Temporal Precision Matters: Brain-Tuning Speech Language Models with Millisecond-Resolution Neural Signals

Zhejun Zhang[1], **Wenqing Zhou**[1], **Haozhe Xu**[1], **Lin Zhang**[2,1], **and Lei Li**[1*]
[1]Beijing University of Posts and Telecommunications, China
[2]Beijing Big Data Center, China
{zhejun.zhang, zwq211, xuhaozhe2022, zhanglin, leili}@bupt.edu.cn

Abstract

Brain-tuning enhances brain alignment and downstream performance by fine-tuning speech language models with neural recordings. However, previous work relies primarily on fMRI, whose temporal resolution integrates neural activity over seconds, blending distinct processing stages into a single supervision signal and precluding temporally targeted training. We introduce ECoG-tuning, which leverages electrocorticography's millisecond precision to train speech language models. We design temporally targeted windows—a speech window capturing acoustic-phonetic encoding and a language window capturing higher-order linguistic processing—grounded in neuroscientific findings about temporal encoding hierarchies. Evaluating three models on the Podcast ECoG dataset, we find that ECoG-tuning significantly improves brain alignment over pretrained and distillation baselines. Notably, full spatiotemporal dynamics yield 7–17% higher alignment than time-averaged supervision across models, and language-window tuning produces larger gains in higher-order language regions, indicating that temporal precision provides additional training value. Moreover, ECoG-tuned models consistently improve or maintain downstream performance. Overall, our work provides initial evidence that electrophysiology is a viable brain-tuning modality, demonstrating how neuroscientific insights into processing hierarchies can inform principled model training strategies. Code is available at https://github.com/Mochizuki-BUPT/ECoG-Tuning-main.

1 Introduction

Language models have emerged as powerful tools for predicting human brain activity during language comprehension, revealing notable alignment between artificial and biological language processing (Wehbe et al., 2014; Jain and Huth, 2018; Toneva and Wehbe, 2019; Schrimpf et al., 2018, 2021; Goldstein et al., 2022, 2024; Caucheteux and King, 2022; Nakagi et al., 2024; AlKhamissi et al., 2025b; Oota et al., 2025; Kriegeskorte et al., 2008). Recent work on brain-tuning—fine-tuning speech language models (SLMs) using human brain recordings—has demonstrated that neural supervision enhances models' semantic understanding and downstream performance (Moussa et al., 2025; Moussa and Toneva, 2025; Negi et al., 2025).

However, existing approaches have relied primarily on functional magnetic resonance imaging (fMRI). While fMRI provides broad spatial coverage, its temporal resolution is limited: each sample integrates neural activity over ∼2 s, conflating distinct processing stages into an undifferentiated supervision signal and precluding training strategies that target specific cognitive processes. Electrocorticography (ECoG) offers millisecond precision and signal-to-noise ratios (SNR) substantially exceeding non-invasive methods, enabling investigation of rapid neural dynamics. Critically, ECoG research has shown that language comprehension unfolds along a temporal hierarchy (Goldstein et al., 2025a). Goldstein et al. (2025b) demonstrated that speech encoding peaks at ∼54 ms post-word-onset while language encoding peaks at ∼247 ms—a ∼200 ms separation invisible to fMRI but resolvable with ECoG. This raises a natural question: can brain-tuning benefit from this temporal precision?

In this work, we introduce ECoG-tuning, leveraging intracranial electrophysiology for SLM training. ECoG's millisecond resolution enables two methodological advances: (1) word-level supervision yielding several-fold more training samples, and (2) temporally targeted training that supervises models on neural responses from distinct processing stages. We design two windows anchored to word onset: a speech window targeting acoustic-phonetic encoding, and a language window targeting higher-order linguistic processing. This design is motivated by the temporal separation observed

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41208–41226

in ECoG encoding studies.

We evaluate ECoG-tuning on three pretrained SLMs (Wav2Vec 2.0, HuBERT, and Whisper) using the Podcast ECoG dataset (Zada et al., 2025). Our main contributions are:

1. **Temporally targeted neural supervision.** We introduce a training strategy that supervises models on neural responses from distinct processing stages, translating neuroscientific findings into actionable methodology.

2. **Empirical validation of temporal precision.** We find that temporal structure carries additional value: full temporal dynamics yield 7–17% higher alignment than time-averaged supervision, and language-window tuning produces larger gains in language-responsive regions compared to speech-window tuning.

3. **Preserved downstream utility.** ECoG-tuned models consistently improve or maintain performance on speech understanding tasks, indicating that temporally targeted neural supervision enhances brain alignment without compromising practical capabilities.

2 Related Work

2.1 Brain-Tuning and Neural Supervision

Fine-tuning pretrained models to predict human brain responses was pioneered by Schwartz et al. (2019) using BERT and fMRI data. Moussa et al. (2025) extended this paradigm to SLMs, showing that fMRI-based brain-tuning improves semantic understanding and downstream performance. Subsequent works expanded along multiple dimensions: multi-participant training (Moussa and Toneva, 2025), bilingual brain data (Negi et al., 2025), multimodal audio-video models (Policzer et al., 2025), and parameter-efficient fine-tuning (PEFT) via LoRA (Vattikonda et al., 2025). Related approaches improve model-brain alignment through alternative mechanisms. Cognitive feature injection methods incorporate eye-tracking or electroencephalography (EEG) features during pretraining (Ren and Xiong, 2021; Ding et al., 2022), while associative memory mechanisms enhance alignment without direct neural supervision (Yin et al., 2025b). Unlike these approaches, brain-tuning—and our ECoG-tuning—directly optimize neural encoding as the primary training objective.

Previous brain-tuning research has relied primarily on fMRI, which offers wide spatial coverage but has a temporal resolution of approximately 2 seconds. In contrast, ECoG provides millisecond precision, presenting a complementary method that allows for temporally targeted supervision. This approach enables training on neural responses from specific stages, which we introduce here as a novel strategy for brain-tuning.

2.2 Temporal Dynamics in Neural Language Processing

ECoG provides distinctive capabilities for studying rapid neural dynamics during language processing (Goldstein et al., 2022; Mischler et al., 2024; Bhattacharjee et al., 2026). Recent ECoG studies have revealed a temporal encoding hierarchy, with speech and language encoding peaking at distinct latencies separated by approximately 200 ms (Goldstein et al., 2025b). This temporal structure corresponds to hierarchical processing in language models, where earlier layers align with earlier neural responses and deeper layers with later responses (Millet et al., 2022; Mischler et al., 2024; Gwilliams et al., 2025; Raugel et al., 2025; He et al., 2025a).

The Podcast ECoG dataset (Zada et al., 2025) provides a public resource for such investigations, and prior work has shown that ECoG captures fine-grained temporal structure relevant to language model alignment, from shared response modeling (Bhattacharjee et al., 2026) to disentangled embeddings isolating distinct cognitive processes (He et al., 2025b). Our work leverages this temporal structure as a training signal, rather than for alignment evaluation alone.

