nature computational science

Article

https://doi.org/10.1038/s43588-025-00900-y

Aligning brains into a shared space improves their alignment with large language models

Received: 28 May 2024

Accepted: 25 September 2025

Published online: 18 November 2025

Check for updates

Arnab Bhattacharjee[1], Zaid Zada[2], Haocheng Wang[2], Bobbi Aubrey[2], Werner Doyle[3], Patricia Dugan[3], Daniel Friedman[3], Orrin Devinsky[3], Adeen Flinker[3,4], Peter J. Ramadge[1], Uri Hasson[2], Ariel Goldstein[5,6,8] & Samuel A. Nastase[2,7,8]

Recent research demonstrates that large language models can predict neural activity recorded via electrocorticography during natural language processing. To predict word-by-word neural activity, most prior work evaluates encoding models within individual electrodes and participants, limiting generalizability. Here we analyze electrocorticography data from eight participants listening to the same 30-min podcast. Using a shared response model, we estimate a common information space across participants. This shared space substantially enhances large language model-based encoding performance and enables denoising of individual brain responses by projecting back into participant-specific electrode spaces—yielding a 37% average improvement in encoding accuracy (from $r = 0.188$ to $r = 0.257$). The greatest gains occur in brain areas specialized for language comprehension, particularly the superior temporal gyrus and inferior frontal gyrus. Our findings highlight that estimating a shared space allows us to construct encoding models that better generalize across individuals.

Recent advances in the field of natural language processing have showcased the exceptional performance of large language models (LLMs) across various natural language tasks[1–3]. 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[4–8]. In these works, researchers use encoding models to estimate a linear mapping between internal representations—namely, 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[6,9–14].

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[15–17]. However, while a coarse alignment exists across individual brains[18,19], the finer cortical topographies for language representation exhibit notable idiosyncrasies among individuals[20–23]. To address this, hyperalignment techniques have been developed in fMRI research to aggregate information across participants into a unified information space while overcoming the misalignment of functional topographies across participants[24–28]. ECoG presents a more difficult

[1]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA. [2]Department of Psychology and Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [3]New York University Grossman School of Medicine, New York, NY, USA. [4]New York University Tandon School of Engineering, Brooklyn, NY, USA. [5]Google Research, Mountain View, CA, USA. [6]Business School, Data Science Department and Cognitive Department, Hebrew University, Jerusalem, Israel. [7]Department of Psychology and Center for Computational Language Sciences, University of Southern California, Los Angeles, CA, USA. [8]These authors contributed equally: Ariel Goldstein, Samuel A. Nastase. e-mail: arnab@princeton.edu

https://doi.org/10.1038/s43588-025-00900-y

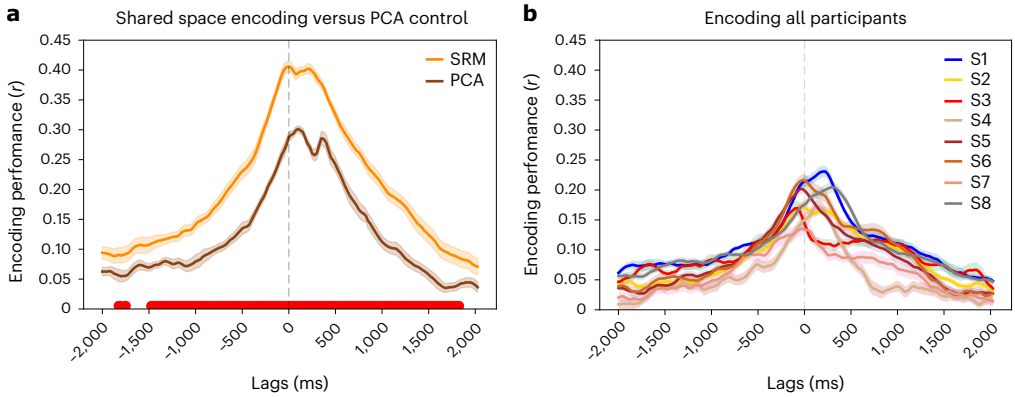

Fig. 1 | Improving model-based encoding performance with SRM. a, Encoding model performance based on SRM (orange) and control analysis based on PCA (blue) with matched dimensionality ($k = 5$). As a control analysis, PCA aggregates neural signals across participants with the same dimensionality reduction and the same orthogonality constraint, but does not align neural response trajectories across participants. Encoding performance is averaged across features. Red markers at the bottom indicate lags where the encoding performance of SRM and PCA differs significantly across test folds ($P < 0.01$, two-sided t-test, FDR corrected). **b**, Encoding model performance based on the original neural activity in each participant ($N = 8$). Encoding performance is averaged across electrodes within each participant. Error bands indicate the standard error of the mean across cross-validation folds for all the plots.

correspondence problem than fMRI because each participant 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[29]. For this reason, encoding models are typically constructed separately at each electrode within individual participants and are not assessed for their generalization to new participants[4,6].

In this Article, we measured the neural responses of eight ECoG participants implanted with invasive intracranial electrodes while they listened to a natural language stimulus. We develop a shared response model[25] (SRM) to aggregate neural activity and isolate a stimulus-driven 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[30,31]. We show that the SRM yields substantially higher encoding performance than the original individual-specific electrodes. Moreover, we show that we can use this shared space to 'denoise' individual participant responses by projecting from the shared space back into the individual electrode space. We find that the SRM-reconstructed data yield 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 participants.

