Fine-tuning speech models on slow fMRI improves prediction for fast ECoG

Aditya R. Vaidya
Department of Computer Science
The University of Texas at Austin, USA
avaidya@utexas.edu

Richard J. Antonello
Zuckerman Mind Brain Behavior Institute
Columbia University, USA
rja2163@columbia.edu

Alexander G. Huth
Departments of Neuroscience and Statistics
University of California, Berkeley, USA
alex.huth@berkeley.edu

Abstract

Neuroscientists have recently turned to intracranial brain recording methods, like electrocorticography (ECoG), for human experiments because of the fine spatial and temporal resolution that they afford. Models trained on this data, however, are fundamentally restricted by the patient populations that can receive the implants necessary for recording. We propose using non-invasive fMRI to bridge the gap in training data. Using spoken language representations fine-tuned on fMRI, we build encoding models of ECoG. These representations show improved prediction performance in ECoG, even though the temporal resolution of fMRI is two orders of magnitude worse. Prediction improves in frequency bands well beyond what is directly measured in fMRI. Next, to test the procedure's generalization ability, we fine-tune models on fMRI responses that are temporally downsampled by a factor of 2. Despite the loss in resolution, these models are able to predict fMRI and ECoG responses at levels comparable to the original fMRI-tuned models. Finally, we show that ECoG performance steadily scales with the amount of fMRI-tuning data. Our results show that "slow" data like fMRI can be a valuable resource for building better models of "fast" brain data like ECoG. In the future, integrating across multiple recording methods may further improve performance in other applications, like decoding. Code is available at: HuthLab/fmri-ecog-transfer

1 Introduction

Many neuroscience experiments use data recorded from intracranial electrodes that have been surgically implanted in human patients as part of clinical treatment. In electrocorticography (ECoG) the electrodes are arranged in grids that are applied directly to the surface of the brain. Electrodes in such close proximity to the brain tissue record neural activity with high temporal and spatial precision, enabling the construction of detailed encoding models that predict brain activity elicited by a stimulus. These encoding models can then be used to answer neuroscience research questions [1, 2] or in downstream applications like brain-computer interfaces (BCIs) [3, 4].

The downside of intracranial recordings is that datasets are small and rare. ECoG electrodes are only implanted when clinically necessary, and are typically removed within a few days [5, 6]. The number of electrodes and their placement also differs from patient to patient, depending on their clinical needs. This severely limits the amount of data collected from each individual, making it challenging to build elaborate encoding models. In contrast, non-invasive measurements like functional magnetic

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

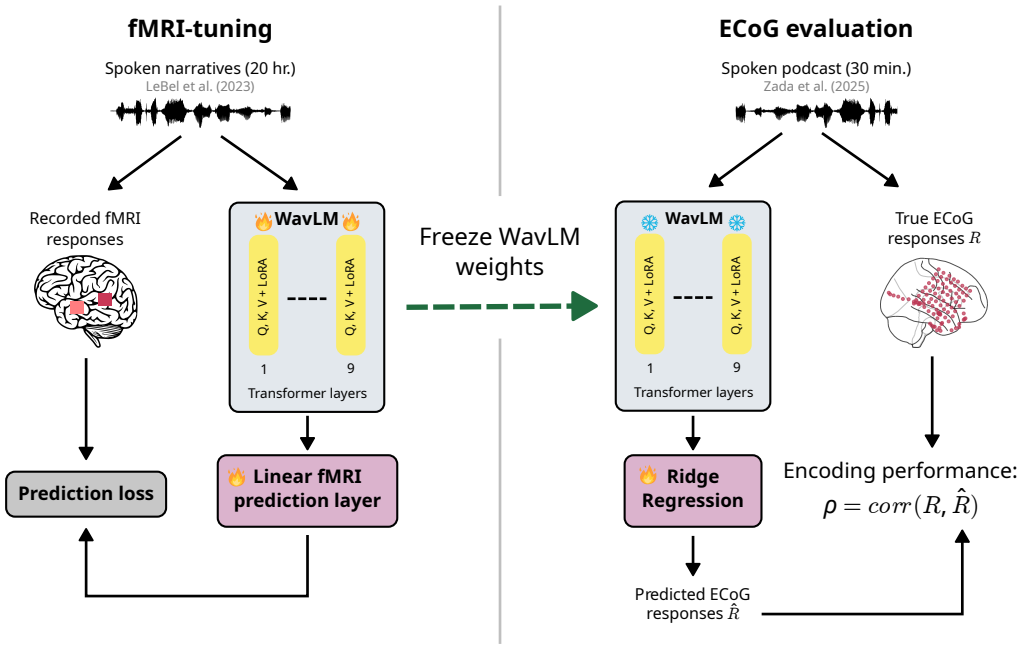


Figure 1: **fMRI-to-ECoG transfer via fMRI-tuning.** We fine-tune layers 1–9 of a deep speech representation model, WavLM Base+ [21], to predict fMRI responses (measured at 0.5 Hz) to spoken language. We then freeze the weights of the WavLM model and use its representations to build linearized encoding models of ECoG responses (analyzed at 20 Hz) to speech from a separate dataset. Successfully performing this task requires learning representations that are useful across brain recording methods and robust to new subjects and stimuli.

resonance imaging (fMRI)—while having much lower temporal resolution than ECoG—are easy to acquire repeatedly and have whole-brain coverage. Can plentiful fMRI data be exploited to improve encoding models for intracranial data?

Modern encoding models are typically based on pretrained deep neural networks. Language encoding models often use neural network language models pretrained on large text corpora [7, 8] or speech models pretrained on audio corpora [9, 10, 11]. These networks have mostly been used as frozen feature extractors [12, 13], but in more recent "brain-tuning" studies the networks are also fine-tuned on brain data to improve encoding performance [14, 15]. fMRI-tuned models increase prediction performance relative to pretrained models [16], and also generalize to fMRI data from new subjects or different brain areas [17]. While the limited size of ECoG datasets makes fine-tuning challenging, it may be possible that fMRI-tuned models, by virtue of learning more brain-like representations, generalize to ECoG. This would enable ECoG encoding model performance to scale with available fMRI datasets.

The idea that ECoG encoding models could benefit from fine-tuning on fMRI data is surprising but not impossible. ECoG measures electrical activity in the brain that varies at the scale of milliseconds, while fMRI measures blood flow that varies over seconds—at least 2 orders of magnitude slower. Yet despite its low sampling rate, fMRI signals are sensitive to fast timescale stimulus properties. In auditory cortex, for example, fMRI responses can depend upon temporal modulations in sound that are much faster than the fMRI signal itself [18, 19]. More recent studies have shown that auditory cortex fMRI responses to speech are well-modeled using networks like HuBERT [20], WavLM [21], or Whisper [22]. These models' fMRI prediction performance is improved by fine-tuning on fMRI data [14, 16]. And the same networks can also very effectively predict ECoG responses to speech [23]. These results all support the possibility that fMRI-tuned models could improve ECoG prediction performance.

In this work, we demonstrate that speech models fine-tuned on fMRI can generalize to new subjects, stimuli, and recording methods (Figure 1). Models fine-tuned on fMRI improved prediction performance in ECoG, despite the difference in temporal resolution between the two methods. We stress test the "temporal resolution generalization" of this procedure by fine-tuning on downsampled fMRI responses. Despite a sampling rate of only 0.25 Hz, fine-tuning on these responses still yielded a significant improvement in ECoG prediction performance compared to the pretrained model. We show that this slow-to-fast generalization even applies within fMRI data; models fine-tuned on downsampled fMRI responses were able to predict the original fMRI responses better than the pretrained model. Finally, we show that ECoG performance steadily increases with the amount of fMRI fine-tuning data, demonstrating that scaling fMRI language datasets could benefit ECoG models for applications in neuroscience or BCI, despite the vast difference in temporal resolution.

