
Supplementary Materials: Brain-tuning Improves Generalizability and Efficiency of Brain Alignment in Speech Models

Contents

A	Additional Methodology Details	2
A.1	Brain ROI Details	2
A.2	Noise Ceiling Calculation	2
A.3	Loss Functions and Training Details	2
A.4	Downstream tasks	3
B	Additional Results	4
B.1	Extended LoRA Rank Ablations	4
B.2	Additional Brain Alignment Plots	4
B.3	Brain Alignment Generalization of LLM-tuning and Stimulus-tuning	4
C	HuBERT Results	5
C.1	Generalization Results	5
C.2	Downstream Results	6

1 A Additional Methodology Details

2 A.1 Brain ROI Details

3 Glasser Atlas for human cerebral cortex parcellation has 180 labeled ROIs per hemisphere [Glasser
4 et al., 2016]. From these labels, we extract the following regions to be used during brain-tuning:
5 Angular gyrus, lateral temporal cortex, inferior frontal gyrus, and middle frontal gyrus [Oota et al.,
6 2024, Desai et al., 2023]. It also has the primary auditory and the early auditory regions. Fig. 5
7 highlights the ROIs used for brain-tuning on the right hemisphere. Table 1 details each region and the
8 ROI labels that cover it from the parcellation atlas.

Table 1: Brain regions and corresponding ROI labels.

Region	Labels
Angular gyrus (AG)	PFm, PGs, PGi, TPOJ2, TPOJ3
Lateral temporal cortex (LTC)	STSda, STSva, STGa, TE1a, TE2a, TGv, TGd, A5, STSdp, STSvp, PSL, STV, TPOJ1
Inferior frontal gyrus (IFG)	44, 45, IFJa, IFSp
Middle frontal gyrus (MFG)	55b
Primary auditory cortex (A1)	A1
Early auditory regions	A1, PBelt, MBelt, LBelt, RI, A4

9 A.2 Noise Ceiling Calculation

10 Noise in fMRI data is very common and can impair brain-tuning and brain alignment estimation, so
11 it is important to estimate the noise ceiling of each voxel in the fMRI recordings. The voxel-wise
12 noise ceiling is estimated for all participants’ fMRI data based on the preferred method by the
13 original dataset paper [LeBel et al., 2023]. This method leverages repetitions of the same story for
14 the participant (e.g., a story is repeated 10 times), then uses repetitions to compute the maximum
15 explainable variance for each voxel. This noise ceiling value estimates the amount of explainable
16 variance in the brain signal, ranging from 0 to 1. We use this estimated noise ceiling to normalize the
17 brain alignment during brain alignment estimation, as mentioned in Section 3.4.1 of the main paper.

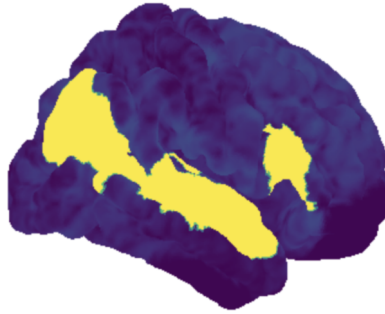


Figure 5: Brain-tuning ROIs. Yellow-highlighted regions are used for brain-tuning (Section 3.3.2).

18 A.3 Loss Functions and Training Details

19 Here, we detail the formulations of loss functions compared in Sections 3.3 and 4.4 of the main paper.
20 We then compare different training techniques for Multi-Brain-tuning.

21 To define the loss functions, assume that we have a batch B of audio-fMRI pairs for Participant i ,
22 where the ground-truth fMRI responses are R^i and the predicted fMRI responses are \hat{R}^i .