Results

We recorded neural activity in eight participants using ECoG while they listened to a 30-min podcast comprising roughly 5,000 words. In keeping with prior work[6], we extracted the high-gamma power across 184 electrodes selected for sensitivity to language (see 'Electrode selection' section in the Methods; Supplementary Table 1). We averaged neural activity in 200-ms windows at 25-ms lag increments ranging from −2,000 ms to +2,000 ms relative to word onset. Using a transcript of the stimulus, we extracted token-level contextual embeddings from the widely used LLM generative pretrained transformer (GPT-2) XL (ref. 32) and averaged these into a single embedding per word. We then reduced these high-dimensional word embeddings to 50 dimensions using principal component analysis (PCA). Finally, we mapped these embeddings onto neural activity separately at each lag using linear regression with tenfold cross-validation[33,34]. Encoding models were estimated to predict fluctuations in the neural signal across words at a given lag, and separate models were fit at each lag. Encoding models were trained on nine temporally contiguous segments of the stimulus, then tested on the tenth left-out segment. We evaluated encoding performance by computing the correlation between the model-predicted and actual neural activity across words at a given electrode or shared feature. In the following analyses, we evaluate the extent to which aggregating electrode signals into a shared space across participants improves encoding performance.

Linguistic encoding in shared space

We estimated a SRM[25] using tenfold cross-validation (training on a subset of the nine out of ten segments of the story) with $k = 5$ shared features. Given that k cannot be larger than the total number of electrodes in any given participant, we chose $k = 5$ as a round number lower than the fewest number of electrodes in any single participant (eight electrodes in participant S4). This fitting procedure yields a shared space and the corresponding participant-specific weights (W_i). We projected each participant's training data into the shared space and averaged the reduced-dimension data across participant (S_{train}). This S_{train} is the averaged word-by-word time series across participants projected into the reduced-dimension shared space: $S_{\text{train}} = 1/m \sum_{i=1}^{m} W_i^T X_i^{\text{train}}$. Next, we estimated encoding models using S_{train}. We used least-squares regression to estimate a weight matrix to predict the shared neural activity from GPT-2 XL embeddings (reduced to 50 dimensions using PCA). To evaluate encoding model performance, we first use the participant-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 then average across participants: $S_{\text{test}} = 1/m \sum_{i=1}^{m} W_i^T X_i^{\text{test}}$.

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 test segment of the story stimulus. We evaluate these model-based predictions by computing the correlation between predicted and actual neural activity for each shared feature. In this way, both the SRM and the encoding models are estimated and evaluated within the same tenfold cross-validation scheme[30,31]. We repeated this analysis in 200-ms windows for lags ranging from 2,000 ms before word onset to 2,000 ms after with a 25-ms stride. Both SRM and encoding models were estimated and evaluated separately at each lag relative to word onset.

When using contextual embeddings to predict shared features, we observed strong encoding performance with peak accuracy (averaged across shared features) of $r = 0.405$ roughly 200 ms after word articulation (Fig. 1a). SRM dramatically outperformed typical electrode-wise encoding performance using the same embeddings and cross-validation scheme (Fig. 1b). This improvement, however,