2 Data and methods

2.1 fMRI data

From the public dataset released by LeBel et al. (2023) [24] and Tang et al. (2023) [3], we used preprocessed fMRI data from 3 participants who were scanned while listening to 94–103 naturally spoken narrative stories (17.8 h–19.7 h per participant). These stories capture a range of responses that may be easier to capture in whole-brain recordings (like fMRI) than in more limited intracranial datasets. Three of the stories were presented to the subjects multiple times: two ("fromboyhoodtofatherhood" and "onapproachtopluto") were heard five times each, and one ("wheretheressmoke") ten times. We averaged across the responses of the repeated presentations to reduce measurement noise.

2.2 Speech encoding models

In this work, we build linearized speech encoding models that aim to estimate the response R to a stimulus S as:

$$\hat{R}_t = f(S_t; \theta)\beta \tag{1}$$

where $f : \mathbb{S} \to \mathbb{R}^F$ is a non-linear transform of the stimulus S at time t with parameters θ, and $\beta \in \mathbb{R}^{F \times C}$ is a linear projection of the F features to each of C channels of the brain response. Response channels are voxels in fMRI and electrodes in ECoG. When possible, encoding models use a finite impulse response (FIR) structure to capture temporal properties of the response. For example, fMRI responses derive from the blood-oxygen-level-dependent (BOLD) signal which, after an impulse of neural activity, rises to a peak over the course of 3-4 s and then falls back to baseline over another 4-6 s [25]. To capture this behavior, our fMRI encoding models use concatenated stimulus features from several timepoints (TRs $t-4$, $t-3$, $t-2$, and $t-1$) to predict the responses at time t. For an FIR model with d delays, this expands the feature space to $d \cdot F$ dimensions. In ECoG, responses are also delayed by upstream neural processing, often spanning 200-800 ms after a stimulus transient [26, 27]. Because the sampling rate of ECoG data is also much higher, capturing this temporal response could require $d = 20$ or more. This quickly becomes computationally intractable for large feature spaces. To circumvent this issue, we followed earlier studies [28, 29] and fit ECoG encoding models that use only a single delay hyperparameter τ, yielding $\hat{R}_t = f(S_{t-\tau}; \theta)\beta$.

Here, we parametrize the non-linear stimulus transform f using WavLM [21], a neural network that operates on the waveform of the stimulus. Layer 9 has been previously shown to have the highest encoding performance in fMRI [30], so we only extract features from this layer. Following previous work, we extract features by sliding a 4 s window over the waveform, feeding the windowed stimulus into the model, and saving the hidden state of the final token of the layer.

For modeling fMRI, we extract features with a 0.25 s stride, and the result is downsampled to 0.5 Hz, the sampling rate of the fMRI responses. For modeling ECoG, we extract features with a 0.05 s stride to match the 20 Hz sampling rate used for the ECoG responses.

2.3 fMRI-tuning of audio models

To induce a brain-like bias in the speech representations, we fine-tune the underlying WavLM model to predict fMRI responses in a procedure we call "fMRI-tuning". We adopt the procedure of Vattikonda et al. (2025) [16] to fine-tune separate WavLM models to predict the fMRI responses of three subjects

from the fMRI dataset, starting from the WavLM Base+ checkpoint. To reduce the likelihood of overfitting, we use rank-4 low-rank adaptation (LoRA [31]) on the W^Q, W^K, W^V matrices of each of the first nine Transformer layers, and we constrain the final linear projection from the ninth layer of WavLM to fMRI voxels to be rank-100. We use the Adam optimizer [32] with a learning rate of 5×10^{-4} to optimize the LoRA matrices and linear projection to minimize a spatial correlation loss.

We use two stories as a validation set ("fromboyhoodtofatherhood" and "onapproachtopluto") and, as we do for the pretrained model, evaluate encoding performance on one test story ("wheretheressmoke"). We fine-tuned each model for 30 epochs with a batch size of 10 TRs using the same feature extraction parameters as described in Section 2.2. To select the best epoch, we evaluated encoding performance on the validation set. Using the ridge parameters from the pretrained model, we fit ridge regression encoding models using features from each of the 30 epochs, and we selected the epoch with the best validation encoding performance. For this epoch, we then re-ran cross-validation to select the best ridge parameters, and we evaluated its encoding performance on the unseen test story.

Fine-tuning one WavLM model took 30 hours on a single 48GB NVIDIA RTX A6000.

To test the generality of our results, we also fine-tuned the encoder of `whisper-large-v3` [22] on fMRI responses to audiovisual stimuli from the CNeuroMod *Friends* dataset [33, 34]. We used seasons 1–5 and did not use LoRA.[1] This fMRI-tuned Whisper model differs from the fMRI-tuned WavLM models in several ways. First, Whisper is originally trained for speech transcription, rather than being self-supervised. Second, the model is simultaneously fine-tuned on multiple fMRI subjects instead of one. Third, the responses in NeuroMod are to audiovisual stimuli rather than purely audio as in LeBel et al., though we only use the audio for fMRI-tuning.

2.4 ECoG data and evaluation

We use the "Podcast" dataset [29] to evaluate the generalization of our fMRI-tuned models to intracranial data. In this dataset, nine patients listened to a 30-minute podcast while electrical brain activity was recorded with electrocorticography (ECoG). The dataset contains 1,268 electrodes across all patients, who are distinct from the participants in the fMRI dataset.

As is conventional for ECoG, we used the high-gamma power (provided by [29]) in each ECoG electrode as the response of interest [35, 36]. Power in the high-gamma band was computed as the analytic amplitude of the Hilbert transform of ECoG signals that had been band-pass filtered in the 70-200 Hz band. High-gamma power is thought to represent pooled activity across synchronously active neurons in the region near the electrode. We further downsampled the high-gamma signals to 20 Hz to reduce computational burdens.

As mentioned earlier, the high temporal resolution of ECoG and high dimensionality of the speech representations make it computationally difficult to fit FIR encoding weights. Instead, we offset the stimulus different amounts relative to the response, and build separate linear models for each lag with ridge regression. 81 lags are evenly spaced between -2 and $+2$ seconds. ECoG encoding performance is calculated with 4-fold cross-validation on the 30-minute stimulus. The encoding performance for a model is that of the best performing lag, selected independently per electrode.

Though fitting the FIR encoding model on the full feature space is too expensive, we also fit some FIR models with a reduced feature space. Using principal components analysis (PCA), we select the first 100 dimensions of the WavLM features and fit 40-delay FIR models to independently predict each ECoG electrode.

Along with encoding performance (Pearson correlation coefficient) on the overall signal, we also evaluate performance within individual frequency bands by computing the power spectral density (PSD) of the residual of the model's predictions. The residual PSD for a model is chosen to be that of the lag with the highest encoding performance.

Fitting all ECoG encoding models for one fMRI-tuned model took 10 hours on a single NVIDIA RTX A6000.

