). We projected each subject's training data into the shared space and averaged the reduced-dimension data across subjects (S_{train}). Next, we estimated encoding models using the same training data, comprising the average word-by-word time series across subjects in the reduced-dimension shared space. We used linear least-squares regression to estimate a weight matrix to predict the shared neural activity from contextual embeddings (reduced to 50 dimensions using PCA) extracted from GPT-2 XL (Radford et al., 2019). To evaluate encoding model performance, we first use the subject-specific weights W_i to project the test data (corresponding to the left-out test segment of story stimulus) into the shared space estimated from the training data, and average across subjects: $S_{test} = 1/m \sum_{i=1}^{m} W_i^T X_i^{test}$.

We then use the encoding weights estimated from the training set to generate model-based predictions of neural activity in shared space from the contextual embeddings for the left-out training segment of the story stimulus. We evaluate these model-based predictions by computing the Pearson correlation between predicted and actual neural activity for each shared feature. In this way, both the shared response model and the encoding models are estimated and evaluated within the same 10-fold cross-validation procedures (Van Uden et al., 2018; Nastase et al., 2020). We repeated this analysis for lags from 2000 ms before word onset to 2000 ms after with a 25 ms stride, fitting and evaluating separate encoding models at each lag.

When using contextual embeddings to predict shared features, we observed strong encoding performance with peak accuracy (averaged across shared features) of 0.38 roughly 200ms after word articulation (Fig. 1A). SRM dramatically outperforms typical electrode-wise encoding performance using the same embeddings and cross-validation scheme (Fig. 1B). This improvement, however, could be driven by the fact SRM reduces dimensionality by aggregating signals across electrodes. As a control analysis, we instead aggregated electrodes across subjects using principal component analysis (PCA) with dimensionality k = 5 and reassessed encoding performance in the PCA-based shared

space. PCA similarly reduces dimensionality with the same orthogonality constraint as SRM, without aligning individual subjects into a shared feature space. We found that SRM achieves significantly higher encoding performance than PCA ($p < .01$, FDR corrected; Fig. 1A). This control analysis shows that the stronger encoding performance is not simply due to the decreased dimensionality of the shared space.

3.2 Reconstructing Electrode Activity via the Shared Space

We hypothesize that projecting an individual subject's neural activity into the reduced-dimension shared space and then back into electrode space will effectively denoise the individual-subject data and increase encoding model performance. First, we transform the individual subject data into the reduced-dimension shared subspace S_i by multiplying it with the learned, subject-specific weights from SRM training. We then use the transpose of the subject-specific weights to reconstruct their electrode data for both the training and the test sets, as shown in equation 2. Then, we perform an encoding analysis for each subject using the SRM-reconstructed data and compare it with the encoding performances using the subjects' original neural data (Fig. 2). The SRM-reconstructed data significantly improved encoding performance at numerous lags for each subject ($p < .01$ for all subjects, FDR corrected), with an average 38% improvement in peak model performance across subjects.