https://doi.org/10.1038/s43588-025-00900-y

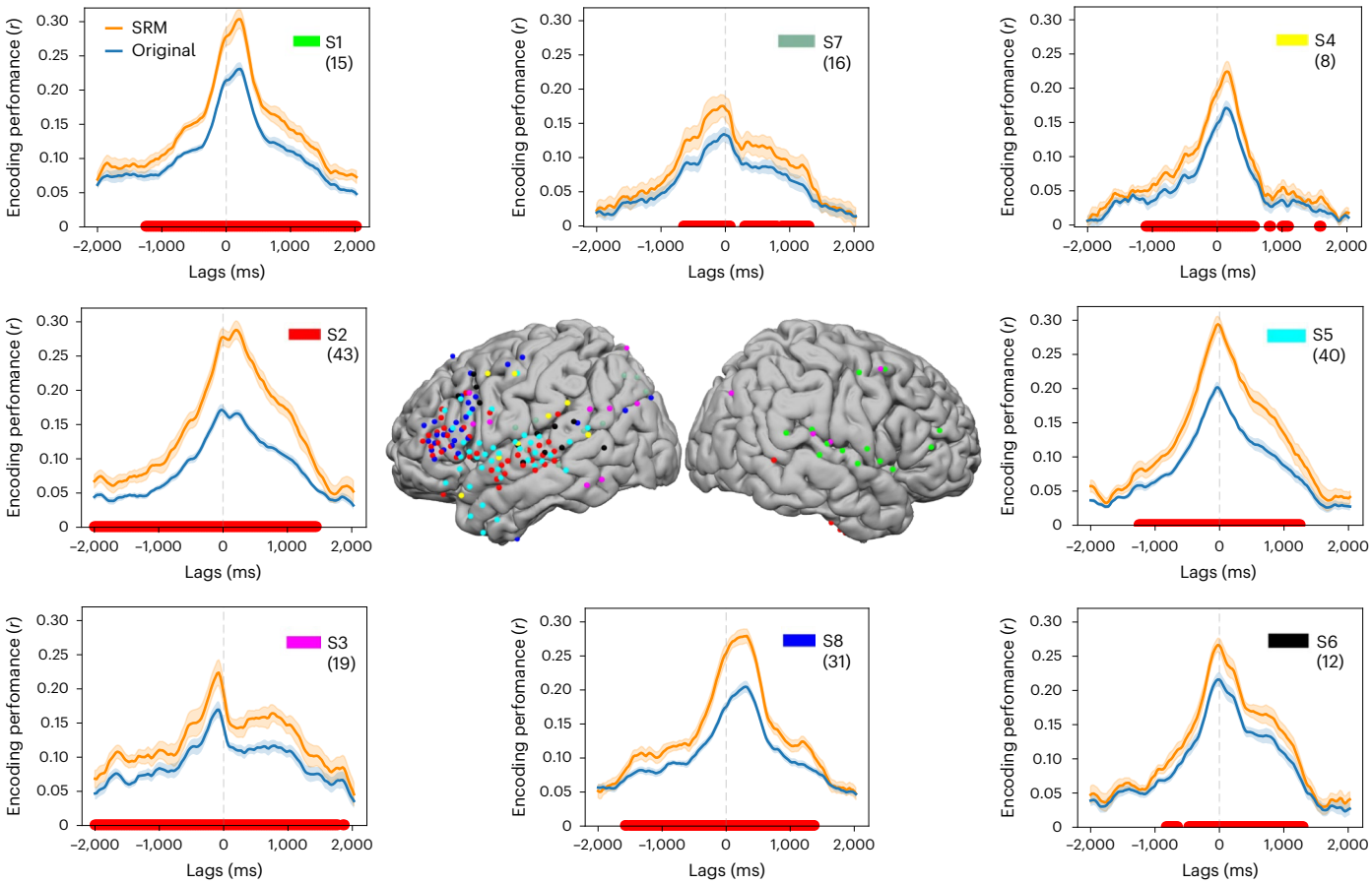

Fig. 2 | Reconstructing electrode activity via the shared space. At center, electrode placement is shown for all participants ($N = 8$). Electrode-wise encoding performance is shown for each participant based on reconstructed electrode activity from the shared space (orange) and original electrode activity (blue). Encoding performance is averaged across electrodes within each participant. Error bands indicate the standard error of the mean encoding performance across folds. Red markers at the bottom indicate lags with a significant difference in encoding performance between SRM-reconstructed and original electrodes across test folds ($P < 0.01$, two-sided t-test, FDR corrected). In each subplot, the number in brackets indicates the number of electrodes for that participant.

could be driven by the fact SRM reduces dimensionality by aggregating signals across electrodes.

As a control analysis, we instead aggregated electrodes across participants using PCA with dimensionality $k = 5$. When performing this PCA-based control analysis, we concatenated all the participants along the electrode axis (because all participants have the same number of words/samples due to the shared podcast stimulus), resulting in a $w \times E$ matrix, where w is the number of words and E is the total number of electrodes across participants. We then estimated PCA from the training set of this matrix, effectively reducing the total number electrodes across participants to $k = 5$. We projected the test data onto the principal components estimated from the training data and reassessed encoding performance in this PCA-based reduced dimensional space using the same procedure as used with SRM. PCA similarly reduces dimensionality with the same orthogonality constraint as SRM, without aligning individual participants into a shared feature space. We found that SRM achieves significantly higher encoding performance than PCA ($P < 0.01$, false discovery rate (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.

Taken together, these findings indicate that using SRM to isolate stimulus-driven neural activity that is shared across participants yields improved alignment to the LLM embeddings. In a follow-up analysis focusing on a subset of three participants with the largest number of electrodes, we found that encoding performance is stable at higher shared dimensionality k (Supplementary Fig. 1); that said, increasing the shared dimensionality does yield a larger number of features with moderate encoding performance (Supplementary Table 2). We also visualized the encoding performance for each of the $k = 5$ shared features separately (Supplementary Fig. 2), revealing somewhat different temporal profiles of encoding performance for each feature. Finally, in a control analysis, we shuffled the embeddings to disrupt the pairing between neural activity and the corresponding stimulus words. With permuted embeddings, encoding performance drops to approximately zero (Supplementary Fig. 3), indicating that encoding performance is driven by word-specific features of the embeddings.

Reconstructing electrode activity via the shared space

We hypothesized that projecting an individual participant's neural activity into the reduced-dimension shared space and then back into electrode space will effectively denoise the individual participant data and increase encoding model performance. First, we transformed the individual participant data into the reduced-dimension shared subspace S_i by multiplying it with the learned, participant-specific weights from SRM training. We then used the transpose of the participant-specific weights to reconstruct a given participant's electrode data for both the training and the test sets, as shown in equation (2). Next, we performed an encoding analysis for each participant using the SRM-reconstructed data and compared it with the encoding performances using the participants' original neural data (Fig. 2). The SRM-reconstructed data significantly improved encoding performance