[1] The model checkpoint is available here:
`https://huggingface.co/RAntonello/whisper-large-v3-braintuned`

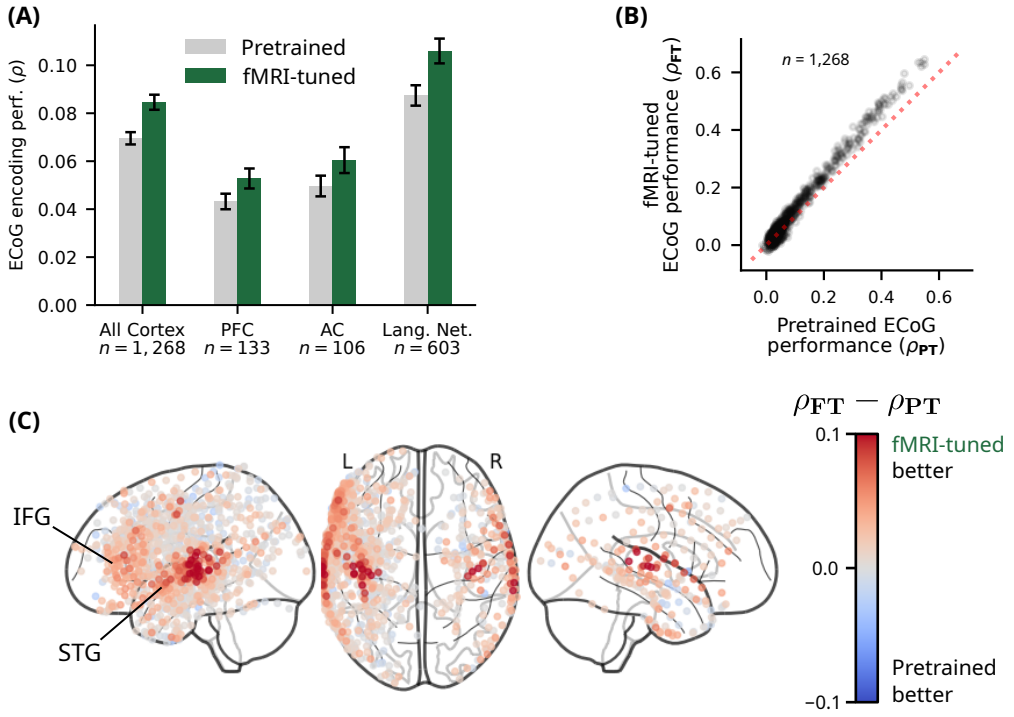


Figure 2: **fMRI-tuned models for predicting ECoG.** (**A**) ECoG encoding performance averaged across all electrodes within regions of interest (ROIs). Error bars show the standard error of the mean (SEM) across electrodes. The Destrieux atlas [38] was used to find electrodes in each ROI: prefrontal cortex (PFC), primary auditory cortex (AC), and the language network as described in Lipkin et al. (2022) [39] (see Appendix A.1 for details). fMRI-tuned models outperform pretrained models in every ROI ($p < 1 \times 10^{-8}$ for paired t-tests across electrodes within the ROI). (**B**) Encoding performance of each electrode using the pretrained vs. fMRI-tuned models. fMRI-tuned models consistently outperform pretrained models, especially for electrodes that were already well-modeled ($p < 1 \times 10^{-87}$ for a paired t-test across all electrodes). (**C**) Change in encoding performance for each electrode after fMRI-tuning, visualized on a brain atlas. Electrodes with the greatest improvement are found in bilateral auditory cortex. Smaller gains are seen in other components of the language network, including the superior temporal gyrus (STG) and inferior frontal gyrus (IFG). In Appendix A.2, we separate the electrodes by subject (Figure 6).

2.5 Signal-to-noise ratio of fMRI responses

fMRI data comprise both signal and noise. To tease the two apart one must use data collected while the same stimulus is played several times. Averaging the response timecourse across repeats gives an estimate of the "signal" portion of the data. Subtracting that average from each individual recording then gives an estimate of the "noise" part. Computing the variance of the signal and noise components then enables one to directly estimate the total signal to noise ratio (SNR). We estimate the SNR within each frequency band in a similar fashion: compute the power spectral densities (PSD) of the signal and noise components, and then compute the ratio at each frequency [37]. We use the 10 individual presentations of the test story ("wheretheressmoke") to compute band-wise SNR in our fMRI data.

3 Results

3.1 Fine-tuning on fMRI improves ECoG performance

Following Vattikonda et al. (2025) [16], we fine-tuned separate WavLM models on responses from each fMRI participant. Using features from these fine-tuned models, we fit new encoding models to ECoG responses from the "Podcast" dataset.

Averaged across cortex, speech models fine-tuned on fMRI outperform the pretrained WavLM model when predicting ECoG (Figure 2A). fMRI-tuning shows significant improvements for 7 of 9 subjects (see Appendix A.2). Allowing for the expressivity of an FIR encoding model, rather than a single delay, fMRI-tuning still outperforms the pretrained model (see Appendix A.3 for details). In Appendix A.4, we compare the WavLM models against other fMRI and stimulus-only baselines. fMRI-tuned Whisper also improves performance in predicting ECoG, from 0.0683 to 0.0702 (2.8%), despite differences in initial training objective, number of fMRI-tuning subjects, and stimulus modality.

Improvements are most pronounced in electrodes within auditory cortex (AC) and the "language network", a set of areas hypothesized to be selective for language processing [39]. Electrodes that did not improve after fine-tuning had poor pretrained performance as well, suggesting their response is not stimulus-driven (Figure 2B).

In Figure 2C, we show the change in performance after fMRI-tuning for each electrode on an atlas brain map. Despite evaluating on a different recording modality, the cortical areas that improve in ECoG after fine-tuning largely overlap with previous fMRI-specific findings [15, 16, 17].

3.2 fMRI-tuning improves prediction in high-frequency bands of high-gamma

Responses in auditory cortex (AC) tend to follow spectral features of the stimulus and can fluctuate quickly [26, 40]. Interestingly, electrodes in AC also showed the greatest improvements after fMRI-tuning. This led us to ask what parts of the neural response were better predicted after fMRI-tuning. We characterized the spectral properties of the model's performance by examining the power spectral density of the residuals from the encoding model predictions. Because fMRI responses are temporally smooth, one might expect fMRI-tuning to only improve predictions for slowly-varying components of the ECoG response at or below the fMRI Nyquist rate (0.25 Hz). However, these fMRI signals still ultimately derive from neural responses to rapidly-changing speech stimuli, which we know are well-modeled by the WavLM features. Thus it is also possible that fine-tuning captures or accentuates brain-relevant information at much finer timescales, leading to improvements in ECoG prediction performance above 0.25 Hz.

Surprisingly, we observed that fMRI-tuning improved the prediction of ECoG across the entire spectrum (Figure 3). The biggest improvements were at frequencies below 0.25 Hz, but substantial and significant improvements were seen at frequencies up to 1 Hz. Even in the 1-10 Hz band we found small but consistent improvement from fMRI-tuning. The power of the error reduced by 1.16% below 0.25 Hz and 0.602% above it. This result indicates that models tuned to predict fMRI data also predict ECoG at timescales that fMRI cannot resolve.

3.3 Models fine-tuned on downsampled fMRI generalize to faster responses

In the previous sections we showed that fMRI-tuned models generalize to the higher temporal resolution of ECoG. How far can we push this temporal generalization? To further stress-test our approach we created a new dataset at even lower temporal resolution by downsampling the fMRI data to 0.25 Hz (4 seconds between timepoints). We then fine-tuned new models on these downsampled fMRI responses as before.

First, we compared how well the original fMRI-tuned and downsampled models predict held-out fMRI data. Figure 4A shows the difference in model prediction performance in one fMRI subject (S3). We found that models fine-tuned on downsampled responses perform similarly to the original models overall, with slight decreases in performance in some higher-level areas like prefrontal cortex (PFC) and the boundaries of the occipital and temporal lobes. Overall prediction performance was comparable, and both models outperformed the pretrained model in all brain areas (Figure 4B).