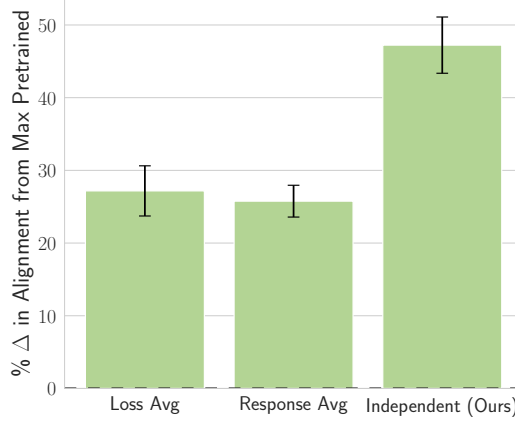


Figure 6: Multi-Brain-tuning for independent predictions vs loss and response average. 3.3.3).

23 **L_2 Loss.** We compute the L_2 loss between R^i and \hat{R}^i as follows:

$$\mathcal{L}_{l2} = \frac{1}{|B|} \sum_{b=0}^{|B|-1} (R_b^i - \hat{R}_b^i)^2 \quad (1)$$

24 **Correlation Loss.** We compute the correlation loss between R^i and \hat{R}^i as follows (where corr is the
25 correlation over voxels):

$$\mathcal{L}_{\text{corr}} = \frac{1}{|B|} \sum_{b=0}^{|B|-1} (1 - \text{corr}(R_b^i, \hat{R}_b^i)) \quad (2)$$

26 **Cosine + L_2 Loss.** We compute the Cosine + L_2 loss between R^i and \hat{R}^i as follows (where cos is
27 the cosine similarity over voxels):

$$\mathcal{L}_{\text{cos}} = \frac{1}{|B|} \sum_{b=0}^{|B|-1} (1 - \cos(R_b^i, \hat{R}_b^i)) \quad (3)$$

28 Then we use it alongside the L_2 Loss (with $\lambda = 0.5$) as follows:

$$\mathcal{L}_{\text{cos-l2}} = \mathcal{L}_{\text{cos}} + \lambda \mathcal{L}_{l2} \quad (4)$$

29 **Comparing Multi-Brain-tuning techniques.** Fig. 6 shows the change in performance from
30 pretrained Wav2Vec2.0 for 3 different methods of incorporating responses from multiple participants
31 (detailed in Section 3.3.3). In the Loss Avg method, we compute the loss for each participant, then
32 average it across participants before updating the parameters. On the other hand, for the Response
33 Avg method, we average responses over participants then compute the loss and update the parameters.
34 Finally, the method we use in the paper is predicting each participant response independently (but
35 using the same projection head), compute the loss, and update the model parameters. This is found to
36 work better than these alternatives as it allows the models to learn more robustly from information
37 across participants.

38 A.4 Downstream tasks

39 We detail here the datasets and the formulation of the downstream tasks mentioned in Section 3.4.4.

40 **Phonemes Prediction.** Phoneme recognition is done as a multi-label classification problem, following
41 the work of [Moussa et al., 2025]. A linear classifier projects the layer representation to a set of 39
42 possible phonemes that occurred in the original input audio segment. We use the TIMIT dataset
43 Garofolo [1993] because of its phonetically rich audio snippets. The final performance measure is
44 the classifier’s F1-score on the held-out test set. We report the F1-score averaged over the upper
45 middle-layers as done for brain alignment in Section 3.4.1.