https://doi.org/10.1038/s43588-025-00900-y

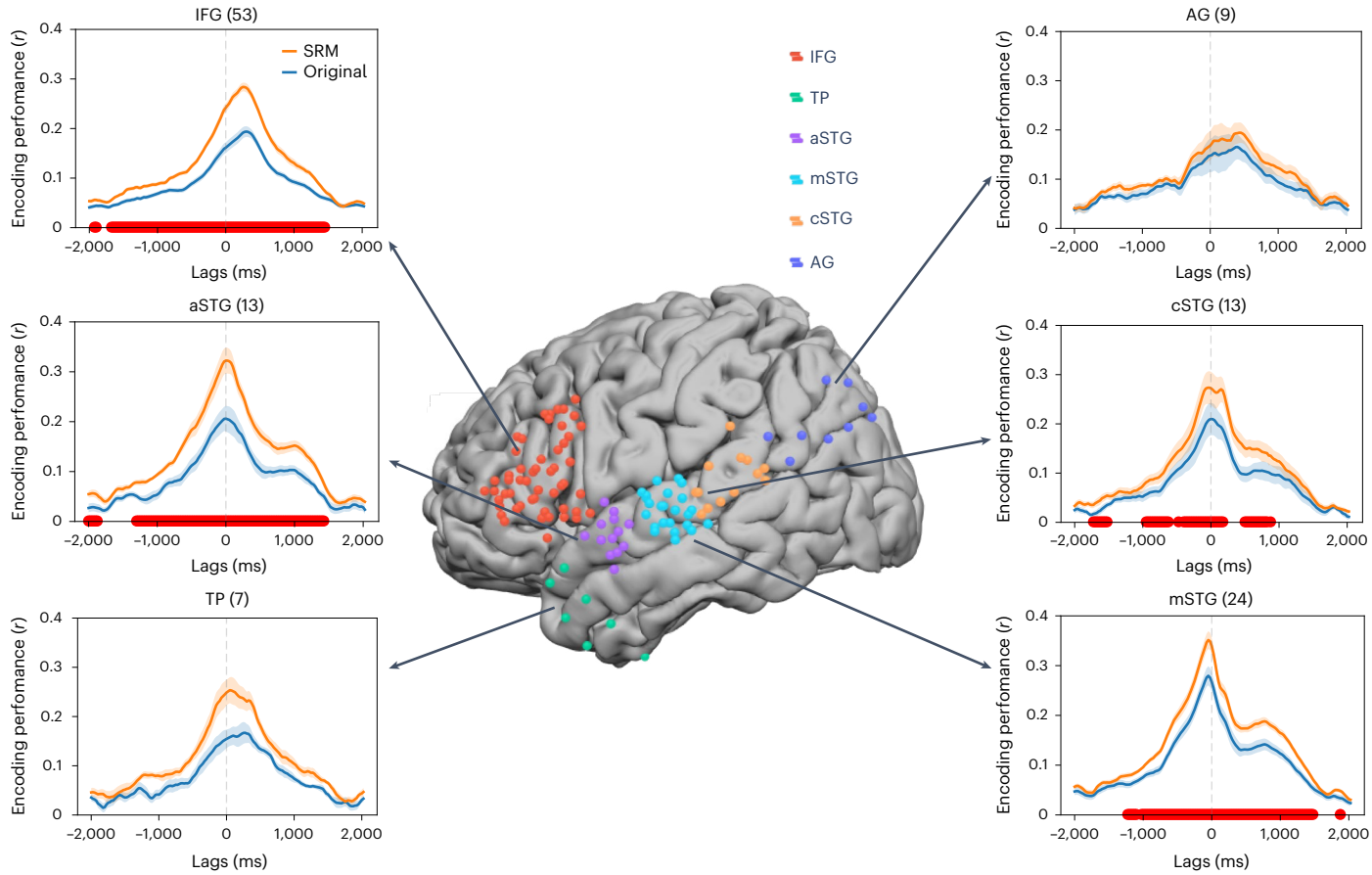

Fig. 3 | Comparison of encoding performance for SRM-reconstructed data and original electrode data for different regions of the language network. At center, electrode placement is shown for all participants ($N = 8$). Encoding performance values for lags spanning −2,000 ms to +2,000 ms lags are shown 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 the standard error of the mean encoding performance across folds. Red markers at the bottom indicate lags with a significant difference in encoding performance between SRM-reconstructed and original electrodes across folds ($P < 0.01$, two-sided t-test, FDR corrected). For each subplot title, the number in brackets indicates the electrode number of the respective brain area.

at numerous lags for each participant ($P < 0.01$ for all participants, FDR corrected), with an average 37% improvement (from $r = 0.188$ to $r = 0.257$) in peak model performance across participants (see Supplementary Table 3 for participant-level improvements).

Localizing improved encoding performance with SRM

To map out which brain regions improve most while reconstructing electrode activity from the shared space, we quantified the difference in peak encoding model performance between SRM-reconstructed data and the original data for each electrode separately (Supplementary Fig. 4). Qualitatively, the largest improvements were found in the inferior frontal gyrus (IFG) and the superior temporal gyrus (STG) (see Supplementary Table 4 for the number of electrodes at varying ranges of improvement). To compare encoding performance with and without SRM reconstruction at the electrode level, we performed paired t-tests at each electrode across the ten cross-validation test sets. Out of 184 electrodes, encoding performance significantly improved in 131 electrodes (all $P < 0.05$, FDR corrected) when reconstructed from the shared space. We further examined improvements in encoding performance of SRM-reconstructed data for different areas of the language network[35] (Fig. 3). We observed that SRM reconstruction yields significantly better encoding performance compared with the original electrode data in IFG, anterior STG (aSTG), and middle STG (mSTG) ($P < 0.01$, FDR corrected). While caudal STG (cSTG), angular gyrus (AG) and temporal pole (TP) show nominal improvements in encoding performance, this may be partly due to the relatively small number of electrodes in these regions.

Generalizing encoding model across participants via shared space

In the previous analyses, we showed that SRM can improve encoding performance—but, like prior work[6], we estimated and evaluated those encoding models in individual participants. SRM, however, should therefore allow us to build encoding models that generalize to new participants who have received the same stimulus[30,31]. To test this hypothesis, we estimated both SRM and encoding models in a subset of training participants (for a training segment of the story stimulus), then evaluated encoding model performance on a left-out participant (for the left-out test segment of the story). We first estimated a shared space (S) for $N-1$ training participants based on the training segments of the story. In this case, SRM training data did not include the neural data of the test participant. Then, we estimated encoding models from the $N-1$ training participants in this shared space.

Next, we estimated a transformation to project the test participant's data into the shared space derived from the training participants. The shared space (S) derived from the training participants is used as a template, and we calculate a participant-specific weight matrix W_j to rotate the left-out participant 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