While surprising, this result can be explained by the signal-to-noise ratio (SNR) properties of fMRI responses to natural stimuli. To explore this idea, we measured the SNR of the fMRI data across different frequency bands using a spectral approach [37] (see Section 2.5). This showed that not all frequency bands of the fMRI response are equally informative. Most of the usable signal lies in the band between 0.01 and 0.1 Hz, while very little falls in the highest frequencies (Figure 4C). Thus, by downsampling and truncating the noisiest frequencies, we actually *raise* the overall SNR of the

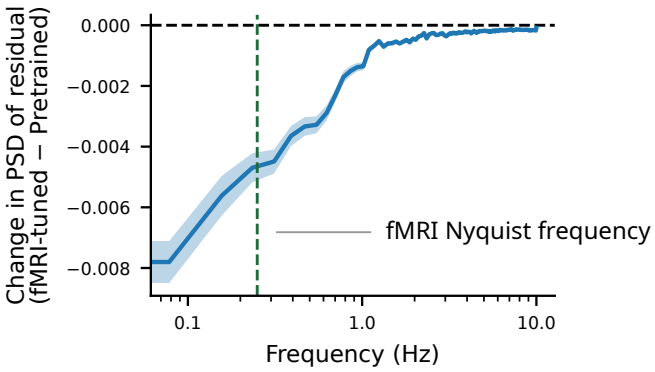


Figure 3: **Frequency-binned improvement of fMRI-tuned models.** Change in the power spectral density (PSD) of the ECoG encoding model residual after fMRI-tuning (lower is better), averaged across all electrodes. The shaded area shows the standard error across electrodes. The green dashed line at 0.25 Hz shows the Nyquist frequency of the fMRI responses. fMRI-tuning improves model fit (reduces the residual power) overall, with improvement both below and above the fMRI Nyquist frequency (two-sided t-test across electrodes between pretrained and fMRI-tuned PSD, $p < 5 \times 10^{-5}$ for every frequency). The power of the residual below 0.25 Hz decreases by 1.16% ($p < 1 \times 10^{-23}$, two-sided t-test), and above 0.25 Hz the power decreases by 0.602% ($p < 1 \times 10^{-31}$, two-sided t-test). This demonstrates that models fine-tuned on fMRI data can generalize and improve prediction performance of responses much faster than can be measured in fMRI.

fMRI responses. Fine-tuning on these "cleaner" downsampled responses allows the representations to generalize to the original signal.

Finally, we tested whether the models fine-tuned on downsampled fMRI data could still generalize to ECoG data. Figure 4D compares ECoG encoding model performance using fMRI-tuned and downsampled-fMRI-tuned models. As was the case predicting fMRI data, we found virtually no difference between the two models at predicting ECoG data. This demonstrates even more extreme temporal generalization than before: models fine-tuned on fMRI data with a Nyquist rate of 0.125 Hz improve performance on ECoG data.

3.4 ECoG prediction scales with fMRI data

Several works have examined the scaling laws of encoding models in brain data [41, 42]. In particular, Antonello et al. (2023) [30] showed that language encoding models in fMRI scale logarithmically with the amount of training data. Moussa et al. (2025) [17] followed up this finding by fine-tuning the underlying models, and they found that fine-tuning reduces the amount of data needed to generalize to new subjects. We next examined whether this effect holds across modalities as well. To test scaling, we fine-tuned models using subsets of the LeBel et al. fMRI dataset, and then tested them on the "Podcast" ECoG dataset [29].

We found that improvements in overall ECoG prediction scaled logarithmically with the number of stories used for fMRI-tuning (Figure 5A). This result mirrors the literature [30, 17], but with smaller improvement magnitudes due to transferring across modalities. (In Appendix C.1, we additionally compare the scaling behavior of linear and fMRI-tuned models on fMRI encoding.)

We next quantified the scaling law of fMRI-tuning data on ECoG. Adapting the formulation of Antonello et al. (2023), for each electrode we fit a linear model $\Delta\rho_e \approx m_e \log_2 N$ to predict the change in encoding performance $\Delta\rho_e$ from the number of fMRI-tuning stories N. In Figure 5B, we show the scaling coefficient m_e for all electrodes. Our trend was consistent across electrodes; after filtering for language-responsive electrodes, 176 of 219 electrodes across cortex had a positive relationship between encoding performance and the number of stories used for fMRI-tuning (Figure 5B). Scaling effects were largest in areas with strong baseline encoding performance (Figure 5C).