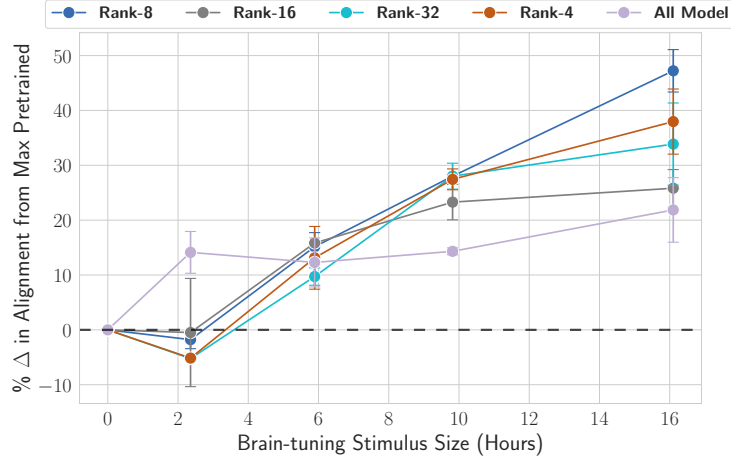


Figure 7: Effect of the number of trainable parameters for Wav2Vec2.0 (Extending Section 4.4 of the main paper). Increasing the number of trainable parameters with a higher rank (e.g., 32) or by fine-tuning the entire model doesn’t lead better scaling than rank-8 updates.

46 **Phonetic Sentence Type Prediction.** Predicting the phonetic sentence type can be used to evaluate
 47 a model’s phonetic understanding beyond single phonemes and words. The TIMIT dataset Garo-
 48 folo [1993] has three phonetic types for each utterance: SA (for utterances that cover all English
 49 phonemes), SX (for phonetically balanced utterances that have extensive phone coverage but few
 50 words), and SI (for natural and phonetically diverse utterances). Each of the three types (SA, SX,
 51 SI) highlights specific speech dialectal, phonetic, or coverage aspects. To evaluate performance on
 52 this task, we follow [Moussa et al., 2025] to predict the phonetic sentence type. We add a projection
 53 classification head to predict the sentence type from the given layer’s representation. The performance
 54 is measured by the F1-score on the held-out test set, averaged across the upper-middle layers.

55 B Additional Results

56 B.1 Extended LoRA Rank Ablations

57 Fig. 7 extends Fig. 4 by adding Rank-32 updates and all model updates (fine-tuning all transformer
 58 parameters). It supports our finding that we don’t need more than rank-8 updates for our method
 59 to work well. Moreover, it shows that updating the entire model scales more slowly than LoRA,
 60 indicating that it needs more data to reach the same performance.

61 B.2 Additional Brain Alignment Plots

62 Here, we visualize the impact of brain-tuning on brain alignment for the remaining heldout partici-
 63 pants. Fig. 11 extends Fig. 2B) by showing the remaining 4 participants. Similarly to Fig. 2B, we
 64 observe a widespread improvement across the brain for these participants, especially the frontal and
 65 parietal regions, while the auditory cortex shows a slight decrease in alignment. We attribute this
 66 decrease in auditory cortex alignment to the fact that we report alignment over the upper-middle and
 67 later layers of the models, which are known to be more semantic (refer to Section 4.2 in the main
 68 paper for more details).

69 B.3 Brain Alignment Generalization of LLM-tuning and Stimulus-tuning

70 In this section, we elaborate on the training details of LLM-tuning and Stimulus-tuning, then report
 71 on how they scale with more tuning data.

72 **LLM-tuning.** For tuning, we use representations from layers 18 to 24 of the Llama2-7B Model
 73 [Touvron et al., 2023] instead of brain signals. These layers are used because they show the best
 74 alignment with late language regions. We then apply a similar fine-tuning pipeline to that of brain-

tuning (detailed in Section 3.3.4 of the main paper). We use LoRA rank-8 updates, a learning rate of 10^{-4} with linear decay, and a batch size of 128 samples of (audio, fMRI response) pairs. For LLM-tuning, we found that it takes longer to converge than brain-tuning; we train them for 250 epochs, which takes around 10h on two NVIDIA A40 48GB GPUs.

Stimulus-tuning. This baseline aims to test the benefits of brain-tuning against simply fine-tuning using stimulus audio. This highlights any improvements in the model that would be solely due to seeing more data. We follow the training setting in [Moussa et al., 2025] for stimulus-tuning. The same pretraining losses (the diversity loss and the contrastive loss) with the same hyperparameters of [Baevski et al., 2020] are used. The model is then fine-tuned for 300 epochs using a base learning rate of 2×10^{-5} with a warm-up for the first 10% of the updates followed by a linear decay schedule. It takes around the same amount of time to train as LLM-tuning.