3 Methodology

3.1 Speech Language Models

We evaluate three pretrained transformer-based SLMs: Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023). We use the Base versions of Wav2Vec 2.0 and HuBERT, and only the encoder of Whisper-small. The fine-tuned components of all three have comparable sizes (∼95–102M parameters), 12 transformer layers, and an embedding dimension of 768. Wav2Vec 2.0 and HuBERT are self-supervised models that employ a convolutional neural network (CNN) feature extractor to produce frame-level features at 20 ms intervals, while Whisper converts 30-second audio segments into log-mel spectrograms. Following prior work (Moussa et al., 2025), we freeze the CNN feature extractor

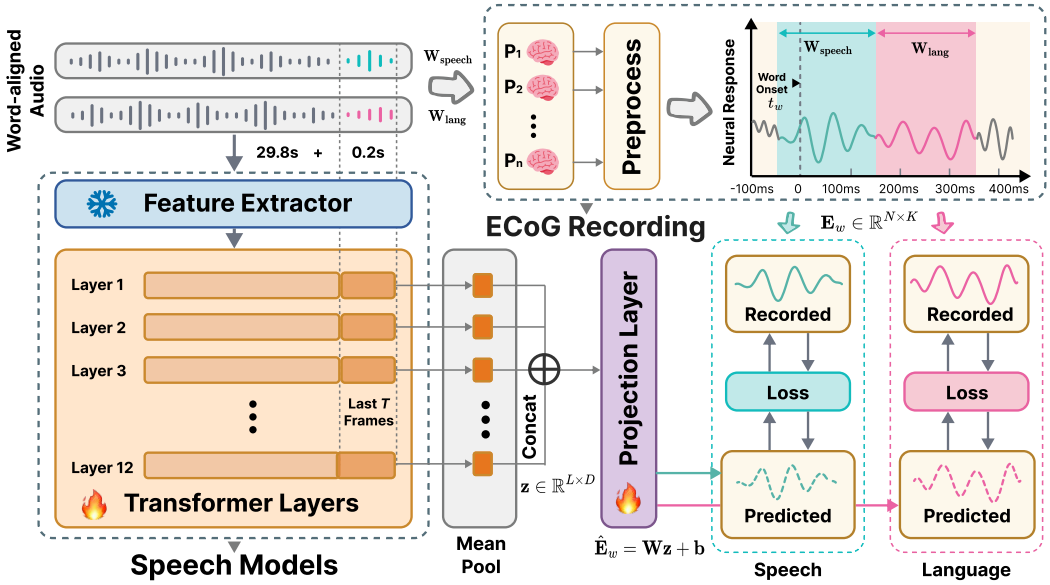

Figure 1: **Overview of the temporally targeted ECoG-tuning framework.** Audio context (30 s) ending 200 ms post-word-onset is processed through the speech encoder. The last 10 frames (200 ms) from each layer are pooled and concatenated, then projected to predict the spatiotemporal ECoG response.

for Wav2Vec 2.0 and HuBERT, and fine-tune only the transformer encoder. For Whisper, we fine-tune the encoder while keeping the decoder frozen.

3.2 ECoG Data

We use the Podcast dataset (Zada et al., 2025), a rare publicly available naturalistic speech comprehension resource that contains intracranial recordings from 9 participants listening to a 30-minute podcast (5,137 words with word-level timestamps). Electrode placement was determined by clinical needs, yielding variable spatial coverage across participants. After quality control, 1,268 electrodes were retained, comparable to those in prior intracranial studies (Mischler et al., 2024; He et al., 2025b).

ECoG's millisecond resolution enables word-level supervision: each word constitutes one training sample paired with its spatiotemporal neural response, yielding ∼5,000 samples per participant from 30 minutes of audio (Table 9). This contrasts with the fMRI-based approach, which operates at ∼2 s per sample. We extract high-gamma band power (70–200 Hz) as the neural target—the standard electrophysiological index of local population firing rates in human intracranial language research (Crone et al., 2006; Zada et al., 2025). Preprocessing details are in Appendix A.1. For region-specific analyses, we map electrodes to the Glasser parcellation (Glasser et al., 2016) and identify language-responsive regions (see Appendix A.2).

3.3 ECoG-Tuning Framework

3.3.1 Temporally Targeted Design

One contribution of this work is utilizing the temporal precision of ECoG to target distinct processing stages in the brain. The window design draws on findings from a neural encoding study: during speech comprehension, speech encoding in STG peaks at ∼54 ms post-word-onset, while language encoding in IFG peaks at ∼247 ms (Goldstein et al., 2025b). Anchored to the word onset t_w, we define two windows separated at 150 ms—approximately the midpoint of the two encoding peaks. This design aims to reduce, though not eliminate, the overlap between the two stages (Figure 1):

$$W(t_w, \tau) = \begin{cases} [t_w - 50\,\text{ms},\ t_w + 150\,\text{ms}] & \text{if } \tau = \text{speech} \\ [t_w + 150\,\text{ms},\ t_w + 350\,\text{ms}] & \text{if } \tau = \text{lang} \end{cases} \quad (1)$$

Here τ labels the neural processing stage rather than the model representation. We adopt a 200 ms duration for each word, as the same study systematically compared fixed and adaptive window lengths in their encoding models and found that 200 ms windows achieve encoding performance comparable to variable-length alternatives. While their finding pertains to the encoding direction (embedding → neural signal), we validate this choice for our fine-tuning setting through a sensitivity analysis (Appendix D.1). For each word w, we extract

$\mathbf{E}_w \in \mathbb{R}^{N\times K}$, where N is the number of electrodes (participant-specific) and $K = 102$ time points (200 ms at 512 Hz). While the framework extends to additional configurations, we focus on these two theoretically motivated cases that target distinct processing stages.

3.3.2 Training Objective

We design the training objective to preserve ECoG's spatiotemporal structure: models predict the full response matrix $\mathbf{E}_w \in \mathbb{R}^{N\times K}$ for each word, capturing both spatial patterns across N electrodes and temporal dynamics across K time points. Figure 1 illustrates the ECoG-tuning framework.

Audio input. Whisper requires 30-second input segments; we adopt this context window across all three models to maintain a consistent framework:

$$\mathbf{A}_w = \mathbf{A}[\max(0, t_e - \Delta) : t_e] \tag{2}$$

where $\Delta = 30$ s and $t_e = t_w + 200$ ms.

Feature aggregation. Let $\mathbf{H}^{(l)} \in \mathbb{R}^{M\times D}$ denote hidden states from encoder layer l. We pool over the final $T = 10$ frames (200 ms at 50 Hz) and concatenate across all $L = 12$ layers:

$$\mathbf{z} = \bigoplus_{l=1}^{L} \left(\frac{1}{T} \sum_{m=M-T+1}^{M} \mathbf{H}_m^{(l)} \right) \in \mathbb{R}^{LD} \tag{3}$$

Spatiotemporal prediction. A linear projection maps the aggregated representation to the target:

$$\hat{\mathbf{E}}_w = \mathbf{W}\mathbf{z} + \mathbf{b} \in \mathbb{R}^{N\times K} \tag{4}$$

Loss function. We minimize mean squared error (MSE) over the spatiotemporal response:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{w\in\mathcal{B}} \left\| \hat{\mathbf{E}}_w - \mathbf{E}_w \right\|_F^2 \tag{5}$$

where $\mathcal{B}$ is the training batch. We validate this choice against correlation-based and hybrid alternatives in Appendix D.3.

We adopt a linear projection and MSE loss intentionally to isolate neural signal contribution from architectural complexity, so that observed improvements can be more directly attributed to the neural supervision signal. The procedure is summarized in Algorithm 1.

Algorithm 1 ECoG-Tuning

Require: Audio $\mathbf{A}$, ECoG recordings $\mathbf{E}$, word onsets $\{t_1, \ldots, t_W\}$, pretrained encoder $\boldsymbol{\theta}_{\text{enc}}$, window type $\tau \in \{\text{speech}, \text{lang}\}$ (corresponding to $\mathbf{W}_{\text{speech}}$, $\mathbf{W}_{\text{lang}}$)
Ensure: Fine-tuned encoder $\boldsymbol{\theta}^*_{\text{enc}}$

```
Initialize projection head θ_proj
Freeze feature extractor θ_feat
for epoch = 1 to max_epochs do
    for each batch B of words do
        for each word w ∈ B with onset t_w do
            [t_s, t_n] ← W(t_w, τ) {Neural window bounds}
            A_w ← A[max(0, t_e − Δ) : t_e] {30 s context}
            E_w ← ExtractWindow(E, t_s, t_n)
            {H^(l)}_{l=1}^L ← Encoder(A_w; θ_enc)
            z ← ⊕_{l=1}^L MeanPool(H^(l)_{−T:}) {Last T frames}
            Ê_w ← Linear(z; θ_proj)
        end for
        L ← (1/|B|) Σ_{w∈B} ||Ê_w − E_w||_F^2
        Update θ_enc, θ_proj via Adam
    end for
    if no improvement for patience epochs then
        break
    end if
end for
return θ*_enc
```

3.3.3 Training Protocol

We train separate models for each participant, as clinically-determined electrode placement varies across individuals ($N = 72$–235; Appendix A.2); cross-participant training is validated in Appendix D.7. Data are split into training (80%), validation (10%), and held-out test (10%) sets for final evaluation. Training uses Adam with learning rates 1×10^{-5} (encoder) and 2×10^{-5} (projection), with early stopping (patience = 5). Complete hyperparameters are provided in Appendix B.1.