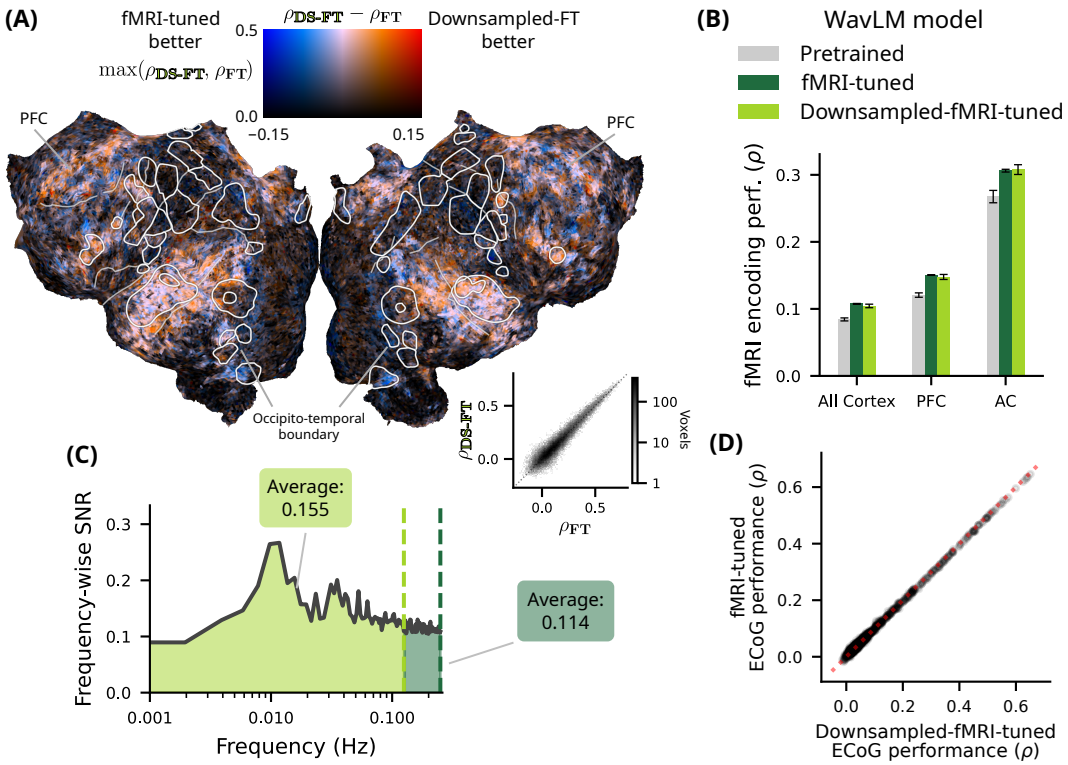


Figure 4: **Fine-tuning on downsampled fMRI responses.** (**A**) Flattened cortical surface of fMRI subject S3 that compares the encoding performance of models fine-tuned on original fMRI responses (blue) and downsampled fMRI responses (red). Voxel brightness is proportional to overall model performance. The 2-dimensional histogram shows that voxels have similar performance across fine-tuning conditions. We show subjects S1 and S2 in Appendix B. (**B**) fMRI encoding performance, averaged within ROIs, after fine-tuning on original or downsampled fMRI responses. Error bars show standard error across subjects. There was no significant effect between the two fine-tuning conditions (paired t-test across subjects had p-values 0.424, 0.565, and 0.798 for each respective ROI). (**C**) Signal-to-noise ratio (SNR) of the fMRI responses, averaged across subjects. The dark and light green vertical lines are the Nyquist frequency of the original and downsampled fMRI responses at 0.25 Hz and 0.125 Hz, respectively. The average SNR is 0.155 ± 0.0143 below the 0.125 Hz threshold and 0.114 ± 0.000534 above it. The original signal had an overall SNR of 0.139 ± 0.00895. This shows why downsampling the fMRI data before fine-tuning had little effect on model performance: most of the responses between $0.125 - 0.25$ Hz are noise. (**D**) Effect on ECoG performance after fine-tuning on downsampled or original fMRI responses. We see no significant change in performance due to fine-tuning data ($p = 0.514$, two-sided paired t-test across all electrodes).

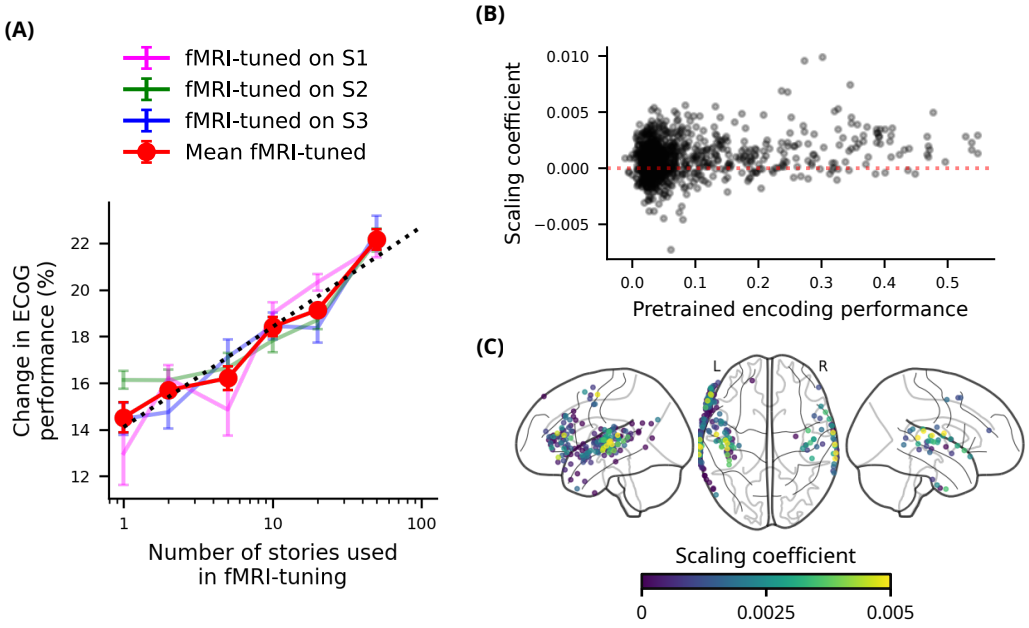


Figure 5: **Scaling of fMRI-tuning for ECoG.** (**A**) ECoG encoding performance as a function of the number of fMRI fine-tuning stories. Error bars indicate the standard error over bootstraps of the fMRI stories. We show the scaling performance with the full fMRI-tuning dataset in Appendix C.2. (**B**) Pretrained encoding performance vs. scaling coefficient m_e for all electrodes. The scaling law is measured per electrode as the slope of a linear regression between (1) its encoding performance and (2) log-fMRI story count. The electrodes that tend to improve have strong pretrained performance. (**C**) Electrodes were thresholded by pretrained encoding performance $\rho > 0.1$, yielding $n = 219$ electrodes, and plotted on a brain atlas.

Our results extend the literature by showing that fMRI-tuned models not only generalize across subjects, but also across recording methods. These results suggest that increasing amounts of fMRI data could be used to mitigate the limitations of intracranial data collection.

4 Conclusion

In this work, we showed that speech representations fine-tuned with fMRI data transfer to ECoG prediction, a recording modality with orders of magnitude higher temporal resolution. Transfer succeeds even when the fMRI data are further downsampled to 0.25 Hz, and performance scales logarithmically with the amount of fMRI fine-tuning data.

How is this possible? High-gamma power, the ECoG signal that we model, is thought to be the closest neural correlate to the BOLD signal measured in fMRI [43]. Thus it is plausible that fine-tuning on fMRI accentuates features that are relevant for high-gamma, even at response frequencies invisible to fMRI. This shared underpinning may be sufficient to enable slow-to-fast model transfer.

ECoG datasets are typically small: clinically necessary electrodes are implanted for a few days and then removed. fMRI datasets, in contrast, face no such constraints. Thus, as fMRI datasets continue to grow, our scaling results suggest that leveraging them will provide better and better ECoG encoding models. This is relevant for BCI applications where encoding model quality limits performance.

Still, several limitations apply to these results. We primarily focused on two models (WavLM Base+ and Whisper large-v3) and the single task domain of spoken language. Our approach is probably not fully data-optimal: there is also a rich landscape of training recipes involving nonlinear encoding adapters for ECoG data which we do not explore in this work, and there are other known means of fine-tuning acoustic representations to be more brain-like, such as fine-tuning on phonemes or semantic representations [16]. A wider search of this landscape would help clarify the optimal

means by which fMRI data can be used to predict ECoG data. Yet it also seems likely that the optimal approach would combine all available data modes: fine-tune on fMRI *and* ECoG *and* other potential information sources (like electroencephalography (EEG)) in a joint training cocktail. Future work (and larger datasets) will be required to fully explore this space. Broadly, our work presents strong evidence that brain-like representations are highly modality invariant, and that out-of-modality training data can be a core component of the training mix for future brain encoding models.

Acknowledgements

This project was supported by NIDCD grant R01DC014279. Computing resources were provided by the Zuckerman Institute at Columbia University and the Texas Advanced Computing Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343(6174):1006–1010, February 2014.

[2] Menoua Keshishian, Gavin Mischler, Samuel Thomas, Brian Kingsbury, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. Parallel hierarchical encoding of linguistic representations in the human auditory cortex and recurrent automatic speech recognition systems. *Nature Machine Intelligence*, 8(2):257–269, February 2026.

[3] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, May 2023.

[4] Kaylo T. Littlejohn, Cheol Jun Cho, Jessie R. Liu, Alexander B. Silva, Bohan Yu, Vanessa R. Anderson, Cady M. Kurtz-Miott, Samantha Brosler, Anshul P. Kashyap, Irina P. Hallinan, Adit Shah, Adelyn Tu-Chan, Karunesh Ganguly, David A. Moses, Edward F. Chang, and Gopala K. Anumanchipalli. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature Neuroscience*, pages 1–11, March 2025.

[5] Edward F. Chang. Towards Large-Scale, Human-Based, Mesoscopic Neurotechnologies. *Neuron*, 86(1):68–78, April 2015.