https://doi.org/10.1038/s43588-025-00900-y

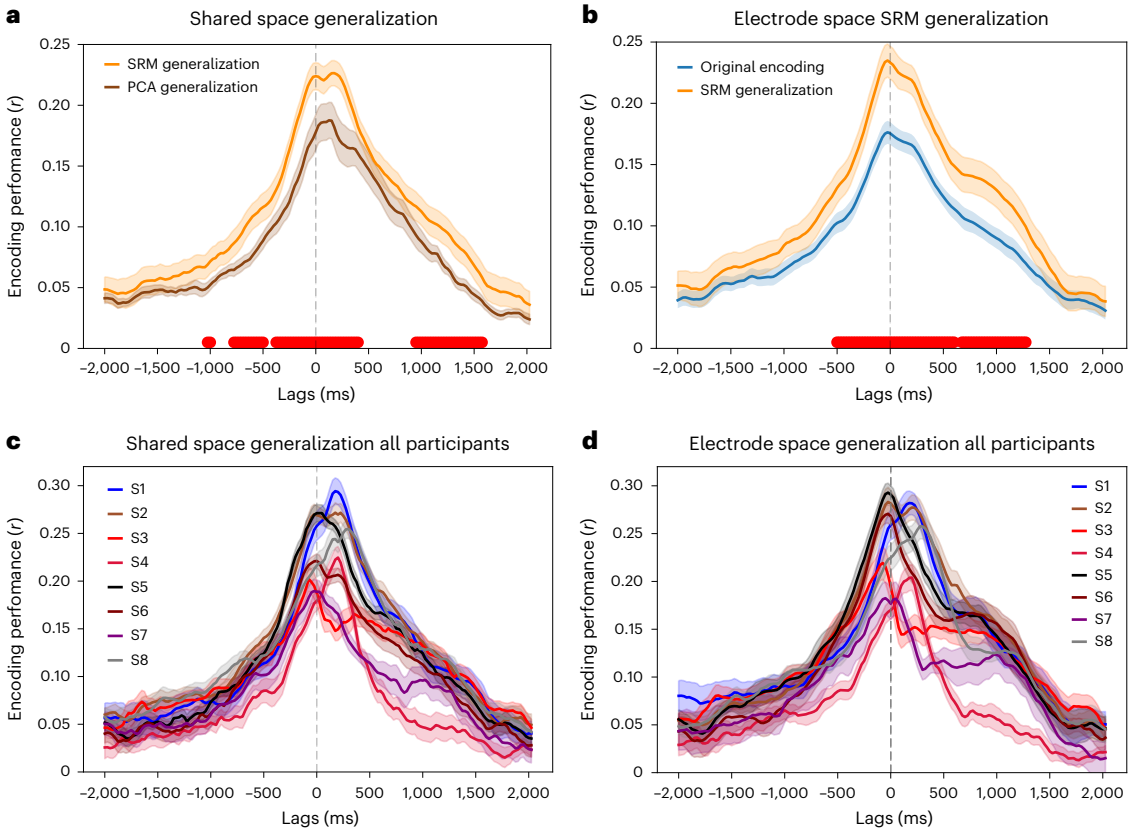

Fig. 4 | Cross-participant encoding performance via the shared space. In cross-participant encoding, both SRM and encoding models are estimated from $N-1$ participants and model-based predictions are tested against a left-out participant. **a**, Cross-participant encoding performance in SRM-based shared space (orange) compared with PCA control. Encoding performance is averaged across test participants and features in shared space. **b**, Cross-participant encoding performance in the test participant's SRM-reconstructed electrode space compared with within-participant encoding performance in original electrode space. Encoding performance is averaged across test participants and electrodes. Error bands indicate the standard error of the mean encoding performance across participants. Red markers at the bottom indicate lags with a significant difference between SRM and PCA-based generalization (**a**), or between SRM and original encoding (**b**), across participants ($P < 0.05$, two-sided t-test, FDR corrected). **c**, Cross-participant encoding performance in shared space for individual participants. **d**, Cross-participant encoding performance in SRM-reconstructed electrode space for individual participants. Error bands indicate the standard error of the mean encoding performance across test folds.

find W_j. The shared space S is not affected by aligning a left-out participant in this way. Now, we transform the test participant's neural activity for the test segment of the story into the shared space (estimated from other participants) using W_j estimated from the training segments of the story: $S_j^{\text{test}} = W_j^T X_j^{\text{test}}$. Finally, we evaluate the encoding models estimated from other participants' data and training segments of the story. We use the encoding weights trained on the shared space (S) to generate predictions for the left-out participant (S_j^{test}). We carry out this process for each lag for all participants (leave-one-participant-out).

Using SRM, we obtain cross-participant encoding performance (Fig. 4a, orange) comparable to the performance observed when encoding models are estimated and evaluated in individual participants (Fig. 1b). For a more direct comparison, we implemented a control analysis using PCA: we estimate PCA across $N-1$ training participants 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 participant onto reduced-dimension PCA reduced space using the left-out participant's training story segments. Finally, we estimate encoding models in the reduced-dimension PCA space from the training participants and the training story segments. We project the left-out participant's test segment into the shared space to evaluate the model-based predictions. Cross-participant encoding performance is significantly better with SRM than with PCA (Fig. 4a; $P < 0.05$, FDR corrected).

To extend this cross-participant encoding analysis from the reduced-dimension shared space to the original electrode space, we first project $N-1$ participants 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 participant j into the shared space. Now we can use W_j to project data from the estimated shared space back into the test participant's space: $X^{\text{train}} = W_j S$. This allows us to estimate encoding models based strictly on other participants' data in the test participant's original electrode space: $X^{\text{test}} = W_j W_j^T X_j^{\text{test}}$. Cross-participant encoding performance in the test participant's SRM-reconstructed electrode space significantly outperforms within-participant encoding models in the original electrode space (Fig. 4b, $P < 0.05$, FDR corrected). In both of these analyses, we show that an SRM estimated from $N-1$ participants can be used to find a set of shared features that generalize to a new participant 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.

Quantifying shared information across participants

How well can we reconstruct a novel participant's neural responses to a novel stimulus based on the neural activity of other participants? 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 $N-1$ participants, then reconstructed neural activity for

Table 1 | SRM reconstruction performance across participants

Participant	Correlation between SRM-reconstructed and original data
S1	0.336 (0.311–0.362), $P<0.001$
S2	0.268 (0.252–0.285), $P<0.001$
S3	0.258 (0.233–0.283), $P<0.001$
S4	0.251 (0.226–0.277), $P<0.001$
S5	0.313 (0.293–0.333), $P<0.001$
S6	0.368 (0.343–0.392), $P<0.001$
S7	0.216 (0.192–0.240), $P<0.001$
S8	0.260 (0.242–0.277), $P<0.001$
Mean	0.284 (0.262–0.306), $P<0.001$

Electrode activity was reconstructed based on other participants' data transformed via the shared space into the test participant's electrode space. Averaged correlations between SRM-reconstructed and original test data are reported for each participant with 95% confidence intervals and FDR-corrected *P* values across folds. Note that LLM embeddings and encoding models are not used in this analysis.

the left-out test segment in a left-out participant. We then correlated the reconstructed neural activity for the test participant j with the participant's actual neural activity. High correlations indicate that the shared space robustly captures shared information that generalizes across participants. To elaborate, first, we train an SRM model on the training data for $N-1$ participants except j using equation (1). Then, using the training data X_j for participant j, we find the matrix W_j mapping participant 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 $N-1$ participants except j using $S_{\text{test}} = 1/m \sum_{i \neq j} W_i^T X_i^{\text{test}}$. With this shared response for the test data, we reconstruct the test data for participant j (based strictly on data from $N-1$ participants) by $X_{\text{test}}^r = W_j S_{\text{test}}$. Finally, we calculate the correlation across words between the reconstructed X_{test}^r and the actual data X_{test}^j for each electrode. We repeat this process for each test participant for all the test segments. We find that SRM-based reconstruction based on the neural activity of other participants yields 0.284 correlation on average (Table 1).