3.4 Control Conditions

To isolate the contributions of stimulus-aligned temporal dynamics in ECoG signals, we implement four control conditions, each targeting a distinct aspect of the training signal:

Permuted-ECoG (PE). ECoG responses are block-permuted to disrupt audio–neural correspondence while preserving signal statistics. This tests

whether stimulus alignment—rather than the presence of realistic neural-like signals—drives the alignment gains.

Temporal-Mean (TM). The spatiotemporal target $\mathbf{E}_w \in \mathbb{R}^{N\times K}$ is replaced by its temporal average $\bar{\mathbf{E}}_w \in \mathbb{R}^N$. This isolates the contribution of millisecond-resolved dynamics beyond time-averaged spatial patterns. As a finer-grained complement, we also evaluate a Temporal-Shuffled (TS) variant that retains the $N \times K$ target but randomly permutes its time points within each electrode, separating temporal structure from target dimensionality (Appendix D.2).

BigSLM-Tuned. Representations from larger versions of SLMs (∼1 B) serve as training targets, testing whether distillation from a larger model of the same family provides comparable benefits.

LLM-Tuned. Text representations from Mistral-7B serve as targets, testing whether text-derived linguistic features—which capture semantic structure without neural dynamics—can substitute for direct neural supervision.

3.5 Evaluation

3.5.1 Brain Alignment

We evaluate model-brain alignment using encoding analysis adapted from Goldstein et al. (2025b). For each electrode n, we fit a linear encoding model mapping representations to neural responses. Model embeddings are standardized and used to train a ridge regression encoder (replacing the original OLS) to predict the neural response. Brain alignment is quantified as the Pearson correlation between predicted and actual responses:

$$r_n = \text{corr}(\hat{\mathbf{e}}_n, \mathbf{e}_n) \tag{6}$$

Aggregate alignment across electrodes is:

$$B = \frac{1}{N}\sum_{n=1}^{N} r_n \tag{7}$$

Results are reported on the held-out test partition (see Appendix C.1 for complete details).

3.5.2 Downstream Tasks

To assess whether ECoG-tuning improves linguistic representations beyond brain alignment, we evaluate on three speech understanding tasks following Moussa et al. (2025): Phonemes Prediction, Phonetic Sentence Type Prediction (TIMIT; Garofolo et al., 1993), and Emotion Recognition (CREMA-D; Cao et al., 2014). These tasks span different levels of linguistic abstraction. We train linear probes on frozen model representations and report macro F1-score averaged across encoder layers. Detailed task specifications are provided in Appendix C.2.

4 Results

4.1 ECoG-Tuning Improves Brain Alignment

We first evaluate whether ECoG-tuning enhances the alignment between SLM representations and neural responses. Figure 2 presents comprehensive results across all three model families.

Overall Alignment Improvements. ECoG-tuning consistently improves brain alignment compared to pretrained models across all three architectures (Figure 2 a, d, g). Whisper exhibits the largest gains ($\Delta r = +0.062$, Cohen's $d = 0.72$, $p < 0.001$), and HuBERT and Wav2Vec 2.0 show improvements of $\Delta r \approx +0.04$–0.06 across both temporal windows ($p < 0.05$). Importantly, ECoG-tuned models also outperform the Permuted-ECoG control ($p < 0.05$), indicating that improvements require properly stimulus-aligned neural signals rather than arbitrary neural statistics.

Comparison with Distillation Approaches. ECoG-tuning consistently outperforms both distillation baselines. This suggests that neural supervision provides training signal properties not fully captured by scaling model capacity or leveraging text-derived representations alone. Compared to BigSLM-Tuned, ECoG-tuned models show robust advantages across architectures (Cohen's $d = 0.71$–0.79 for Whisper and HuBERT, $p < 0.05$). Compared to LLM-Tuned, improvements are particularly pronounced for HuBERT and Wav2Vec 2.0 ($d = 0.91$–1.03, $p < 0.05$). These results highlight the contribution of direct neural supervision beyond model scaling.

Layer-wise Patterns. The benefits of neural supervision propagate across the model's representational hierarchy (Figure 2 b, e, h; Appendix D.4 for details). ECoG-tuned representations consistently outperform their pretrained counterparts throughout all encoder layers, indicating that neural supervision refines representations at multiple levels of abstraction. The magnitude of improvement varies across architectures: Whisper exhibits the most pronounced gains with peak improvements of approximately 50% in layers 9–10, while HuBERT and

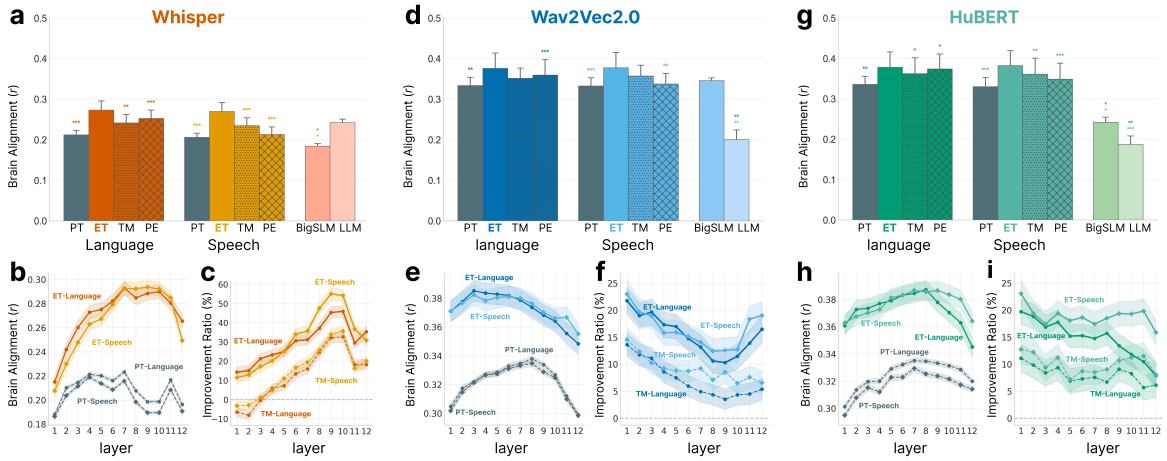

Figure 2: **ECoG-tuning significantly improves brain alignment across all three SLMs (Δr up to +0.062, Cohen's d = 0.72). (a, d, g)** Overall alignment (Pearson r) under ECoG-tuning (ET), pretrained baseline (PT), and controls (TM: Temporal-Mean; PE: Permuted-ECoG). **(b, e, h)** Layer-wise alignment; solid and colored: ECoG-tuned, dashed and gray: pretrained. **(c, f, i)** Layer-wise improvement (%) over pretrained; solid: ECoG-tuned, dashed: Temporal-Mean. Error bars and shaded regions: 95% *CI*.

Wav2Vec 2.0 show more uniform improvements distributed across layers (Figure 2 c, f, i). Notably, all three models show positive improvements even in the early layers, suggesting that ECoG-tuning affects representations throughout the encoder stack.

4.2 Contribution of Temporal Dynamics

We examine whether the temporal precision of ECoG provides value beyond that of time-averaged signals from two perspectives.

Full temporal dynamics outperform time-averaged supervision. To quantify the contribution of temporal information, we compare ECoG-tuning with full spatiotemporal targets against Temporal-Mean, which collapses 200 ms of neural dynamics into a time-averaged vector. Both approaches improve alignment over pretrained baselines on average (Figure 2 c, f, i); however, preserving temporal dynamics consistently outperforms the TM counterpart across all layers, windows, and architectures (Appendix D.5). This advantage averages ∼17% for Whisper, 7%–8% for Wav2Vec 2.0, and 7%–9% for HuBERT (Cohen's $d = 0.20$–0.34, $p < 0.05$ for Whisper and HuBERT). These results suggest that millisecond-resolution temporal dynamics carry information beyond time-averaged patterns. The Temporal-Shuffled control yields lower alignment than intact ECoG-tuning across three SLMs (Appendix D.2), further suggesting that temporal structure, rather than target dimensionality alone, contributes to the advantage.