[6] Manuel R. Mercier, Anne-Sophie Dubarry, François Tadel, Pietro Avanzini, Nikolai Axmacher, Dillan Cellier, Maria Del Vecchio, Liberty S. Hamilton, Dora Hermes, Michael J. Kahana, Robert T. Knight, Anais Llorens, Pierre Megevand, Lucia Melloni, Kai J. Miller, Vitória Piai, Aina Puce, Nick F Ramsey, Caspar M. Schwiedrzik, Sydney E. Smith, Arjen Stolk, Nicole C. Swann, Mariska J Vansteensel, Bradley Voytek, Liang Wang, Jean-Philippe Lachaux, and Robert Oostenveld. Advances in human intracranial electroencephalography research, guidelines and good practices. *NeuroImage*, 260:119438, October 2022.

[7] Shailee Jain and Alexander Huth. Incorporating Context into Language Encoding Models for fMRI. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6628–6637. Curran Associates, Inc., 2018.

[8] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, February 2022.

[9] Aditya R. Vaidya, Shailee Jain, and Alexander Huth. Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech. In *Proceedings of the 39th International Conference on Machine Learning*, pages 21927–21944. PMLR, June 2022.

[10] Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning, June 2022.

[11] Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):e3002366, December 2023.

[12] Alexander J. E. Kell, Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018.

[13] Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8528, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[14] Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving Semantic Understanding in Speech Language Models via Brain-tuning. In *The Thirteenth International Conference on Learning Representations*, October 2024.

[15] Anuja Negi, Subba Reddy Oota, Anwar O. Nunez-Elizalde, Manish Gupta, and Fatma Deniz. Brain-Informed Fine-Tuning for Improved Multilingual Understanding in Language Models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, October 2025.

[16] Nishitha Vattikonda, Aditya R. Vaidya, Richard J. Antonello, and Alexander G. Huth. Brain-WavLM: Fine-tuning Speech Representations with Brain Responses to Language, February 2025.

[17] Omer Moussa and Mariya Toneva. Brain-tuning Improves Generalizability and Efficiency of Brain Alignment in Speech Models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, October 2025.

[18] Tobias Overath, Josh H. McDermott, Jean Mary Zarate, and David Poeppel. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, 18(6):903–911, June 2015.

[19] Marc Schönwiesner and Robert J. Zatorre. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences*, 106(34):14611–14616, August 2009.

[20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv:2106.07447 [cs, eess]*, June 2021.

[21] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *arXiv:2110.13900 [cs, eess]*, October 2021.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

[23] Yuanning Li, Gopala K. Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H. Carney, Junfeng Lu, Jinsong Wu, and Edward F. Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, December 2023.

[24] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, August 2023.

[25] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011.

[26] Patrick W. Hullett, Liberty S. Hamilton, Nima Mesgarani, Christoph E. Schreiner, and Edward F. Chang. Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *Journal of Neuroscience*, 36(6):2014–2026, February 2016.

[27] Liberty S. Hamilton, Erik Edwards, and Edward F. Chang. A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, 28(12):1860–1871.e4, June 2018.

[28] Ariel Goldstein, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel A. Nastase, Harshvardhan Gazula, Aditi Singh, Aditi Rao, Gina Choe, Catherine Kim, Werner Doyle, Daniel Friedman, Sasha Devore, Patricia Dugan, Avinatan Hassidim, Michael Brenner, Yossi Matias, Orrin Devinsky, Adeen Flinker, and Uri Hasson. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, 9(5):1041–1055, May 2025.

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

[30] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems*, 36:21895–21907, December 2023.

[31] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021.

[32] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

[33] Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The Algonauts Project 2025 Challenge: How the Human Brain Makes Sense of Multimodal Movies, January 2025.

[34] Julie Boyle, Basile Pinsard, Valentina Borghesani, Francois Paugam, Elizabeth DuPre, and Pierre Bellec. The Courtois NeuroMod project: Quality assessment of the initial data release (2020). In *2023 Conference on Cognitive Computational Neuroscience*, Oxford, UK, 2023. Cognitive Computational Neuroscience.

[35] Roy Mukamel, Hagar Gelbard, Amos Arieli, Uri Hasson, Itzhak Fried, and Rafael Malach. Coupling Between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. *Science*, 309(5736):951–954, August 2005.

[36] Jeremy R. Manning, Joshua Jacobs, Itzhak Fried, and Michael J. Kahana. Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. *Journal of Neuroscience*, 29(43):13613–13620, October 2009.

[37] Anne Hsu, Alexander Borst, and Frédéric E Theunissen. Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, 15(2):91–109, January 2004.

[38] Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010.

[39] Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, Caitlyn Hoeflin, Alvincé Pongos, Idan A. Blank, Melissa Kline Struhl, Anna Ivanova, Steven Shannon, Aalok Sathe, Malte Hoffmann, Alfonso Nieto-Castañón, and Evelina Fedorenko. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, 9(1):529, August 2022.

[40] Sam V. Norman-Haignere, Laura K. Long, Orrin Devinsky, Werner Doyle, Ifeoma Irobunda, Edward M. Merricks, Neil A. Feldstein, Guy M. McKhann, Catherine A. Schevon, Adeen Flinker, and Nima Mesgarani. Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, pages 1–15, February 2022.

[41] Takuya Matsuyama, Kota S Sasaki, and Shinji Nishimoto. Applicability of scaling laws to vision encoding models, August 2023.

[42] Abdulkadir Gokce and Martin Schrimpf. Scaling Laws for Task-Optimized Models of the Primate Visual Ventral Stream. In *Forty-Second International Conference on Machine Learning*, June 2025.

[43] Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, July 2001.

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023.

A fMRI-tuning improves ECoG performance

A.1 ROI definitions

We used the following using labels from the Destrieux [38] atlas to represent ROIs in Figure 2B:

- Prefrontal cortex (PFC): `G_front_middle`, `G_front_sup`, `S_front_middle`, `S_front_sup`
- Primary auditory cortex (AC): `G_temporal_middle`, `G_temporal_sup`, `S_temporal_transverse`
- Language network (Lipkin et al. (2022) [39]): `Pole_temporal`, `G_temp_sup-Lateral`, `G_temporal_middle`, `G_temp_sup-Plan_polar`, `G_front_inf-Orbital`, `G_front_inf-Triangul`, `Lat_Fis-ant-Horizont`, `S_temporal_transverse`, `G_front_inf-Opercular`, `G_temp_sup-Plan_tempo`, `Lat_Fis-ant-Vertical`, `G_pariet_inf-Supramar`, `S_interm_prim-Jensen`, `S_temporal_sup`, `S_orbital_lateral`, `S_precentral-inf-part`, `G_precentral`, `S_central`

A.2 Subject-level effects in ECoG performance after fMRI-tuning

We analyzed subject-level effects in ECoG by fitting a linear mixed-effects model with the WavLM model type (pretrained or fMRI-tuned) as a fixed effect and subject-specific random intercepts. fMRI-tuned models had significantly higher performance than the pretrained models with $\beta = 0.0117$, SE = 0.0043, 95% CI [0.0034, 0.0201], $p = 0.006$.

In Table 1, we quantify the mean performance (across electrodes) for each ECoG subject, before and after fMRI-tuning. We also compute paired two-sided t-tests across electrodes within each subject. The fMRI-tuned models were significantly better ($p < 0.005$) than the pretrained model for 7 of 9 ECoG patients.

Table 1: Per-subject ECoG performance and significance.