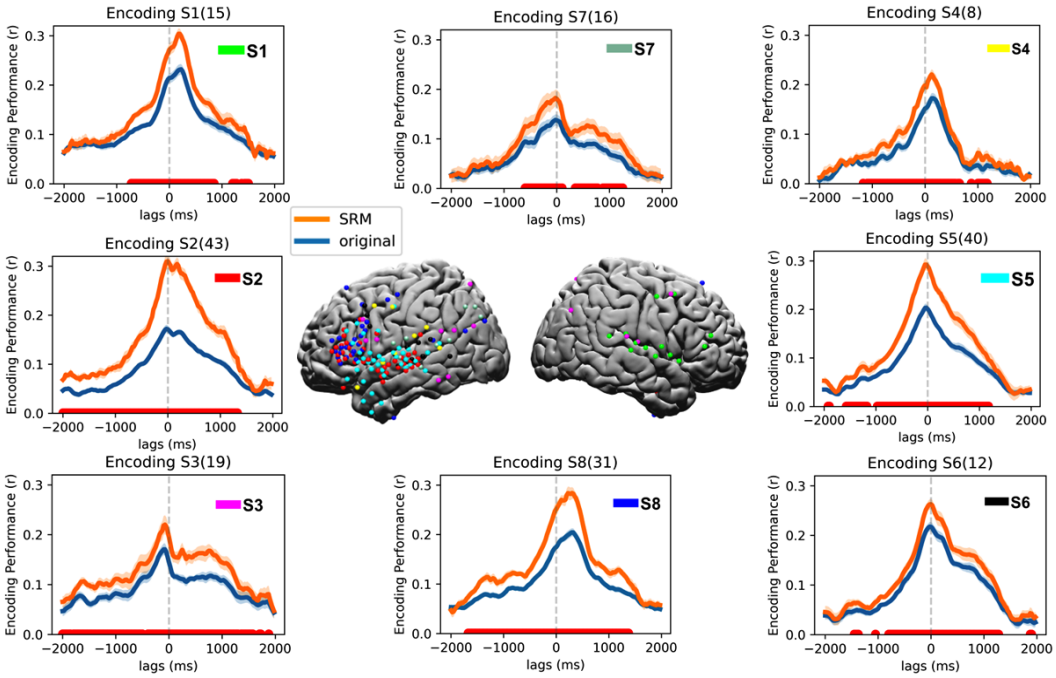

Figure 2: Reconstructing electrode activity via the shared space. At center, electrode placement is shown for all subjects ($N = 8$). Electrode-wise encoding performance is shown for each subject-based electrode activity reconstructed from the shared space (orange) and original electrode activity (blue). Encoding performance is averaged across electrodes within each subject. Error bands indicate standard error of the mean encoding performance across folds. The red markers at bottom indicate lags with a significant difference between encoding performance for SRM-reconstructed and original electrodes across folds (FDR controlled at .01)

3.3 Localizing Improved Encoding Performance with SRM Reconstruction

To map out which brain regions improve most when reconstructing electrode activity from the shared space, we quantified the difference in encoding model performance between SRM-reconstructed data and the original data for each electrode separately (Fig. S1). Qualitatively, the largest improvements were found in the inferior frontal gyrus (IFG) and the superior temporal gyrus (STG). Table

1 reports the number of electrodes at varying ranges of improvement. Out of 184 electrodes, encoding performance nominally improved in 168 electrodes when reconstructed from the shared space, with a maximum improvement of 0.33. We further examined improvements in encoding performance of SRM-reconstructed data for different areas of the language network (Fig. 3). We observed that SRM reconstruction yields significantly better encoding performance compared to the original electrode data in IFG, anterior STG (aSTG), and middle STG (mSTG) ($p < .01$, FDR corrected). While caudal STG (cSTG), angular gyrus (AG), and temporal pole (TP) show nominal improvement in encoding performance, these improvements are not significant after correcting for multiple lags; this may in part be due to the relatively fewer number of electrodes in these areas.