Temporal Dissociation. Comparing $\mathbf{W_{lang}}$ and $\mathbf{W_{speech}}$ reveals stage-specific effects. $\mathbf{W_{lang}}$ tuning yields larger gains in language-responsive regions across models (Table 1), with Whisper and HuBERT showing significance ($p < 0.05$).

Model	Metric	$\mathbf{W_{lang}}$	$\mathbf{W_{speech}}$
Whisper	Δr	**0.054***	0.043
	Rel. Impr.	**24.7%**	15.2%
HuBERT	Δr	**0.058***	0.033
	Rel. Impr.	**22.9%**	1.4%
Wav2Vec 2.0	Δr	**0.045**	0.021
	Rel. Impr.	**10.0%**	7.7%

Table 1: **Alignment gains by window condition**. Δr: absolute improvement; Rel. Impr.: percentage improvement. Bold: higher value; *: $p < 0.05$ (significant).

The spatial distribution of improvements (Figure 3) corroborates this pattern: while both window conditions produce gains throughout the recorded cortex, $\mathbf{W_{lang}}$ tuning yields notably stronger effects in regions linked to lexical-semantic processing—particularly in IFG and MFG for Whisper and HuBERT. This temporal dissociation is consistent with neurophysiological evidence that higher-order language areas exhibit peak encoding within the $\mathbf{W_{lang}}$ interval, though residual acoustic correlates within this interval cannot be fully ruled out. All models achieve robust alignment improvements under both window conditions, though window selection modulates region-specific effects.

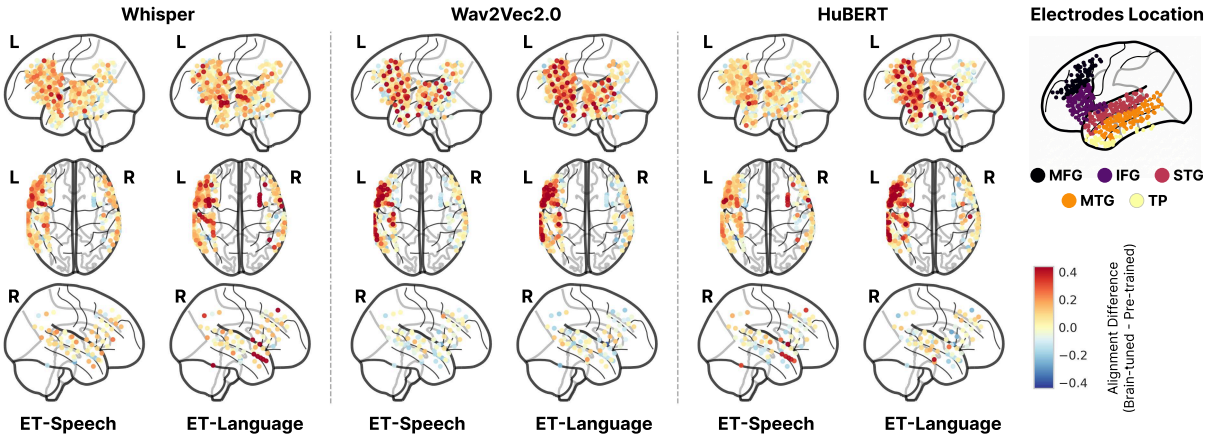

Figure 3: **Spatial distribution of alignment changes across the electrode array.** Glass brain visualization showing alignment difference (Δr = ECoG-tuned - Pretrained) for each model under $\mathbf{W}_{\text{lang}}$ and $\mathbf{W}_{\text{speech}}$. Node color: red = improved alignment, blue = decreased alignment. Electrode positions are displayed in MNI space.

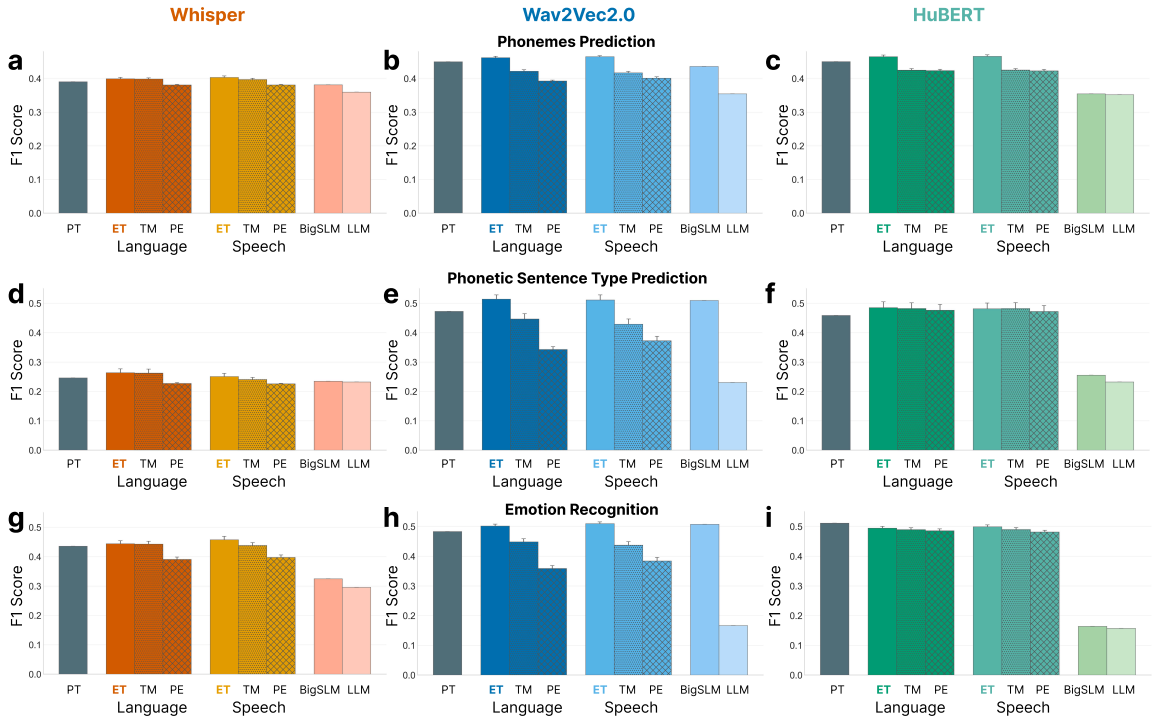

Figure 4: **Downstream task performance of ECoG-tuned models and their comparisons.** Bar colors/styles as in Figure 2 a,d,g. Error bars: 95% *CI* across layers and participants.

4.3 ECoG-Tuning Improves or Maintains Downstream Performance

We evaluate whether the benefits of neural supervision extend beyond brain alignment to practical capabilities. ECoG-tuned models consistently exceed or match pretrained counterparts, indicating that neural supervision does not compromise—and often enhances—practical utility (Figure 4).

Phonemes prediction shows the most consistent improvements: all models outperform pretrained baselines and distillation controls, suggesting that ECoG-tuning refines acoustic-phonetic representations. **Phonetic sentence type prediction** improves for all models relative to pretrained baselines, with Whisper and HuBERT exceeding distillation conditions. **Emotion recognition** shows a nuanced pattern: Whisper and HuBERT improve over baselines, while Wav2Vec 2.0 maintains comparable performance, consistent with findings that emotion recognition relies on prosodic features well-captured during self-supervised pretrain-

ing (Moussa et al., 2025).

Moreover, cross-participant training yields similar improvements (4–7% on phonemes prediction and emotion recognition, +31% on phonetic sentence type prediction; Appendix D.7), and window selection shows task-dependent trends: $\mathbf{W_{lang}}$ tends to yield relatively stronger performance on linguistically demanding tasks while $\mathbf{W_{speech}}$ tends to favor acoustic-phonetic tasks (Appendix D.6).

5 Discussion

Our experiment results provide complementary evidence: ECoG-tuned models achieve substantially improved brain alignment, and ECoG's millisecond precision enables training strategies unavailable to hemodynamic imaging. Together with prior fMRI-based works, these findings suggest that neural supervision benefits are not specific to a single recording modality, while ECoG's temporal resolution opens new methodological possibilities.