Interestingly, we obtained good reconstruction performance in participant S1 ($r = 0.336$), despite the fact that all S1 electrodes were implanted in the right hemisphere, whereas most electrodes in other participants are in the left hemisphere (Fig. 2). To more rigorously test this cross-hemisphere effect, we re-evaluated right-hemisphere reconstruction performance in test participant S1 after excluding all other right-hemisphere electrodes in the training participants. Cross-hemisphere reconstruction performance in S1 was reduced, but still relatively strong at $r = 0.305$. This suggests that the stimulus-driven signals aligned by SRM are relatively bilateral in nature, given that SRM can reconstruct signals across hemispheres. Although language processing is historically associated with lateralization, the higher-level combinatorial and semantic systems for language comprehension may be largely bilateral[36]. This is corroborated by recent work in neuroimaging: for example, interparticipant correlations during naturalistic story listening are remarkably bilateral[19] (Fig. 4a), and semantic encoding during natural language comprehension is also highly bilateral[34,37].

To evaluate how well we can reconstruct a left-out participant's neural activity, we also performed a between-participant time-segment classification analysis[24]. This analysis quantifies how well we can predict one participant's neural activity patterns for specific segments of the stimulus from other participants' brain activity (see 'Time-segment classification' in the Supplementary Information for methodological details). We found that between-participant time-segment classification accuracy nearly doubled with SRM, relative to a PCA-based control (increasing from 37% with PCA to 93% with SRM using 40-word segments; chance 1.6%; Supplementary Fig. 5).

Exploring SRM encoding across different model features

Lastly, we explored encoding performance in the shared space across several different sets of model features. First, we asked whether encoding performance observed with contextual embeddings in the foregoing analyses might be driven by lower-level linguistic features (see 'Syntactic and phonetic embeddings' in the Supplementary Information for methodological details). To test this possibility, we repeated our encoding analysis in the shared space using three different types of stimulus feature[8,14,38]: (1) contextual embeddings extracted from GPT-2 XL used in the preceding analyses; (2) syntactic features capturing part-of-speech and linguistic dependencies; and (3) phonetic/articulatory features (Supplementary Fig. 6). We found that the contextual embeddings outperformed both the syntactic and phonetic features across all five shared features. This suggests that SRM is not simply keying into low-level features; the shared features appear to encode relatively high-level contextual semantic content.

We next 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 −2,000 ms to +2,000 ms relative to word onset. In both cases, we found that intermediate layers yield the highest prediction performance in human brain activity (Supplementary Fig. 7a), consistent with prior work[4–6,8]. Next, we evaluated encoding models for two different types of word embedding: contextual (GPT-2 XL) and noncontextual (GloVe) embeddings (see ref. 39; note that GloVe encoding was initially used to select electrodes). We found that contextual embeddings yield dramatically higher encoding performance than noncontextual embeddings, both for original electrode data and in the shared space (Supplementary Fig. 7b, $P < 0.01$, FDR correction), similarly to prior work[4,6,8]. Finally, we repeated our SRM encoding analysis with several open-source GPT models ranging from 125 million (M) to 20 billion (B) parameters: neo-125M, large-774M, neo-1.3B, XL-1.5B, neo-2.7B and neo-20B. Qualitatively, SRM appears to yield higher encoding performance for all models, and we observe a weak trend where larger models yield better encoding performance (Supplementary Fig. 7c), consistent with previously reported results[13,40]. In general, these exploratory analyses suggest that the improvement in encoding performance with SRM does not interact unexpectedly with different model features; however, further work is needed to investigate the relation between SRM and other factors, such as performance on language benchmark tasks[11], training diet[41] and multimodal architectures[42].

Discussion

Many recent studies have begun to use encoding models to predict neural responses during natural language processing using contextual embeddings derived from LLMs[4–7,9–11,14]. Our study demonstrates that aligning the neural activity in each brain into a shared, stimulus-driven feature space substantially enhances encoding performance. This shared space isolates stimulus-driven latent features in neural activity across both participants and electrodes, while effectively filtering out participant-specific idiosyncrasies[24,25,27]. 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 participants.

SRM and other hyperalignment methods were developed, initially with fMRI, to estimate a shared information space aligned across participants[24–28,30,31,43]. ECoG acquisition presents a more challenging correspondence problem due to varying electrode numbers and placement across participants[29]. 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[6,9,11], electrodes are simply pooled across participants to construct a 'superparticipant'. No mapping from one participant to another is attempted, and, critically, whether encoding models actually generalize across participants has not been investigated. In this Article, 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 participants[25]. 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 37% 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[6,12]. Note that more stimulus-driven areas, such as STG and IFG in the case of language processing, may make larger contributions to the shared space. Similarly, particular participants or brain regions with more electrodes or stronger signal may have a larger impact on the shared space.

The vast majority of prior work fitting electrode-wise linguistic encoding models does not evaluate whether models generalize across individual participants[6,10,11,14]. Our findings show that, by estimating both SRM and encoding models on a subset of training participants and stimuli, the shared space can be used to build encoding models that robustly generalize to new participants and stimuli. We show that cross-participant encoding performance via the shared space matches or even exceeds within-participant encoding performance. This generalization probably hinges on using a rich, naturalistic stimulus (such as a spoken story) to obtain a diverse sampling of brain states, which ultimately yields a more robust, generalizable shared space[24,27]. This kind of generalization—allowing us to precisely predict neural activity in previously unseen participants—can provide a way to circumvent the scarcity of individual participant data, which is particularly egregious with ECoG recordings in patients with epilepsy. Given a shared, naturalistic stimulus, SRM can allow us to leverage previously collected data from a larger group of participants in a single individual's idiosyncratic electrode space—which may accelerate research on individualized brain decoding and brain–computer interfaces[44,45].