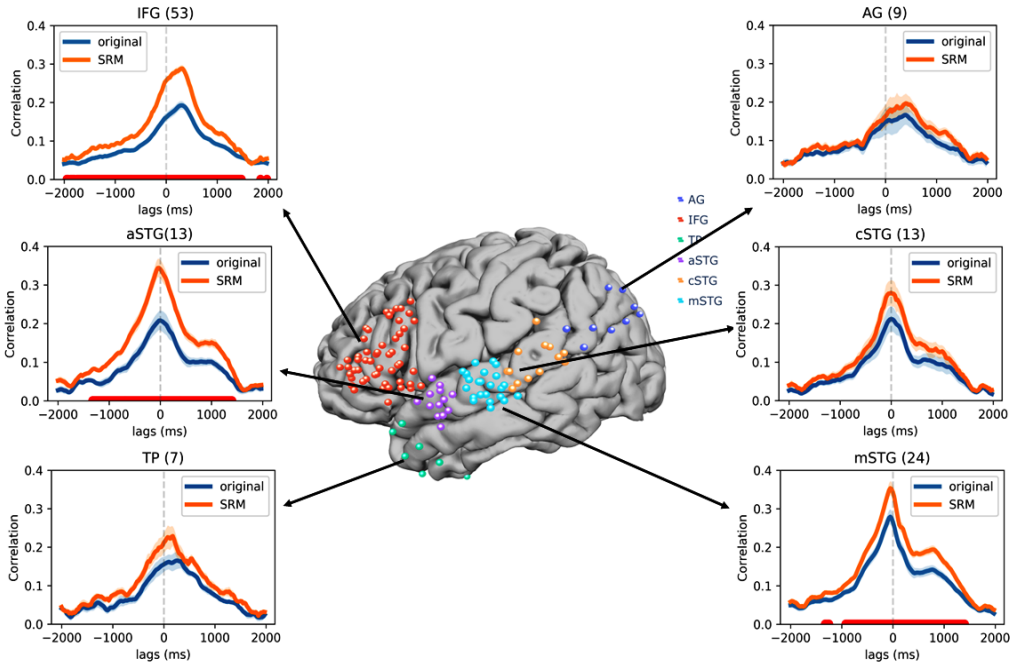

Figure 3: Comparison of encoding performance for SRM-reconstructed data and original electrode data for different regions of the language network. At center, electrode placement is shown for all subjects (N = 8). Electrode-wise encoding performance values for lags spanning -2000 ms to +2000 ms lags are shown for each brain area based on electrode activity reconstructed from the shared space (orange) and original electrode activity (blue). Encoding performance is averaged across electrodes within each brain area. Error bands indicate standard error of the mean encoding performance across folds. The red markers at bottom indicate lags with a significant difference between encoding performance for SRM-reconstructed and original electrodes across folds (FDR controlled at .01 across lags). cSTG: caudal superior temporal gyrus, mSTG: middle superior temporal gyrus, aSTG: anterior superior temporal gyrus, TP: temporal pole, AG: angular gyrus, IFG: inferior frontal gyrus.

Table 1: Improvement in electrode-wise encoding performance with SRM-reconstructed versus original electrode data. Differences were computed between the respective maximum encoding performance values across lags for SRM-reconstructed and original electrode data.

$r_{SRM} - r_{original}$	Number of electrodes
> 0.20	24
0.15–0.20	12
0.10–0.15	43
0.05–0.10	50
0.01–0.05	39

3.4 Generalizing Encoding Models across Subjects via the Shared Space

In the previous analyses, we showed that SRM can improve encoding performance—but, like in prior work (e.g. Goldstein et al., 2022), those encoding models were estimated and evaluated in individual subjects. The shared space captures the shared, stimulus-driven features of brain activity and retains subject-specific mappings to and from the shared space. SRM should therefore allow us to build encoding models that generalize to new subjects who have received the same stimulus (Van Uden et al., 2018; Nastase et al., 2020). To test this hypothesis, we estimated both SRM and encoding models in a subset of training subjects (for a training segment of the story stimulus), then evaluated encoding model performance on a left-out subject (for the left-out test segment of the story). We first estimate a shared space (S) for $N\check{}1$ training subjects based on the training segments of the story. In this case, SRM training data does not include the neural data of the test subject. We can estimate encoding models from the $N\check{}1$ training subjects in this shared space.

Next, we must estimate a transformation to project the test subject's data into the shared space derived from the training subjects. The shared space (S) derived from the training subjects is used as a template, and we calculate a subject-specific weight matrix W_j to rotate the left-out subject j into the pre-existing shared space, using the data X_j^{train} from the training segments of the story. To achieve this, we minimize the mean squared error $min_{W_j, W_j^T W_j = I_k} \|X_j - W_j S\|_F^2$ to find W_j. The shared space S is not affected by aligning a left-out subject in this way. Now, we transform the test subject's neural activity for the test segment of the story into the shared space (estimated from other subjects) using W_j estimated from the training segments of the story: $S_j^{test} = W_j^T X_j^{test}$. Finally, we evaluate the encoding models estimated from other subjects' data and training segments of the story. We use the encoding weights trained on the shared space (S), combined with the embeddings for the test segment to generate predictions for the left-out subject (S_j^{test}). We carry out this process for each lag for all subjects (leave-one-subject-out).

Using SRM, we obtain cross-subject encoding performance (Fig. 4a, orange) comparable to the performance observed when encoding models are estimated and evaluated in individual subjects (Fig. 1b). For a more direct comparison, we implemented a control analysis using PCA: we estimate PCA across $N-1$ training subjects to learn a PCA-based reduced-dimension space (with matching dimensionality and orthogonality constraints to SRM) from the training story segments; we then calculate a W transformation like above to project the test subject onto reduced-dimension PCA reduced space using the left-out subject's training story segments. Finally, we estimate encoding models in the reduced-dimension PCA space from the training subjects and the training story segments. We project the left-out subject's test segment into the shared space to evaluate the model-based predictions. Cross-subject encoding performance is nominally better with SRM than with PCA.