Beyond the methodological contribution, our framework affords interpretive value through an interventional lens: by manipulating the supervision signal—its temporal structure, window placement, and stimulus alignment—and observing the resulting changes in alignment and downstream behavior, we can investigate which components contribute to brain-aligned representations. Regarding layer-wise patterns, we note that strict functional-hierarchy interpretations of layer alignment warrant caution, since individual layers may encode a mixture of processing levels (Niu et al., 2022).

The robust alignment improvements across all three architectures (with large effect sizes) suggest that neural supervision provides information not readily obtainable through model scaling alone: ECoG-tuned models consistently outperform both distillation baselines despite the latter leveraging larger pretrained models. Given that SLMs have been shown to lack the brain-relevant semantics present in text models (Oota et al., 2024), neural recordings—which capture processing across multiple linguistic levels, including semantics—may offer a pathway to narrow this gap. The cross-modality consistency between our ECoG results and prior fMRI findings (Moussa et al., 2025) strengthens the case that brain-tuning reflects meaningful alignment with human language processing rather than modality-specific artifacts.

Two lines of evidence are consistent with the value of this temporal precision. First, preserving full spatiotemporal dynamics outperforms time-averaged supervision, indicating that millisecond-resolution structure carries information beyond the temporally integrated pattern. Second, the temporal window dissociation—where language-window tuning preferentially improves alignment in language-responsive regions—is consistent with established findings that acoustic-phonetic encoding peaks early (∼54 ms) while lexical-semantic processing peaks later (∼247 ms) (Goldstein et al., 2025b). This correspondence suggests that neuroscientific insights into temporal hierarchies can inform principled training strategies, translating discoveries about when the brain encodes different aspects of language into actionable methodologies.

Finally, ECoG-tuned models consistently exceed or match pretrained baselines on downstream tasks, showing that enhanced alignment does not compromise practical utility. Recent work has identified brain-like representations as causally relevant to task performance (AlKhamissi et al., 2025a), providing a possible account for the observed link between alignment and practical gains. The pattern where phonemes prediction shows consistent gains aligns with ECoG's fine-grained temporal structure, suggesting that ECoG-tuning particularly refines acoustic-phonetic representations. While our main experiments use per-participant training to accommodate individual differences in electrode placement, cross-participant experiments yield robust improvements (Appendix D.7).

Looking ahead, the framework may extend to non-invasive electrophysiology: EEG provides analogous millisecond-resolved multichannel recordings, though its lower SNR will necessitate dedicated feature engineering to establish robust cognitive supervision signals; such extensibility could help address the data scarcity inherent to invasive recordings. More broadly, future work could explore multi-participant strategies (Moussa and Toneva, 2025), integration with diverse electrophysiological datasets, multi-window supervision (see preliminary results in Appendix D.1), and multilingual extension. We hope this work contributes to the broader convergence of neuroscience and NLP (Appendix G).

6 Conclusion

We introduced ECoG-tuning, a training methodology that leverages intracranial electrophysiology to supervise speech language models on neural re-

sponses from distinct processing stages. Our experiments show that ECoG's fine-grained temporal dynamics provide valuable supervision: model-brain alignment benefits from neural supervision beyond what distillation provides, temporal structure carries information beyond time-averaged patterns, and targeting different processing stages yields region-specific effects consistent with cortical processing hierarchies. These alignment gains are accompanied by preserved or improved downstream performance, supporting practical utility. By bridging millisecond resolution neural recordings and model training, this work offers a proof of concept for incorporating temporally precise neuroscientific insights into brain-aligned artificial systems.

Limitations

Our study has several limitations that suggest directions for future work. The Podcast ECoG dataset, while rare and valuable, remains constrained in scale—comprising recordings from nine participants listening to a 30-minute stimulus—which limits the linguistic diversity encountered during training. Moreover, because ECoG electrode placement is dictated entirely by clinical needs, cortical coverage is variable and incomplete across participants, making region-of-interest-based tuning a worthwhile area for future exploration. More broadly, expanding to additional ECoG datasets as they become available and extending to more scalable non-invasive modalities are promising directions. Finally, direct comparison between ECoG-tuning and fMRI-based brain-tuning is complicated by fundamental differences in signal origin, temporal and spatial scales, and available datasets. We view them as complementary along these dimensions rather than competing, and jointly leveraging the strengths of both modalities during training remains an open and promising question.

Acknowledgements

This work was supported in part by the National Science and Technology Major Project under Grant 2024YFC3307800 and National Natural Science Foundation of China (Grant No. 62176024).

References

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025a. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, Albuquerque, New Mexico. Association for Computational Linguistics.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025b. From language to cognition: How LLMs outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24332–24350, Suzhou, China. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Arnab Bhattacharjee, Zaid Zada, Haocheng Wang, Bobbi Aubrey, Werner Doyle, Patricia Dugan, Daniel Friedman, Orrin Devinsky, Adeen Flinker, Peter J. Ramadge, Uri Hasson, Ariel Goldstein, and Samuel A. Nastase. 2026. Aligning brains into a shared space improves their alignment with large language models. *Nature Computational Science*, 6:169–178.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

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

Nathan E. Crone, Alon Sinai, and Anna Korzeniewska. 2006. High-frequency gamma oscillations and human brain mapping with electrocorticography. In *Progress in Brain Research*, volume 159, pages 275–295.

Xiao Ding, Bowen Chen, Li Du, Bing Qin, and Ting Liu. 2022. CogBERT: Cognition-guided pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3210–3225, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Berta Franzluebbers, Donald Dunagan, Miloš Stanojević, Jan Buys, and John Hale. 2024. Multipath parsing in the brain. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12215–12229, Bangkok, Thailand. Association for Computational Linguistics.

Changjiang Gao, Jixing Li, Jiajun Chen, and Shujian Huang. 2024. Measuring meaning composition in

the human brain with composition scores from large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11295–11308, Bangkok, Thailand. Association for Computational Linguistics.

John S. Garofolo, Lori F. Lamel, William M. Fisher, David S. Pallett, Nancy L. Dahlgren, Victor Zue, and Jonathan G. Fiscus. 1993. Timit acoustic-phonetic continuous speech corpus.

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, Stephen M. Smith, and David C. Van Essen. 2016. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A. Nastase, Zaid Zada, Eric Ham, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, Werner Doyle, Sasha Devore, Patricia Dugan, Roi Reichart, Daniel Friedman, Michael Brenner, Avinatan Hassidim, and 3 others. 2024. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15(1):2768.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, and 3 others. 2025a. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16(1):10529.

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, and 4 others. 2025b. 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.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, and 13 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Rémi King. 2025. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences*, 122(42):e2422097122.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. 2025a. Large language models as neurolinguistic subjects: Discrepancy between performance and competence. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19284–19302, Vienna, Austria. Association for Computational Linguistics.

Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum, and Nima Mesgarani. 2025b. Far from the shallow: Brain-predictive reasoning embedding through residual disentanglement. *Preprint*, arXiv:2510.22860.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. In *Advances in Neural Information Processing Systems*, volume 31, pages 6629–6638. Curran Associates, Inc.

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

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

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Rémi King. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 33428–33443. Curran Associates, Inc.

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. 2024. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 6(12):1467–1477.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. 2025. Improving semantic understanding in speech language models via brain-tuning. In *The Thirteenth International Conference on Learning Representations*.

Omer Moussa and Mariya Toneva. 2025. Brain-tuning improves generalizability and efficiency of brain alignment in speech models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. 2024. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20313–20338, Miami, Florida, USA. Association for Computational Linguistics.

Anuja Negi, Subba Reddy Oota, Anwar O Nunez-Elizalde, Manish Gupta, and Fatma Deniz. 2025. Brain-informed fine-tuning for improved multilingual understanding in language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3143–3153, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Subba Reddy Oota, Zijiao Chen, Manish Gupta, Bapi Raju Surampudi, Gael Jobard, Frederic Alexandre, and Xavier Hinaut. 2025. Deep neural networks and brain alignment: Brain encoding and decoding (survey). *Transactions on Machine Learning Research*.