Next, we test whether scaling the data for LLM-tuning and Stimulus-tuning leads to improved alignment, as we observe with brain-tuning (refer to Section 4.2). Fig. 8 shows the change in brain-alignment for brain-tuning as well as LLM-tuning and Stimulus-tuning on the training participants. LLM-tuning improves alignment with more data, but it shows a saturation trend and is always lower than brain-tuned models. Stimulus-tuning models doesn’t show improvement over the pretrained counterpart, indicating that seeing more audio data is not the cause for the improved alignment.

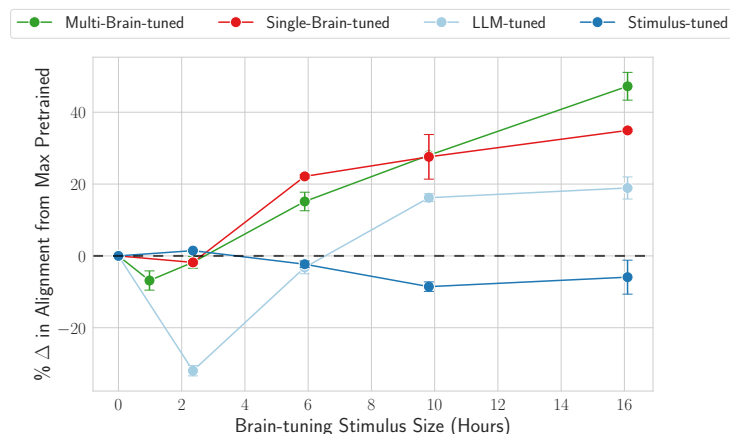


Figure 8: Brain Alignment scaling with LLM-tuning and Stimulus-tuning of Wav2Vec2.0. This figure extends Section 4.2 of the main paper by reporting the improvement in alignment with more data for LLM-tuning and Stimulus-tuning. It shows that LLM-tuning can help improve brain alignment with more data, but it tends to saturate and is always worse than brain-tuning. Stimulus-tuning doesn’t seem to improve alignment and always performs comparably to the pretrained model.

C HuBERT Results

C.1 Generalization Results

We repeat here the same analysis in Section 4.2 but for HuBERT model family. We test how the brain-tuned models generalize against increasing amounts of data during brain-tuning. This is done by measuring brain alignment improvement (relative to pretrained HuBERT) when we scale the data used for Multi- and Single- Brain-tuning. Fig. 9 shows a similar trend to Fig. 2 on both training and heldout participants. When we increase the data used for brain-tuning, Multi-Brain-tuned models tend to perform better than Single-Brain-tuned ones. As for lower data fractions, the improvement of both was comparable. These results (along with their Wav2Vec2.0 parallels in Section 4.2 of the main paper) further confirm the scalability of multi-brain-tuning in improving alignment with new unseen participants. The upward trend also indicates the potential for further improvement if more data is integrated for brain-tuning.