To extend this cross-subject encoding analysis from the reduced-dimension shared space to the original electrode space, we first project $N-1$ subjects to a SRM shared space (S) using the training segments of the story. Next, we calculate the weight matrix W_j to rotate the left-out subject j into the shared space. Now we can use W_j to project data from $N-1$ from the shared space back into the test subject's space: $X^{train} = W_j S$. This allows us to estimate encoding models based strictly on other subjects' data in the test subject's original electrode space: $X^{test} = W_j W_j^T X_j^{test}$. Cross-subject encoding performance in across all the test subject's SRM-reconstructed electrode space nominally outperforms within-subject encoding models in the original electrode space (Fig. 4b). In both of these analyses, we show that an SRM estimated from $N-1$ subjects can be used to find a set of shared features that generalize to a new subject with a different number and placement of electrodes. Given a shared stimulus, SRM can provide a robust enough linkage across disparate, individual-specific electrodes to allow us to build encoding models that generalize to a left-out individual.

3.5 Quantifying Shared Information across Subjects

How well can we reconstruct a novel subject's neural responses to a novel stimulus based on the neural activity of other subjects? To quantify the quality of the shared space without reference to an encoding model, we estimated a shared space based on the training segments of the story in $N-1$ subjects, then reconstructed neural activity for the left-out test segment in a left-out subject. We then correlate the reconstructed neural activity for the test subject j with the subject's actual neural

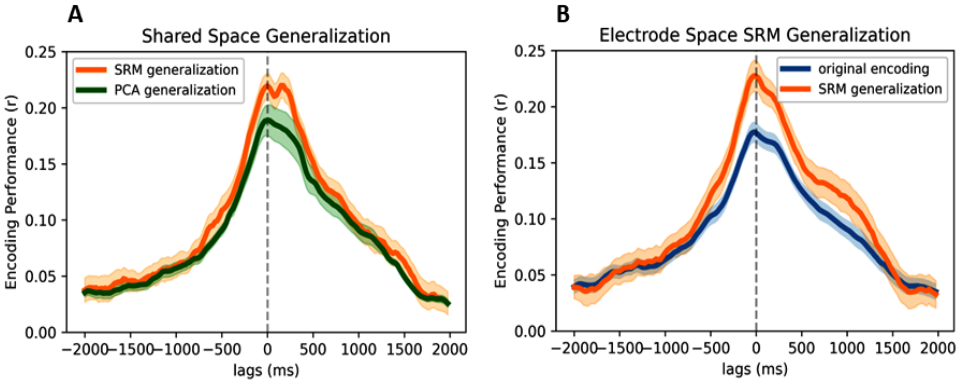

Figure 4: Cross-subject encoding performance via the shared space. In cross-subject encoding, both SRM and encoding models are estimated from $N-1$ subjects and model-based predictions are tested against a left-out subject. (**A**) Cross-subject encoding performance in SRM-based shared space (orange) compared to PCA control. Encoding performance is averaged across features in shared space. (**B**) Cross-subject encoding performance in the test-subject's SRM-reconstructed electrode space compared to within-subject encoding performance in original electrode space. Encoding performance is averaged across subjects and SRM features (a) or electrodes (b). Error bands indicate standard error of the mean encoding performance across subjects.

activity. High correlations indicate that the shared space robustly captures shared information that generalizes across subjects. To elaborate, first, we train an SRM model on the training data for $N-1$ subjects except j using Eq. 1. Then, using the training data X_j for subject j, we find the matrix W_j mapping subject j into the pre-existing shared space by minimizing the mean squared error of $min_{W_j, W_j^T W_j = I_k} \|X_j - W_j S\|_F^2$.

Next, we average the shared responses for the test segment across $N-1$ subjects except j using $S_{test} = 1/m \sum_{i \neq j} W_i^T X_i^{test}$. With this shared response for the test data, we reconstruct the test data for subject j (based strictly on data from $N-1$) by $X_{test}^r = W_j S_{test}$. Finally, we calculate the correlation across words between X_{test}^r and X_{test}^j for each electrode. We repeat this process for each test subject for all the test segments at the word onset. We find that SRM-based reconstruction based on the neural activity of other subjects yields .25 correlation on average (Table 2).