Patient	fMRI-tuned (ρ)	Pretrained (ρ)	p-value
01	**0.096**	0.0781	9.94×10^{-10}
02	**0.0603**	0.0513	1.61×10^{-7}
03	**0.0902**	0.0741	3.42×10^{-19}
04	**0.0839**	0.0707	6.5×10^{-7}
05	0.026	**0.0296**	0.00342
06	**0.104**	0.0885	3×10^{-13}
07	0.0702	**0.0757**	0.061
08	**0.0976**	0.0803	1.28×10^{-6}
09	**0.0958**	0.0764	2.64×10^{-22}

In Figure 6, we visualize the per-electrode effect of fMRI-tuning separately for each subject in the "Podcast" dataset.

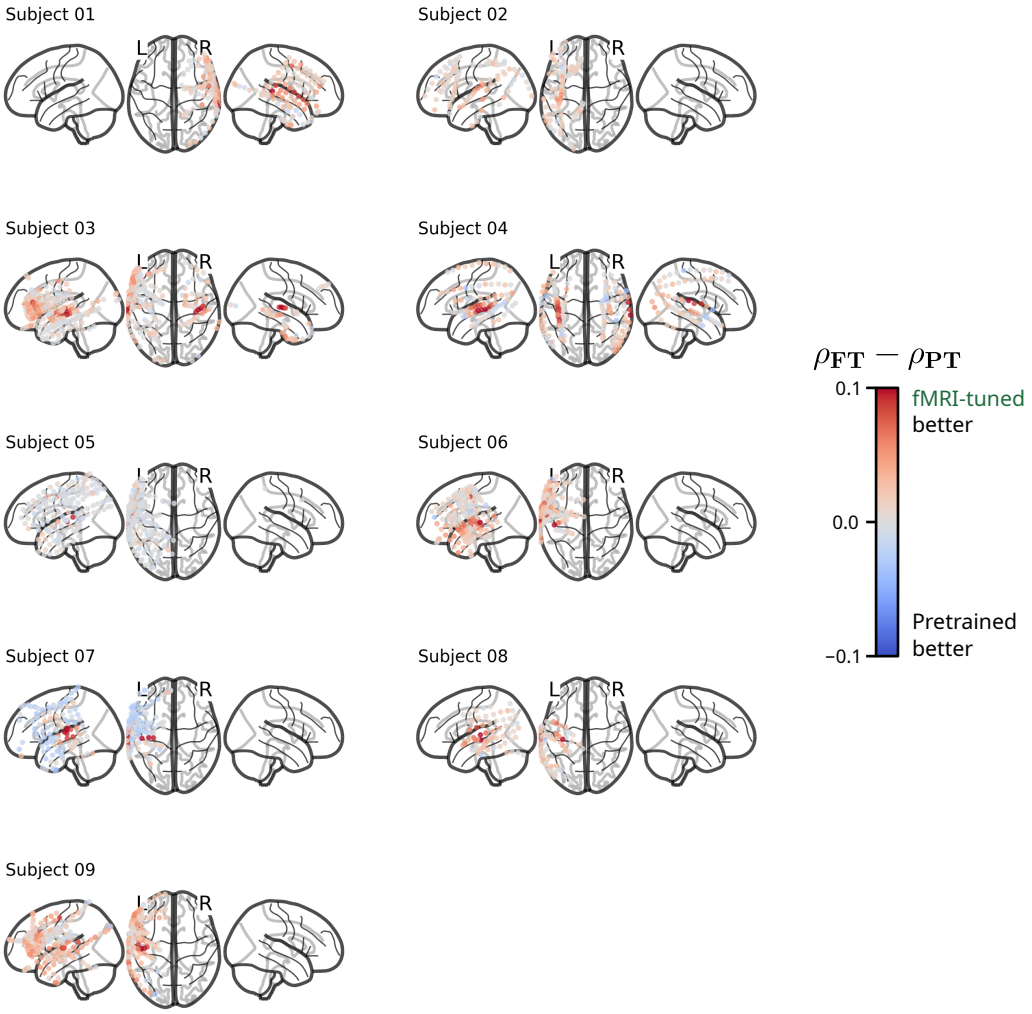


Figure 6: Improvement in encoding performance for all electrodes (Figure 2C), visualized separately for each subject.

A.3 FIR ECoG encoding models

We fit 40-delay FIR models ($t = 50$ ms to $t = 2$ s) on WavLM features after using PCA to reduce them to 100 dimensions. We show their performance in Table 2. With an FIR model, fMRI-tuning outperforms the pretrained model by a larger margin than for the single-delay models. Note that these values cannot be directly compared to those from single-delay models, since here we do not take the maximum performance across delays.

Table 2: FIR ECoG encoding models with a reduced feature space.

Model	Mean encoding perf. (ρ)	SEM	% improvement vs. pretrained
fMRI-tuned	0.05772	0.002968	36.59
Pretrained	0.04226	0.002352	–

A.4 fMRI-tuned models vs. stimulus baselines

To control for simpler speech features, we evaluated whether concatenating the WavLM features with the acoustic envelope, spectrogram, or phonetic features improved ECoG performance. We used FIR models because the temporal response of the WavLM and baseline features may be different. Results are shown in Table 3.

The fMRI-tuned models outperform all three baselines individually, and importantly, they explain an orthogonal component of the response relative to the baselines. For example, concatenating the fMRI-tuned representations with the speech envelope improved performance by 0.01861, indicating that the fMRI-tuning improved the models in non-acoustic ways. Because gains are also seen over the spectral and phonetic features, the fMRI-tuned models are better in ways not explained by any of these baselines.

Table 3: Variance partitioning analysis of fMRI-tuned models and known speech baselines.

Model	WavLM only	with envelope	with spectrogram	with phonetic
Baseline only	N/A	0.04034 ± 0.002177	0.04591 ± 0.002427	0.03722 ± 0.002
Pretrained WavLM	0.04226 ± 0.002352	0.05334 ± 0.002874	0.06289 ± 0.003025	0.05205 ± 0.002815
fMRI-tuned WavLM	0.05772 ± 0.002968	0.05895 ± 0.003048	0.0666 ± 0.003136	0.05808 ± 0.002987

We fine-tuned additional WavLM models on stimulus-only features: LLaMA [44] hidden states (semantics), word rate (acoustic), and articulatory features (phonetic). We also fine-tuned models on fMRI responses shuffled in 10-TR blocks to show that the temporal structure in the fMRI responses is necessary. We report the results in Table 4. The shuffled-fMRI-tuned models show a small improvement over the pretrained model, but are substantially worse than the normal fMRI-tuned models. The three stimulus-tuned models do perform well, but, when factoring in the results from Table 3, learning on these noiseless signals likely improves models along different axes than what fMRI-tuning does.

Table 4: fMRI-tuning compared against more baselines for single-delay models. Performance is averaged across all electrodes.

Model	Mean encoding perf. (ρ)	SEM
fMRI-tuned	0.08107	0.003183
shuffled-fMRI-tuned	0.07233	0.002976
LLaMA-tuned	0.08822	0.003248
Word rate-tuned	0.08739	0.003228
Articulation-tuned	0.08665	0.003185
Pretrained	0.06955	0.002559

B fMRI-tuning on downsampled fMRI responses

Figure 7 compares the fMRI encoding performance of models fine-tuned on the original fMRI responses and the downsampled responses for additional subjects.