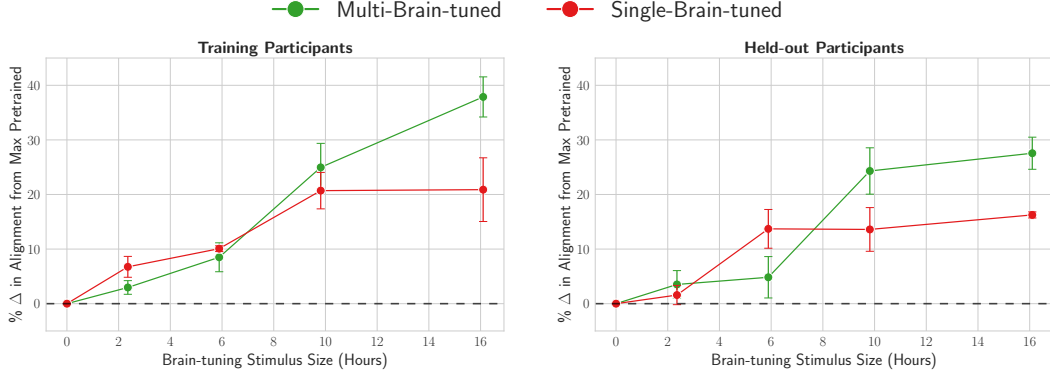


Figure 9: Impact of scaling the tuning data of brain-tuning on brain alignment for HuBERT. Similar to brain-tuned models of Wav2Vec2.0 (Section 4.2 of the main paper), when the data used for brain-tuning scales up, Multi-Brain-tuned models tend to perform better than Single-Brain-tuned ones, on both training and heldout participants.

104 C.2 Downstream Results

105 We report here the downstream performance of HuBERT on the same tasks detailed in Section 4.3
 106 and Supp. A.4. Fig. 10 shows similar findings to Fig. 3 of the main paper. When the amount of tuning
 107 data increases, brain-tuned models eventually reach the same level of performance as the LLM-tuned
 108 one (which was shown to substantially improve performance on similar tasks by [Moussa et al., 2025,
 109 Vattikonda et al., 2025]). Moreover, for all data sizes, brain-tuned models never perform worse than
 110 their pretrained counterparts. These results (alongside their Wav2Vec2.0 equivalents in Section 4.3 of
 111 the main paper) further confirm that our brain-tuning approach doesn't lead to catastrophic forgetting;
 112 to the contrary, it leads to a strong improvement in downstream performance.

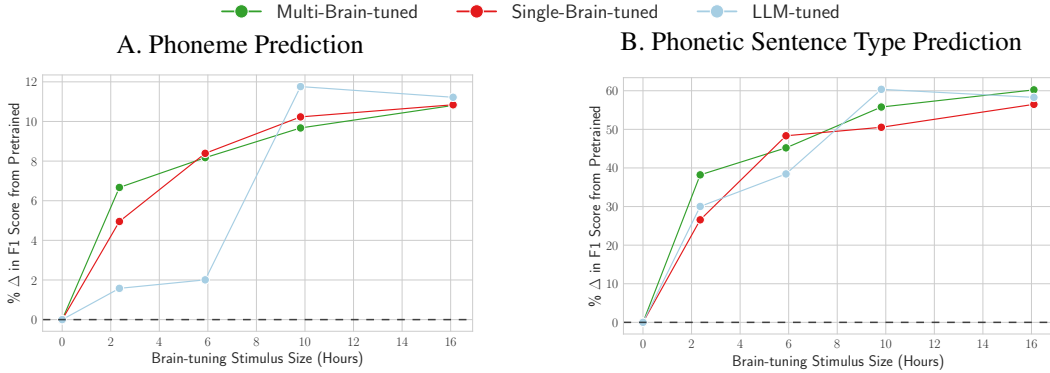
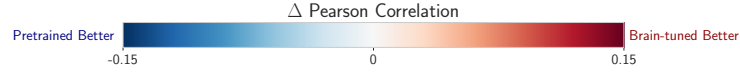
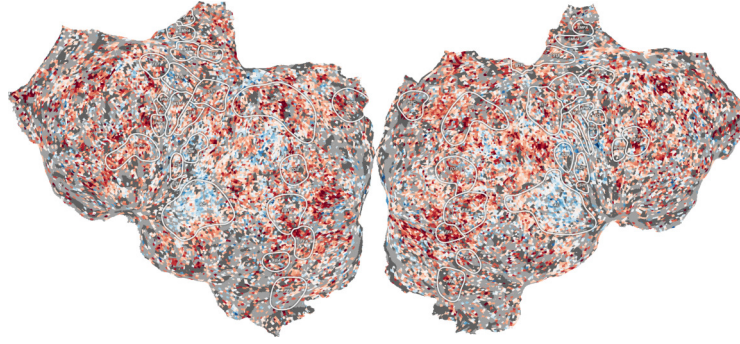