Nico Policzer, Cameron Braunstein, and Mariya Toneva. 2025. The one where they brain-tune for social cognition: Multi-modal brain-tuning on friends. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Joséphine Raugel, Jérémy Rapin, Stéphane d'Ascoli, Valentin Wyart, and Jean-Remi King. 2025. Scaling and context steer LLMs along the same computational path as the human brain. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yuqi Ren and Deyi Xiong. 2021. CogAlign: Learning to align textual neural representations to cognitive language processing signals. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3758–3769, Online. Association for Computational Linguistics.

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

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*. Preprint.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. In *Advances in Neural Information Processing Systems*, volume 32, pages 14123–14133. Curran Associates, Inc.

Padakanti Srijith, Khushbu Pahwa, Radhika Mamidi, Bapi Raju Surampudi, Manish Gupta, and Subba Reddy Oota. 2025. Aligning text/speech representations from multimodal models with MEG brain activity during listening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34457–34474, Suzhou, China. Association for Computational Linguistics.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32, pages 14954–14964. Curran Associates, Inc.

Nishitha Vattikonda, Aditya R. Vaidya, Richard J. Antonello, and Alexander G. Huth. 2025. Brainwavlm: Fine-tuning speech representations with brain responses to language. *Preprint*, arXiv:2502.08866.

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

Congchi Yin, Qian Yu, Zhiwei Fang, Changping Peng, and Piji Li. 2025a. Rethinking cross-subject data splitting for brain-to-text decoding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5686–5700, Suzhou, China. Association for Computational Linguistics.

Congchi Yin, Yongpeng Zhang, Xuyun Wen, and Piji Li. 2025b. Improve language model and brain alignment via associative memory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 986–999, Vienna, Austria. Association for Computational Linguistics.

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. 2025. The "Podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12(1):1135.

A Data and Preprocessing

A.1 ECoG Preprocessing Pipeline

Table 2 summarizes the ECoG preprocessing parameters. We use the preprocessed high-gamma data provided by the Podcast dataset (Zada et al., 2025), which follows established ECoG methodology (Crone et al., 2006). High-gamma band (70–200 Hz) power serves as a reliable index of local neuronal population activity.

Step	Parameter	Value
Input	Raw sampling rate	512 Hz or 2,048 Hz
	Resampled rate	512 Hz
	Electrode status	"good"
	Electrode type	"ECOG"
	Localization	Valid MNI152 coordinates
Artifact Removal	Despiking threshold	>4 quartiles above median
	Despiking interpolation	pchip
	Re-referencing	Common average
Filtering	High-gamma band	70–200 Hz
	Filter type	Butterworth IIR
	Notch filter	60, 120, 180, 240 Hz
Envelope	Method	Hilbert transform
Normalization	Method	Z-score per electrode
Time Windows	Duration	200 ms (K=102 samples)
	$\mathbf{W}_{\text{speech}}$	$[t_w-50, t_w+150]$ ms
	$\mathbf{W}_{\text{lang}}$	$[t_w+150, t_w+350]$ ms

Table 2: **ECoG preprocessing pipeline parameters.**

The preprocessing pipeline proceeds as follows: (1) downsample to 512 Hz if necessary; (2) despike and interpolate high-amplitude outliers; (3) apply common average re-referencing; (4) notch filter at 60, 120, 180, and 240 Hz to remove power line noise; (5) bandpass filter to 70–200 Hz; (6) compute amplitude envelope via Hilbert transform; (7) z-score normalize per electrode across the session.

A.2 Electrode Coverage and Selection

Table 3 presents electrode counts and regional distribution across subjects.

Selection criteria. Electrodes were included based on three criteria: (1) status marked as "good" in the original dataset metadata; (2) type classified as "ECOG" (excluding depth electrodes and auxiliary channels); (3) valid MNI152 coordinates

Part.	Total	Loc.	Lang.	Aud.	Motor	Other
P01	124	99	43	0	7	49
P02	114	90	22	3	2	63
P03	264	235	110	4	5	116
P04	174	143	50	6	9	78
P05	167	159	55	0	22	82
P06	178	166	70	1	24	71
P07	138	116	44	1	14	57
P08	91	72	29	2	4	37
P09	205	188	97	0	17	74
Total	**1,455**	**1,268**	**520**	**17**	**104**	**627**

Table 3: **Electrode distribution across subjects**. **Total**: all channels; **Loc.**: electrodes meeting inclusion criteria.

Category	Parameter	Value
Model	Architectures	Whisper, Wav2Vec 2.0, HuBERT
	Encoder layers	12 (trainable)
	Hidden dimension D	768
	Feature extractor	Frozen
	Projection head	$L \times D \to N \times K$
	Projection dropout	0.3
Optimization	Optimizer	Adam (β_1=0.9, β_2=0.999)
	LR (encoder)	1×10^{-5}
	LR (projection)	2×10^{-5}
	Weight decay (enc.)	1×10^{-4}
	Weight decay (proj.)	2×10^{-4}
	LR schedule	Linear warmup (10%) + linear decay
	Gradient clipping	max_norm = 1.0
Training	Batch size	32
	Maximum epochs	30
	Audio sample rate	16 kHz
	Audio context	$[\max(0,\ t_e - 30\text{s}),\ t_e]$
Early Stopping	Metric	Validation correlation
	Patience	5 epochs
	Min improvement δ	0.001
	Overfitting threshold	0.03 (train–val gap)
Data Split	Train:Val:Test	8:1:1

Table 4: **Hyperparameter configuration**.

available. Of the 1,455 total channels, 187 were excluded (80 without localization, 31 with noisy signals, 76 non-brain channels).

Regional mapping. Electrodes were mapped to the Glasser cortical parcellation (Glasser et al., 2016) using MNI152 coordinates. Language responsive regions include: middle temporal gyrus (MTG), inferior frontal gyrus (IFG), angular gyrus (AG), middle frontal gyrus (MFG), superior temporal gyrus (STG; excluding primary and early auditory cortices) and temporal pole (TP).

B Model Training

B.1 Hyperparameter Configuration

Table 4 provides the complete hyperparameter settings for ECoG-tuning.

B.2 Computational Resources

Table 5 summarizes the computational resources required for the training phase of ECoG-tuning experiments.

C Evaluation Details

C.1 Brain Alignment Computation

Representation Extraction. For each word w, we extract hidden states from the last T=10 frames

Resource	Specification
Hardware	
GPU	NVIDIA RTX 4090 (24 GB) × 2
CPU	Intel Xeon Gold 6459C
RAM	72 GB
Training Time	
Per model (average)	∼2 h
Total per architecture	∼112 GPU hours
All experiments (3 arch.)	∼330 GPU hours
Software	
Framework	PyTorch 2.7
Transformers	HuggingFace 4.57
CUDA / Driver	12.6 / 580.76

Table 5: **Computational resources for ECoG-tuning** (training phase only). Total time includes main experiments (9 participants × 2 windows), shuffle control, time-averaged control, and distillation control. Training times vary with participant electrode count.

(corresponding to the last 200 ms of the 30 s audio input) and mean-pool them to form the word-level representation:

$$\tilde{\mathbf{z}}_w^{(l)} = \frac{1}{T} \sum_{m=M-T+1}^{M} \mathbf{H}_m^{(l)} \in \mathbb{R}^D \tag{8}$$

where $D = 768$ is the hidden dimension, yielding a 768-dimensional vector for each layer l (Eq. 3).

Embedding Preprocessing. Within each cross-validation fold, embeddings $\{\tilde{\mathbf{z}}_w^{(l)}\}$ are standardized on the training partition and applied to the test data to prevent information leakage.

Cross-Validation. We segment the test partition of each participant into temporally contiguous folds and perform 10-fold cross-validation. This design aims to mitigate leakage from temporal autocorrelation in neural time series.

Encoding Model. For each encoder layer l and electrode n, we fit an independent ridge regression model to predict the time-averaged neural response:

$$\hat{e}_{n,w}^{(l)} = \boldsymbol{\beta}_n^{(l)\top} \tilde{\mathbf{z}}_w^{(l)} + b_n^{(l)} \tag{9}$$

where $\tilde{\mathbf{z}}_w^{(l)} \in \mathbb{R}^D$ ($D = 768$) is the standardized layer-l embedding for word w, and $\boldsymbol{\beta}_n^{(l)}, b_n^{(l)}$ are learned parameters. The regularization parameter λ is selected via cross-validation on the training set from a logarithmic grid $\{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$.