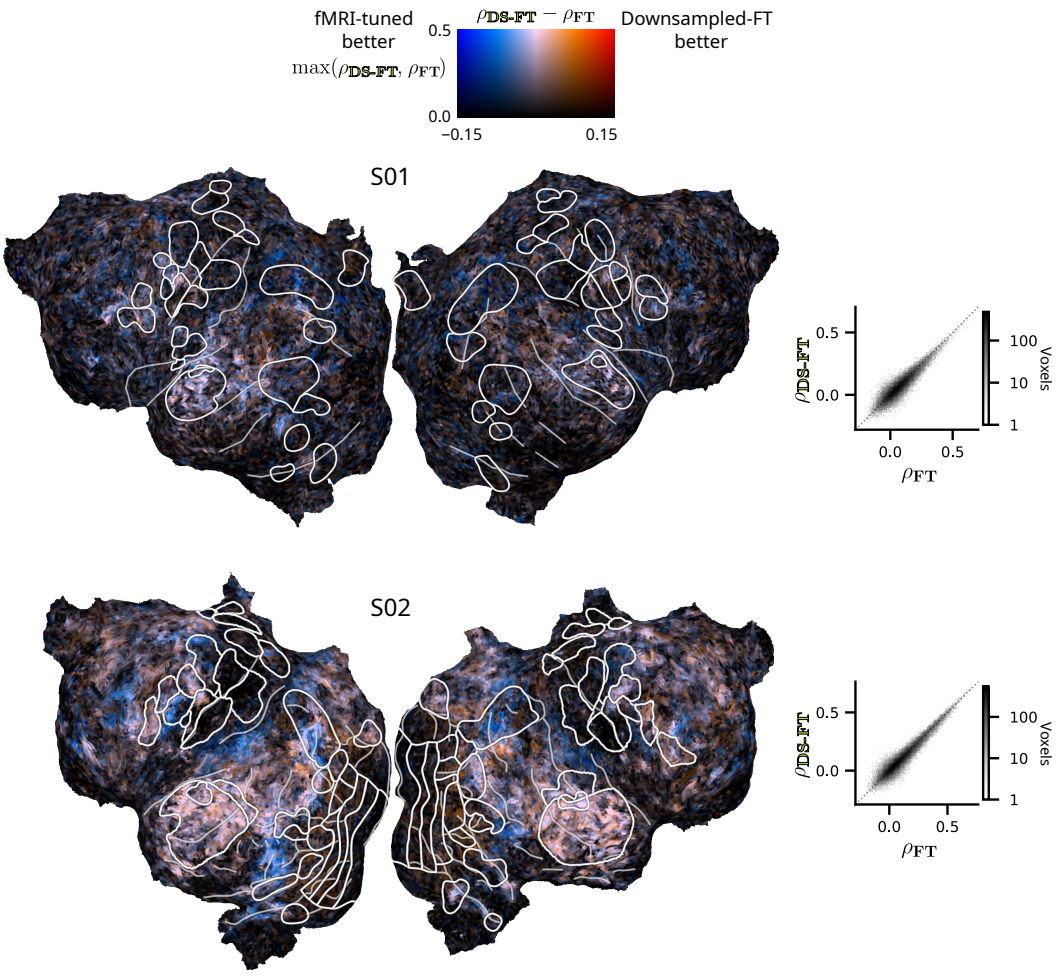


Figure 7: Flattened cortical surfaces of subjects S1 and S2 from LeBel et al. (2023) [24] that compare the encoding performance of the original fMRI-tuned models against the downsampled-fMRI-tuned models. On the right, 2-dimensional histograms show that individual voxel performance is stable across fine-tuning conditions. (See Figure 4A for S3.)

C Scaling of fMRI-tuning

C.1 Scaling of fMRI-tuning for fMRI prediction

Using the models fMRI-tuned in Section 3.4, we show how within-modality prediction scales with fine-tuning dataset size in Figure 8. Fine-tuned models improve at a faster rate than the pre-trained model, and they show gains even at small dataset sizes.

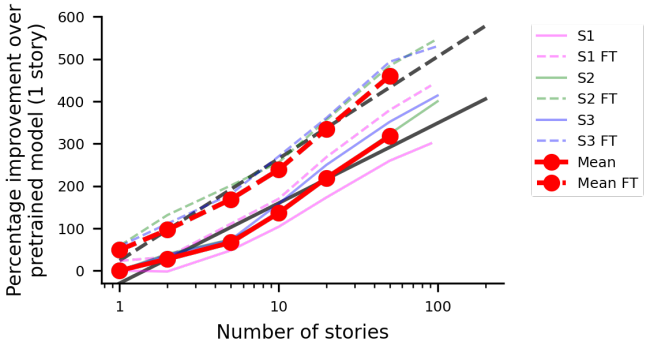


Figure 8: Within-subject fMRI encoding performance scales with the size of the fine-tuning dataset. Pink, green, and blue show scaling for each fMRI subject, and red shows the mean of the three. Solid lines show the scaling performance of the pre-trained model, while dashed lines show scaling of fine-tuned models. Black lines show a best linear fit to the mean across subjects. Fine-tuned models outperform the pretrained model at all dataset sizes and also improve faster.

C.2 Scaling of fMRI-tuning for ECoG prediction

In Figure 9, we extend Figure 5A by showing fMRI-tuning performance using the full fMRI training set.

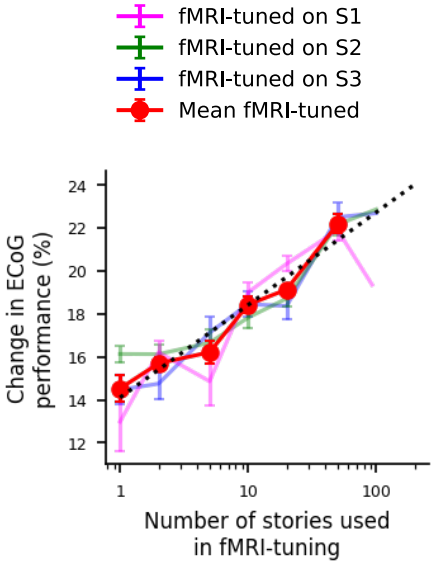


Figure 9: ECoG encoding performance as a function of the number of fMRI fine-tuning stories, including the full training set (93 stories for S1, and 100 stories for S2 and S3). Error bars indicate the standard error over bootstraps of the fMRI stories.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present empirical evidence in Section 3 to support all claims in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not make any theoretical claims.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have thoroughly documented our methods and evaluations in Section 2, and we document the procedure for further analyses in each sub-section of Section 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets (fMRI stories [24], CNeuroMod [33], and ECoG "Podcast" [29]) and models (WavLM [21]) used in this paper are already publicly accessible, including code for fitting linear models within the two brain datasets. Code used for this paper is available at: `https://github.com/HuthLab/fmri-ecog-transfer`

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We discuss training and evaluation data in Section 2 and general model fitting procedure in Section 2.2. We discuss our train-val-test split and all other fine-tuning details in Section 2.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars and statistical significance, as well as how they were computed, in Figures 2, 3, and 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss compute requirements in Section 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and believe this work complies with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We describe the societal impacts of brain encoding models in Sections 1 and 4, particularly as they apply to decoding and brain-computer interfaces (BCIs).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not present artifacts that have a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Section 2, we cite the papers that describe the public brain datasets we use, as well as the paper for the pretrained model that we use for fMRI-tuning.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have released the fine-tuning and evaluation code needed to reproduce our results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We use previously published brain data recorded from human subjects. Refer to the original papers for details: LeBel et al. (2023) [24], Tang et al. (2023) [3], and Zada et al. (2025) [29].

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper uses previously published data from human subjects. For details, refer to the original papers: LeBel et al. (2023) [24], Tang et al. (2023) [3], and Zada et al. (2025) [29].

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We fine-tune speech-audio representation models, which could be considered LLMs. We describe how they are used primarily in Section 2.3. We do not use any other LLMs or agents.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.