Table 2: SRM reconstruction quality based on other subjects' data transformed via the shared space into the test subject's electrode space. Correlation between SRM-reconstructed and original test data (with standard error of the mean correlation across test sets).

Subject	Correlation between SRM-reconstructed and original data
S1	0.28 ± 0.008
S2	0.21 ± 0.006
S3	0.21 ± 0.012
S4	0.23 ± 0.011
S5	0.29 ± 0.008
S6	0.32 ± 0.009
S7	0.20 ± 0.009
S8	0.23 ± 0.007
Average	0.25 ± 0.009

3.6 Exploring SRM Encoding across Different Models Parameters

Lastly, we explored encoding performance in the shared space across several different sets of model features. We first examined how encoding model performance varies across layers for GPT-2 XL:

we extracted contextual embeddings from all 48 layers of GPT-2 XL and repeated our encoding analysis for both shared features and original electrodes at lags ranging from –2000 ms to +2000 ms relative to word onset. In both cases, we found that intermediate layers yield the highest prediction performance in human brain activity (Fig. S2a,S2b), consistent with prior work (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein, Ham, et al., 2023; Kumar et al., 2022). Next, we evaluated encoding models for two different types of word embeddings: contextual (GPT-2 XL) and non-contextual (GloVe) embeddings (Pennington et al. 2014; note that GloVe encoding was initially used to select electrodes). We found that contextual embeddings yield dramatically higher encoding performance that non-contextual embeddings, both for original electrode data and in the shared space (Fig. S2c,S2d; similarly to prior work, e.g., Schrimpf et al. 2021; Goldstein et al. 2022; Kumar et al. 2022. Finally, we repeat our SRM encoding analysis with several open source GPT models ranging from 125M to 20B parameters: neo-125M, large-774M, neo-1.3B, XL-1.5B, neo-2.7B, neo-20B. SRM yields improved encoding performance for all models and we observe a weak trend consistent with previously reported results (Antonello et al., 2023) where larger models yield better encoding performance (Fig. S2e, S2f).

4 Discussion

Many recent studies have begun to employ encoding models to predict neural responses during natural language processing using contextual embeddings derived from LLMs (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Toneva et al., 2022; Cai et al., 2023; Goldstein, Wang, et al., 2023; Mischler et al., 2024; Zada et al., 2023). Our study demonstrates that aligning the neural activity in each brain into a shared, stimulus-driven feature space significantly enhances encoding performance. This shared space isolates stimulus-driven latent features in neural activity across both subjects and electrodes, while effectively filtering out subject-specific idiosyncrasies (Haxby et al., 2011; Chen et al., 2015; Haxby et al., 2020). Our results illustrate that this shared space exhibits stronger alignment with LLM embeddings than a control model using PCA (with matching dimensionality) to aggregate signals across subjects.

SRM and other hyperalignment methods were developed, initially with fMRI, to estimate a shared information space aligned across subjects (Haxby et al., 2011; Chen et al., 2015; Guntupalli et al., 2016; Feilong et al., 2018; Van Uden et al., 2018; Haxby et al., 2020; Nastase et al., 2020; Feilong et al., 2023). ECoG acquisition presents a more challenging correspondence problem due to varying electrode numbers and placement across subjects (Owen et al., 2020). Electrode placement is often arbitrary, based on clinical considerations, yielding both redundancies and gaps in coverage, which can hamper model generalization. In most ECoG research (e.g., Goldstein et al., 2022; Cai et al., 2023; Mischler et al., 2024), electrodes are simply pooled across subjects to construct a "supersubject". No mapping from one subject to another is attempted, and, critically, whether encoding models actually generalize across subjects has not been investigated. In the current manuscript, we extend SRM to ECoG data and demonstrate several ways in which aggregating electrode signals into a shared information space can improve encoding model performance.

The shared features estimated by SRM are linear combinations of signals across electrodes and subjects (Chen et al., 2015). To more easily interpret these signals, we reconstructed electrode-space activity from the reduced-dimension shared space. This allows us to "denoise" individual data via the shared space. We found that SRM-reconstruction improves encoding performance in most electrodes (a mean 38% improvement), particularly in brain areas associated with language processing, such as the IFG and STG. These areas both (a) contain more electrodes than other areas and (b) may be most closely entrained to linguistic features of the shared stimulus (e.g. Goldstein et al., 2022; 2024).