Correlation Computation. For each electrode, we compute the Pearson correlation between predicted and actual neural responses on the held-out

test fold:

$$r_n^{(l)} = \text{corr}(\hat{\mathbf{e}}_n^{(l)}, \mathbf{e}_n) \tag{10}$$

where $\hat{\mathbf{e}}_n^{(l)}$ and $\mathbf{e}_n$ are vectors of predicted and actual responses across all test words. Layer-wise alignment is then averaged across the 10 test folds.

Aggregate Alignment. We report brain alignment at multiple granularities. Overall alignment aggregates across all N electrodes:

$$B_{\text{all}} = \frac{1}{N} \sum_{n=1}^{N} r_n \tag{11}$$

To examine region-specific effects, we also compute alignment separately for electrodes within each region of interest (ROI). For a given ROI $\mathcal{R}$ containing $|\mathcal{R}|$ electrodes:

$$B_{\mathcal{R}} = \frac{1}{|\mathcal{R}|} \sum_{n \in \mathcal{R}} r_n \tag{12}$$

We report B_{lang} for language-responsive regions and B_{all} for all recorded electrodes, enabling comparison of ECoG-tuning effects across functionally distinct cortical areas.

C.2 Downstream Task Specifications

We evaluate ECoG-tuned models and controls on three downstream tasks spanning different levels of linguistic abstraction (Moussa et al., 2025).

Phonemes Prediction (TIMIT). Multi-label classification of 39 phonemes using the TIMIT Acoustic-Phonetic Corpus (Garofolo et al., 1993). Given an audio segment, the classifier predicts which phonemes are present. We use linear probes on frozen layer representations and report macro F1-score on the standard test partition.

Phonetic Sentence Type Prediction (TIMIT). Three-way classification of phonetic sentence types: SA (dialectal sentences designed to cover all English phonemes), SX (phonetically balanced sentences with extensive coverage using minimal words), and SI (phonetically diverse, naturalistic sentences). This task requires understanding phonetic structure beyond individual sounds. Evaluation uses macro F1-score with linear probes.

Emotion Recognition (CREMA-D). Six-way classification (Anger, Disgust, Fear, Happy, Neutral, Sad) on the CREMA-D dataset (Cao et al., 2014), comprising 7,442 clips from 91 actors. This task primarily relies on prosodic and paralinguistic features rather than lexical content. Evaluation uses macro F1-score with linear probes.

For all tasks, we train linear classifiers on frozen representations from each encoder layer and report the mean F1-score (Moussa et al., 2025).

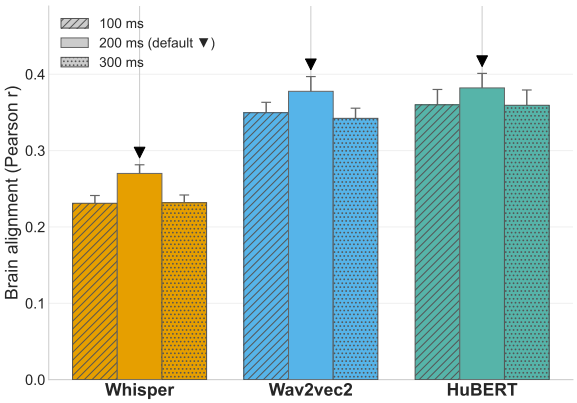

Figure 5: Window duration sensitivity under $\mathbf{W}_{\text{speech}}$, with window center fixed at $+50$ ms post-word-onset. Error bars: 95% *CI*.

D Supplementary Results

D.1 Window Duration Sensitivity

We examine the robustness of the 200 ms choice under the speech condition, varying the window length while keeping the center fixed at $+50$ ms post-word-onset. Figure 5 presents brain alignment across 100 ms, 200 ms, and 300 ms durations—corresponding to windows $[0, +100]$, $[-50, +150]$, and $[-100, +200]$ ms—for all three SLMs. The 200 ms window configuration yields the highest alignment across Whisper, Wav2Vec 2.0, and HuBERT, supporting the theory-driven default choice.

As a related exploration, we investigated a joint dual-window setting in which each training word contributes two equally-weighted loss terms, one for $\mathbf{W}_{\text{speech}}$ and one for $\mathbf{W}_{\text{lang}}$. The resulting model improves over the pretrained baseline but does not surpass the best single-window condition. How to best reconcile these two supervision signals (e.g., through adaptive weighting or curriculum-based scheduling) is an open question worth dedicated investigation.

D.2 Temporal Structure Ablation

To isolate the contribution of temporal structure from that of target dimensionality, we introduce a Temporal-Shuffled (TS) control: the $N \times K$ target is retained, but the K time points are randomly

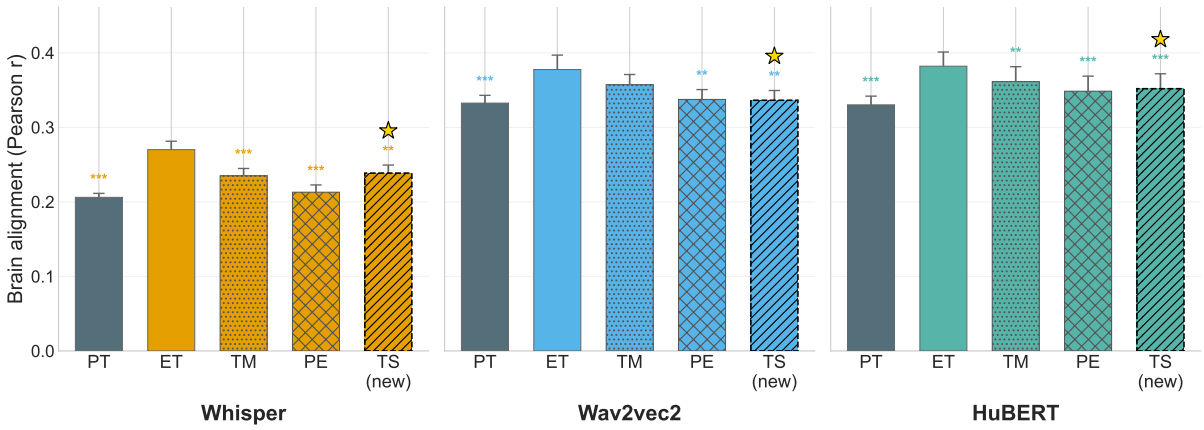

Figure 6: **Temporal structure ablation under $\mathbf{W}_{\text{speech}}$.** TS matches ET in target dimensionality ($N \times K$) but randomly permutes time points within each electrode. Error bars: 95% *CI*.

permuted within each electrode. TS matches ET in target size while destroying stimulus-locked temporal ordering.

Figure 6 reports brain alignment under the $\mathbf{W}_{\text{speech}}$ across the three models. ET outperforms TS consistently (Whisper: Cohen's $d = 0.30$, $p = 0.003$; Wav2Vec 2.0: $d = 0.28$, $p = 0.005$; HuBERT: $d = 0.46$, $p < 0.001$). These results indicate that a stimulus-aligned supervision signal with its original temporal structure preserved is a key contributor to the alignment gains.

D.3 Loss Function Comparison

To validate our choice of MSE loss, we conduct preliminary experiments on Whisper comparing against two alternative objectives motivated by Moussa and Toneva (2025):

Correlation Loss. Minimizes the negative Pearson correlation:

$$\mathcal{L}_{\text{corr}} = \frac{1}{|\mathcal{B}|} \sum_{w \in \mathcal{B}} (1 - r(\hat{\mathbf{e}}_w, \mathbf{e}_w)) \tag{13}$$

where $\hat{\mathbf{e}}_w = \text{vec}(\hat{\mathbf{E}}_w)$ and $\mathbf{e}_w = \text{vec}(\mathbf{E}_w)$ are vectorized predictions and targets, and $r(\cdot, \cdot)$ denotes Pearson correlation.