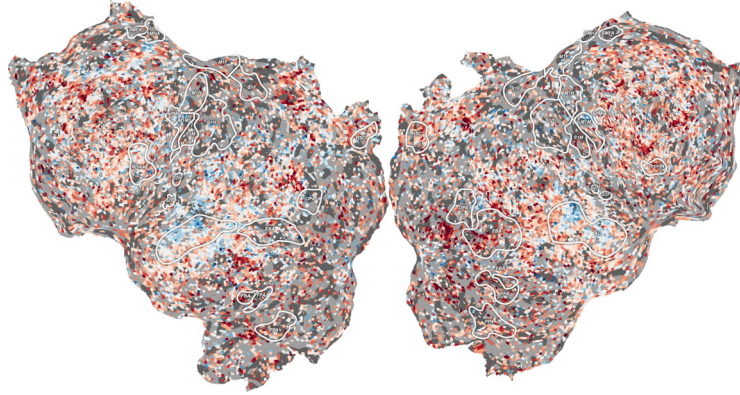
Figure 10: Scaling of downstream performance with tuning data size for HuBERT. Brain-tuned models' performance increases with more data, eventually matching that of the LLM-tuned model.



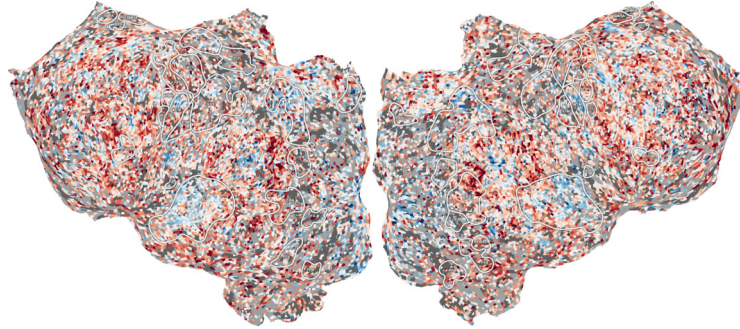
Participant 4



Participant 5



Participant 7



Participant 8

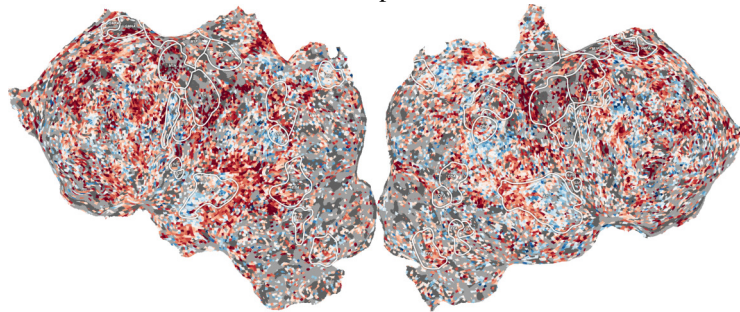


Figure 11: Impact of Multi-Brain-tuning on brain alignment for heldout participants. The figure shows the change in brain alignment (measured by Pearson Correlation) after Multi-Brain-tuning, compared to the pretrained Wav2Vec2.0 model. It shows a widespread improvement across the brain for these participants, especially the frontal and parietal regions.

References

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- Rutvik H Desai, Usha Tadimet, and Nicholas Riccardi. Proper and common names in the semantic system. *Brain Structure and Function*, 228(1):239–254, 2023.
- John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium*, 1993, 1993.
- Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.
- 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, 2023.
- 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*, 2025. URL <https://openreview.net/forum?id=KL8Sm4xRn7>.
- Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. *ACL*, 2024.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Nishitha Vattikonda, Aditya R. Vaidya, Richard J. Antonello, and Alexander G. Huth. Brainwavlm: Fine-tuning speech representations with brain responses to language, 2025. URL <https://arxiv.org/abs/2502.08866>.