The vast majority of prior work fitting electrode-wise linguistic encoding models does not evaluate whether models generalize across individual subjects (Goldstein et al. 2022; Goldstein, Wang, et al. 2023; Mischler et al. 2024; cf. Zada et al. 2023). Our findings show that, by estimating both SRM and encoding models on a subset of training subjects and stimuli, the shared space can be used to build encoding models that robustly generalize to new subjects and stimuli. We show that cross-subject encoding performance via the shared space matches or even exceeds within-subject encoding performance. This generalization likely hinges on using a rich, naturalistic stimulus (like a spoken story) to obtain a diverse sampling of brain states, which ultimately yields a more robust, generalizable shared space (Haxby et al., 2011; 2020). This kind of generalization—allowing us to

precisely predict neural activity in previously unseen subjects—can provide a way to circumvent the scarcity of individual-subject data, which is particularly egregious with ECoG recordings in epilepsy patients. Given a shared, naturalistic stimulus, SRM allows us to leverage previously collected data from a larger group of subjects in a single individual's idiosyncratic electrode space—which may accelerate research on individualized brain decoding and brain-computer interfaces (e.g. Metzger et al., 2023; Willett et al., 2023).

We show that SRM improves model-based encoding performance and provides a basis for robustly generalizing encoding models across individual subjects. SRM is a data-driven, unsupervised algorithm that isolates stimulus-driven features of neural activity and aligns them across individuals. What are these stimulus-driven features? For a naturalistic, spoken language stimulus, we hypothesize that these features likely capture the structure of real-world language supporting comprehension, production, and ultimately communication. Using contextual embeddings derived from an LLM, we confirm this hypothesis by showing that SRM improves model-based encoding performance. That is, we show that by aligning neural signals across subjects, we more closely converge on the shared set of linguistic features encoded by individual brains and LLMs.

REFERENCES

Antonello, R., Vaidya, A., & Huth, A. (2023). Scaling laws for language encoding models in fMRI. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 21895–21907). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/4533e4a352440a32558c1c227602c323-Paper-Conference.pdf

Benjamini, Y., & Hochberg, Y. (1995). 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. doi: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Braga, R. M., DiNicola, L. M., Becker, H. C., & Buckner, R. L. (2020). Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *Journal of Neurophysiology*, *124*(5), 1415–1448. doi: https://doi.org/10.1152/jn.00753.2019

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cai, J., Hadjinicolaou, A. E., Paulk, A. C., Williams, Z. M., & Cash, S. S. (2023). Natural language processing models reveal neural dynamics of human conversation. *bioRxiv*. doi: https://doi.org/10.1101/2023.03.10.531095

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. doi: https://doi.org/10.1038/s42003-022-03036-1

Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fmri shared response model. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2015/file/b3967a0e938dc2a6340e258630febd5a-Paper.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N19-1423

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. doi: https://doi.org/10.1152/jn.00032.2010

Feilong, M., Nastase, S. A., Guntupalli, J. S., & Haxby, J. V. (2018). Reliable individual differences in fine-grained cortical functional architecture. *NeuroImage*, *183*, 375–386. doi: https://doi.org/10.1016/j.neuroimage.2018.08.029

Feilong, M., Nastase, S. A., Jiahui, G., Halchenko, Y. O., Gobbini, M. I., & Haxby, J. V. (2023). The Individualized Neural Tuning Model: precise and generalizable cartography of functional architecture in individual brains. *Imaging Neuroscience*, *1*, 1–34. doi: https://doi.org/10.1162/imag_a_00032

Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., … Hasson, U. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, *15*(1), 2768. doi: https://doi.org/10.1038/s41467-024-46631-y

Goldstein, A., Ham, E., Nastase, S. A., Zada, Z., Grinstein-Dabus, A., Aubrey, B., … Hasson, U. (2023). Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *bioRxiv*. doi: https://doi.org/10.1101/2022.07.11.499562

Goldstein, A., Wang, H., Niekerken, L., Zada, Z., Aubrey, B., Sheffer, T., … Hasson, U. (2023). Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations. *bioRxiv*. doi: https://doi.org/10.1101/2023.06.26.546557