Methods

Data collection and preprocessing

We recorded the neural activity of eight participants (20–48 years) using ECoG. Participants were presented with a 30-min audio podcast 'So a Monkey and a Horse Walk Into a Bar, Act One: Monkey in the Middle' from the *This American Life* podcast. All participants provided oral and written informed consent, and the study was approved by the institutional review board at New York University Langone Medical Center and Princeton University. 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. We also carried out an event-related potential analysis to confirm the precision of the transcribed word onsets (Supplementary Fig. 8). There were a total of 5,013 words in the podcast. Using the Hugging Face environment[46], we supplied the transcript to the LLM GPT-2 XL[32]. We opted to use GPT-2 XL because this model has been shown to perform well in a number of prior studies[4,6,10,13,47,48]. We extracted token-level embeddings, then averaged multiple subword tokens for each word to obtain a single embedding per word. These 1,600-dimensional contextual embeddings were 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. The model was supplied with preceding tokens up to the maximum context length of 1,024 tokens for GPT-2 XL. Note that tokens early in the stimulus are not preceded by a full 1,024 tokens (for instance, the first token is preceded by zero tokens, the second token is preceded by one token and so on), and the absolute duration of the context window (in seconds) will vary based on the speech rate. These embeddings were reduced to 50 dimensions using PCA, in keeping with prior work[6,10], for expedience in using ordinary least-squares (rather than ridge regression) when estimating the encoding models.

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. In keeping with prior work[6], we used high-gamma power as an index of local, stimulus-driven neuronal activity[49,50]. To extract the gamma power time course, we used Morlet wavelets: the power time course was computed in the 70–200 Hz frequency range separately for each frequency with 5-Hz steps. Line noise at 120 and 180 Hz was excluded using a notch filter, and we computed the logarithm of each power time course estimate. The estimates were *z*-scored and averaged across frequencies, yielding the high-gamma power time course. We then applied a 50-ms Hamming window for smoothing. Next, we used despiking to remove signal spikes exceeding four quartiles above and below the median, and used cubic interpolation for sample replacement. We re-referenced the data to account for shared signals across all channels using either common average referencing or an independent component analysis (ICA)-based method (depending on the noise profile of each participant). Finally, we divided the power estimates by the mean value to improve the signal-to-noise ratio[16].

Encoding models

We use contextual word embeddings to predict held-out neural data for individual electrodes or SRM features (see below). First, we averaged neural activity (gamma power) in 200-ms windows at 161 lags ranging from −2,000 ms to +2,000 ms in 25-ms increments for epochs indexed to each word's onset. These parameters were chosen to match prior works[6,10]. 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[33,34]. Encoding models were formulated to predict variance across words separately at each lag relative to word onset. We use tenfold 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 two-sided *t*-tests between the two correlation scores (within participant) across folds for each lag. We also used one-sample *t*-tests to compare reconstruction performance (across participants) against zero. For both within-participant and group-level analyses, we controlled the false discovery rate (FDR) at least at a threshold of 0.05 to correct for multiple statistical tests across varying lags or electrodes (ref. 51).

Electrode selection

To select a subset electrodes involved in language processing, following ref. 6, we first estimated encoding performance using noncontextual GloVe embeddings[39] at 161 lags ranging from −2,000 ms to +2,000 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 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 FDR at 0.01; if an electrode survived FDR correction, it was selected for further analysis. This yielded 184 electrodes (150 in the left hemisphere, 34 in the right hemisphere) across participants (see Supplementary Table 1 for a detailed description of electrode coverage).

Shared response model

Although all the participants 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 SRM[25] to aggregate ECoG data across participants into a common information space that accounts for different electrode placement and functional topographies across individuals. SRM learns participant-specific transformations that map from each participant'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 (see ref. 52; Fig. 4). In the current work, the SRM was estimated so as to recover shared features fluctuating across words. SRM was estimated and evaluated separately at each lag relative to word onset.

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 participants. We use this training dataset to learn participant-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 participant'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 participants. For the solution to be unique, W_i is subject to the constraint of linearly independent and orthonormal columns, $W_i^T W_i = I_k$ (ref. 25). The following optimization problem is solved to estimate W_i and the shared response S:

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

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

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

Reporting summary

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

Data availability

Sample data are available via Zenodo at https://zenodo.org/records/15220273 (ref. 53), and the full raw dataset is publicly available at https://openneuro.org/datasets/ds005574/versions/1.0.0 (ref. 54). Source data are provided with this paper.

Code availability

Code used to analyze the data is publicly available via GitHub at https://github.com/pritamarnab/SRM-Encoding (ref. 55).

References

1. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).
2. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
3. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl Acad. Sci. USA* **117**, 30046–30054 (2020).
4. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
5. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
6. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
7. Toneva, M., Mitchell, T. M. & Wehbe, L. Combining computational controls with natural text reveals aspects of meaning composition. *Nat. Comput. Sci.* **2**, 745–757 (2022).
8. Kumar, S. et al. Shared functional specialization in transformer-based language models and the human brain. *Nat. Commun.* **15**, 5523 (2024).
9. Cai, J., Hadjinicolaou, A. E., Paulk, A. C., Williams, Z. M. & Cash, S. S. Natural language processing models reveal neural dynamics of human conversation. *Nat. Commun.* **16**, 3376 (2025).
10. Goldstein, A. et al. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nat. Hum. Behav.* **9**, 1041–1055 (2025).
11. Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D. & Mesgarani, N. Contextual feature extraction hierarchies converge in large language models and the brain. *Nat. Mach. Intell.* **6**, 1467–1477 (2024).
12. Goldstein, A. et al. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat. Commun.* **15**, 2768 (2024).
13. Hong, Z. et al. Scale matters: large language models with billions (rather than millions) of parameters better match neural representations of natural language. *eLife* **13**, RP101204 (2024).
14. Zada, Z. et al. A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron* **112**, 3211–3222 (2024).
15. Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
16. Honey, C. J. et al. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).
17. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
18. Nastase, S. A., Gazzola, V., Hasson, U. & Keysers, C. Measuring shared responses across subjects using intersubject correlation. *Soc. Cogn. Affect. Neurosci.* **14**, 667–685 (2019).
19. Nastase, S. A. et al. The 'Narratives' fMRI dataset for evaluating models of naturalistic language comprehension. *Sci. Data* **8**, 250 (2021).

20. Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
21. Nieto-Castañón, A. & Fedorenko, E. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage* **63**, 1646–1669 (2012).
22. Braga, R. M., DiNicola, L. M., Becker, H. C. & Buckner, R. L. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *J. Neurophysiol.* **124**, 1415–1448 (2020).
23. Lipkin, B. et al. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Sci. Data* **9**, 529 (2022).
24. Haxby, J. V. et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
25. Chen, P.-H. et al. A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems* Vol. 28 (eds Cortes, C. et al.) (Curran Associates, 2015).
26. Guntupalli, J. S. et al. A model of representational spaces in human cortex. *Cerebral Cortex* **26**, 2919–2934 (2016).
27. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
28. Feilong, M. et al. The Individualized Neural Tuning Model: precise and generalizable cartography of functional architecture in individual brains. *Imag. Neurosci.* **1**, 1–34 (2023).
29. Owen, L. L. W. et al. A Gaussian process model of human electrocorticographic data. *Cereb. Cortex* **30**, 5333–5345 (2020).
30. Van Uden, C. E. et al. Modeling semantic encoding in a common neural representational space. *Front. Neurosci.* **12**, 378029 (2018).
31. Nastase, S. A., Liu, Y.-F., Hillman, H., Norman, K. A. & Hasson, U. Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *NeuroImage* **217**, 116865 (2020).
32. Radford, A. et al. *Language Models Are Unsupervised Multitask Learners* (OpenAI Blog, 2019).
33. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
34. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
35. Desikan, R. S. et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).
36. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
37. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**, 6539–6557 (2017).
38. Honnibal, M. et al. spaCy: industrial-strength natural language processing in Python. *Zenodo* https://doi.org/10.5281/zenodo.1212303 (2020).
39. Pennington, J., Socher, R. & Manning, C. GloVe: global vectors for word representation. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (eds Moschitti, A., Pang, B. & Daelemans, W.) 1532–1543 (Association for Computational Linguistics, 2014).
40. Antonello, R., Vaidya, A. & Huth, A. Scaling laws for language encoding models in fMRI. In *Advances in Neural Information Processing Systems* Vol. 36 (eds Oh, A. et al.) 21895–21907 (Curran Associates, 2023).
41. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 9383 (2024).
42. Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J. & Wehbe, L. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nat. Mach. Intell.* **5**, 1415–1426 (2023).
43. Feilong, M., Nastase, S. A., Guntupalli, J. S. & Haxby, J. V. Reliable individual differences in fine-grained cortical functional architecture. *NeuroImage* **183**, 375–386 (2018).
44. Metzger, S. L. et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037–1046 (2023).
45. Willett, F. R. et al. A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).
46. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, Online, 2020).
47. Kauf, C., Tuckute, G., Levy, R., Andreas, J. & Fedorenko, E. Lexical-semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network. *Neurobiol. Lang.* **5**, 7–42 (2024).
48. Tuckute, G. et al. Driving and suppressing the human language network using large language models. *Nat. Hum. Behav.* **8**, 544–561 (2024).
49. Manning, J. R., Jacobs, J., Fried, I. & Kahana, M. J. Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *J. Neurosci.* **29**, 13613–13620 (2009).
50. Jia, X., Tanabe, S. & Kohn, A. Gamma and the coordination of spiking activity in early visual cortex. *Neuron* **77**, 762–774 (2013).
51. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
52. Cohen, J. D. et al. Computational approaches to fMRI analysis. *Nat. Neurosci.* **20**, 304–313 (2017).
53. Bhattacharjee, A. ECoG data of 8 subjects listening to a podcast. *Zenodo* https://doi.org/10.5281/zenodo.15220273 (2025).
54. Zada, Z. et al. The 'podcast' ECoG dataset for modeling neural activity during natural language comprehension. *Sci. Data* **12**, 1135 (2025).
55. Bhattacharjee, A. Software for the paper titled aligning brains into a shared space improves their alignment to large language models. *Zenodo* https://doi.org/10.5281/zenodo.15644439 (2025).

Acknowledgements

We thank our funders: NIH grant DP1HD091948 (U.H.), NIH grant R01NS109367 (A.F.), NIH CRCNS R01DC022534 (U.H.) and J. Insley Blair Pyne Fund (P.J.R. and U.H.)

Author contributions

Conceptualization, A.B., S.A.N., A.G. and U.H.; data curation, B.A., W.D., D.F., P.D., A.F. and O.D.; formal analysis, A.B.; funding acquisition, P.J.R. and U.H.; methodology, A.B., S.A.N. and U.H.; project administration, A.B., S.A.N., P.J.R. and U.H.; software, A.B.; supervision, S.A.N. and U.H.; visualization, A.B., S.A.N. and U.H.; writing—original draft, A.B. and S.A.N.; writing—review and editing, A.B., S.A.N., Z.Z., H.W. and U.H.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00900-y.

Correspondence and requests for materials should be addressed to Arnab Bhattacharjee.

Peer review information *Nature Computational Science* thanks Mark Lescroart, Jeremy R. Manning and Alex Murphy for their contribution to the peer review of this work. Primary Handling Editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.