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., … others (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. doi: https://doi.org/10.1038/s41593-022-01026-4

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A model of representational spaces in human cortex. *Cerebral Cortex*, *26*(6), 2919–2934. doi: https://doi.org/10.1093/cercor/bhw068

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., … Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416. doi: https://doi.org/10.1016/j.neuron.2011.08.026

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*, e56601. doi: https://doi.org/10.7554/eLife.56601

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., … Hasson, U. (2012). Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, *76*(2), 423–434. doi: https://doi.org/10.1016/j.neuron.2012.08.011

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. doi: https://doi.org/10.1038/nature17637

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., … Nastase, S. A. (2022). Shared functional specialization in transformer-based language models and the human brain. *bioRxiv*. doi: https://doi.org/10.1101/2022.06.08.495348

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. doi: https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., ... Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fmri data from >800 individuals. *Scientific Data*, *9*(1), 529. doi: https://doi.org/10.1038/s41597-022-01645-3

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

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., ... Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 1–10. doi: https://doi.org/10.1038/s41586-023-06443-4

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. *arXiv*. doi: https://doi.org/10.48550/arXiv.2401.17671

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *NeuroImage*, *56*(2), 400–410. doi: https://doi.org/10.1016/j.neuroimage.2010.07.073

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*(6), 667–685. doi: https://doi.org/10.1093/scan/nsz037

Nastase, S. A., Liu, Y.-F., Hillman, H., Norman, K. A., & Hasson, U. (2020). Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage*, *217*, 116865. doi: https://doi.org/10.1016/j.neuroimage.2020.116865

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., ... Hasson, U. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1), 250. doi: https://doi.org/10.1038/s41597-021-01033-3

Nieto-Castañón, A., & Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*, *63*(3), 1646–1669. doi: https://doi.org/10.1016/j.neuroimage.2012.06.065

Owen, L. L., Muntianu, T. A., Heusser, A. C., Daly, P. M., Scangos, K. W., & Manning, J. R. (2020). A Gaussian process model of human electrocorticographic data. *Cerebral Cortex*, *30*(10), 5333–5345. doi: https://doi.org/10.1093/cercor/bhaa115

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. doi: https://doi.org/10.3115/v1/D14-1162

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9. Retrieved from `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. doi: https://doi.org/10.1073/pnas.2105646118

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, *2*(11), 745–757. doi: https://doi.org/10.1038/s43588-022-00354-6

Van Uden, C. E., Nastase, S. A., Connolly, A. C., Feilong, M., Hansen, I., Gobbini, M. I., & Haxby, J. V. (2018). Modeling semantic encoding in a common neural representational space. *Frontiers in Neuroscience*, *12*, 378029. doi: https://doi.org/10.3389/fnins.2018.00437

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., … Henderson, J. M. (2023). A high-performance speech neuroprosthesis. *Nature*, 1–6. doi: https://doi.org/10.1038/s41586-023-06377-x

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2019). HuggingFace's transformers: State-of-the-art natural language processing. *arXiv*. doi: https://doi.org/10.48550/arXiv.1910.03771

Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., … Hasson, U. (2023). A shared linguistic space for transmitting our thoughts from brain to brain in natural conversations. *bioRxiv*. doi: https://doi.org/10.1101/2023.06.27.546708

SUPPLEMENTARY INFORMATION

Table S1: Electrode localization to different brain areas for each subject. STG: superior temporal gyrus, aMTG: anterior middle temporal gyrus, pMTG: posterior middle temporal gyrus, TP: temporal pole, AG: angular gyrus, IFG: inferior frontal gyrus, MFG: middle frontal gyrus, PostCG: postcentral gyrus.

Subjects	Right Hemisphere	Left Hemisphere												
		IFG	STG	AG	TP	Precentral	Parietal	PostCG	aMTG	pMTG	Premotor	MFG	Other	Total
S1	15													15
S2		18	16			1	1		1				6	43
S3	10	2		3		1				2			1	19
S4			3		1	1	2				1			8
S5	2	11	18		5	2			1				1	40
S6		3	6				1			1		1		12
S7	5		3	4			1	1		1			1	16
S8	2	19	1	2	1	1					1	3	1	31
														184

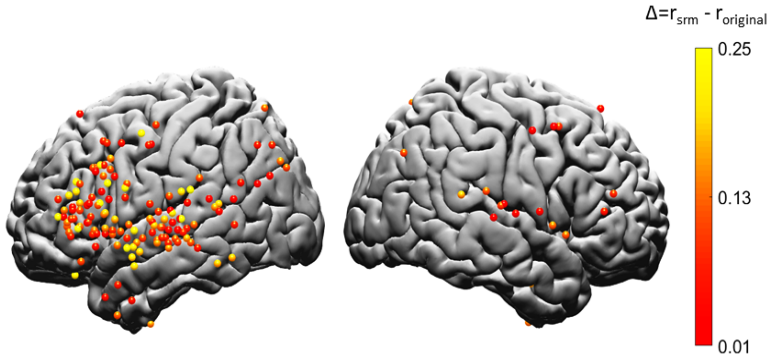

Figure S1: Electrode-wise differences in encoding model performance between SRM-reconstructed data and the original electrode data. Differences were computed between the respective maximum encoding performance values across lags for SRM-reconstructed and original electrode data.

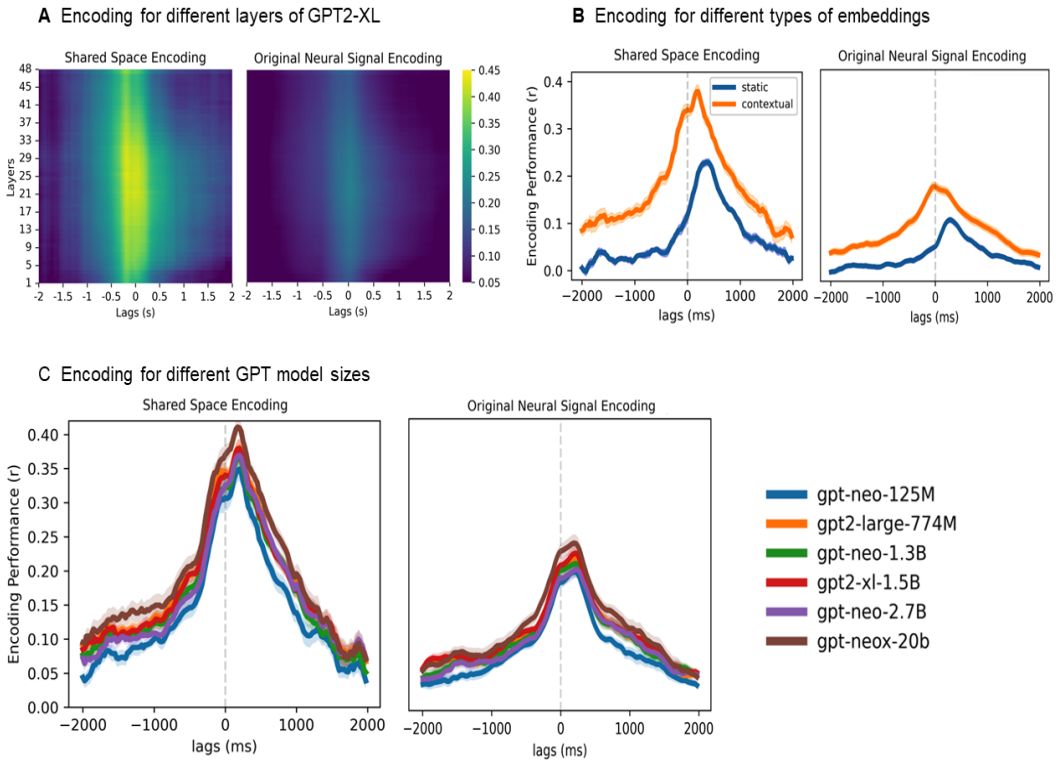

Figure S2: Exploring SRM encoding across different model features. (**A**) Encoding performance across all the layers of GPT-2 XL for both shared space (left) and original electrode data (right). (**B**) Comparison of encoding performance for contextual embeddings from GPT-2 XL (orange) and non-contextual embeddings from GloVe (blue) in both the shared space (left) and original electrode data (right). (**D**) Encoding performance across different sizes of GPT models for both shared space and original electrode data. In all cases, the error bands indicate the standard error of the mean across folds.

Code availability: https://github.com/pritamarnab/SRM-Encoding