Cosine + MSE Loss. Combines cosine similarity with MSE:

$$\mathcal{L}_{\text{cos+MSE}} = \lambda \left(1 - \frac{\hat{\mathbf{e}}_w^\top \mathbf{e}_w}{\|\hat{\mathbf{e}}_w\|_2 \|\mathbf{e}_w\|_2}\right) + (1-\lambda)\mathcal{L}_{\text{MSE}} \tag{14}$$

where $\lambda = 0.5$ for equal weighting.

Figure 7 presents the results of loss function comparisons across both temporal windows. The default MSE loss achieves the best performance

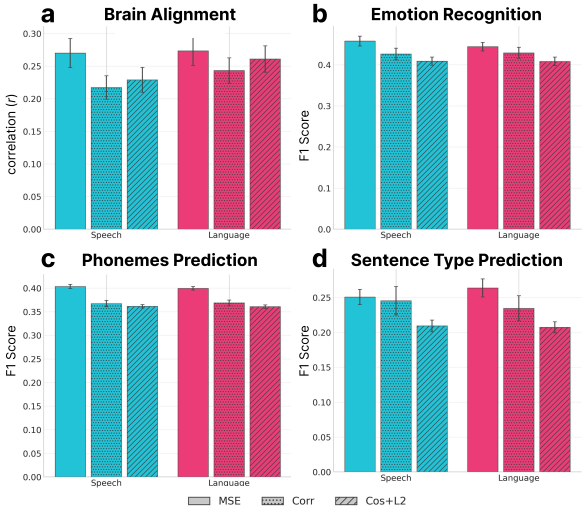

Figure 7: **Loss function comparison results.**

across all metrics, including brain alignment and downstream tasks. Based on these findings, we adopt MSE as the training objective and extend it to all models in our main experiments.

D.4 Layer-wise Brain Alignment

Table 6 presents brain alignment (Pearson r) for each encoder layer across models and conditions.

D.5 Layer-wise Improvement Ratio

Table 7 presents the layer-wise improvement ratio (%) relative to pretrained baselines for ECoG-tuned (ET) and Temporal-Mean (TM) models.

D.6 Downstream Task Performance by Window Condition

Table 8 provides a comparison between $\mathbf{W}_{\text{speech}}$ and $\mathbf{W}_{\text{lang}}$ tuning across downstream tasks.

Layer	Whisper-small				Wav2Vec 2.0-base				HuBERT-base			
	PT-La	ET-La	PT-Sp	ET-Sp	PT-La	ET-La	PT-Sp	ET-Sp	PT-La	ET-La	PT-Sp	ET-Sp
1	0.193	**0.216**	0.192	**0.209**	0.315	**0.372**	0.310	**0.373**	0.310	**0.364**	0.305	**0.366**
2	0.215	**0.242**	0.209	**0.230**	0.329	**0.379**	0.325	**0.379**	0.325	**0.377**	0.319	**0.372**
3	0.219	**0.261**	0.216	**0.250**	0.334	**0.388**	0.333	**0.385**	0.331	**0.379**	0.328	**0.375**
4	0.225	**0.274**	0.223	**0.265**	0.339	**0.386**	0.340	**0.382**	0.332	**0.382**	0.325	**0.378**
5	0.224	**0.278**	0.218	**0.271**	0.339	**0.385**	0.341	**0.384**	0.341	**0.384**	0.338	**0.386**
6	0.219	**0.285**	0.212	**0.283**	0.344	**0.385**	0.343	**0.386**	0.345	**0.387**	0.339	**0.388**
7	0.226	**0.297**	0.219	**0.298**	0.345	**0.383**	0.345	**0.384**	0.346	**0.389**	0.345	**0.392**
8	0.209	**0.288**	0.201	**0.297**	0.350	**0.379**	0.348	**0.380**	0.346	**0.392**	0.339	**0.391**
9	0.200	**0.292**	0.191	**0.298**	0.346	**0.373**	0.342	**0.376**	0.344	**0.386**	0.336	**0.393**
10	0.200	**0.294**	0.191	**0.297**	0.339	**0.370**	0.339	**0.373**	0.342	**0.377**	0.333	**0.389**
11	0.219	**0.284**	0.210	**0.289**	0.321	**0.359**	0.319	**0.372**	0.340	**0.370**	0.329	**0.386**
12	0.198	**0.270**	0.192	**0.254**	0.308	**0.351**	0.307	**0.358**	0.329	**0.353**	0.326	**0.368**
Mean	0.212	**0.274**	0.206	**0.270**	0.334	**0.376**	0.333	**0.378**	0.336	**0.378**	0.330	**0.382**

Table 6: **Layer-wise brain alignment (Pearson r) across models and conditions**. PT-La: pretrained ($\mathbf{W}_{\mathbf{lang}}$); ET-La: ECoG-tuned ($\mathbf{W}_{\mathbf{lang}}$); PT-Sp: pretrained ($\mathbf{W}_{\mathbf{speech}}$); ET-Sp: ECoG-tuned ($\mathbf{W}_{\mathbf{speech}}$). Bold indicates the higher value between pretrained and ECoG-tuned conditions.

Layer	Whisper-small				Wav2Vec 2.0-base				HuBERT-base			
	TM-La	ET-La	TM-Sp	ET-Sp	TM-La	ET-La	TM-Sp	ET-Sp	TM-La	ET-La	TM-Sp	ET-Sp
1	−6.5%	**14.3%**	−3.2%	**11.4%**	13.6%	**21.8%**	14.6%	**23.0%**	11.1%	**19.7%**	12.7%	**23.1%**
2	−8.1%	**15.1%**	−2.9%	**12.9%**	11.7%	**19.0%**	12.3%	**19.7%**	9.9%	**18.8%**	12.2%	**19.4%**
3	−0.4%	**21.3%**	1.0%	**17.2%**	11.1%	**19.7%**	10.3%	**19.0%**	8.4%	**16.9%**	9.4%	**17.4%**
4	5.9%	**23.3%**	4.8%	**20.2%**	8.6%	**17.3%**	9.2%	**15.8%**	9.4%	**17.8%**	11.2%	**19.4%**
5	7.1%	**24.9%**	11.0%	**25.0%**	7.5%	**17.0%**	8.8%	**15.9%**	6.9%	**15.2%**	7.2%	**18.1%**
6	13.4%	**30.6%**	16.6%	**34.0%**	6.0%	**14.8%**	8.7%	**15.1%**	7.3%	**15.3%**	8.7%	**18.6%**
7	15.9%	**31.2%**	19.5%	**35.7%**	4.9%	**13.2%**	9.2%	**14.1%**	7.6%	**14.8%**	8.5%	**17.4%**
8	24.7%	**37.1%**	25.6%	**47.4%**	4.4%	**10.6%**	7.1%	**12.4%**	6.7%	**15.7%**	9.0%	**18.5%**
9	32.2%	**45.3%**	33.8%	**55.0%**	3.5%	**10.3%**	8.5%	**12.6%**	9.1%	**13.4%**	10.6%	**19.4%**
10	32.7%	**45.9%**	35.6%	**54.1%**	4.3%	**11.5%**	6.5%	**12.8%**	8.3%	**11.8%**	11.2%	**19.3%**
11	18.2%	**29.5%**	16.3%	**36.7%**	4.5%	**13.9%**	7.6%	**18.4%**	5.7%	**10.5%**	11.5%	**19.8%**
12	18.2%	**35.3%**	20.0%	**30.7%**	5.4%	**16.5%**	6.6%	**19.0%**	6.2%	**7.9%**	8.1%	**15.9%**
Mean	12.8%	**29.5%**	14.9%	**31.7%**	7.1%	**15.5%**	9.1%	**16.5%**	8.0%	**14.8%**	10.0%	**18.9%**

Table 7: **Layer-wise improvement ratio relative to pretrained baseline**. Improvement ratios are computed per participant and averaged across participants. TM-La: Temporal-Mean ($\mathbf{W}_{\mathbf{lang}}$); TM-Sp: Temporal-Mean ($\mathbf{W}_{\mathbf{speech}}$). Bold indicates the higher value between TM and ET conditions. ET consistently outperforms TM at every layer.

The pattern of task performance provides suggestive evidence that temporal targeting may selectively enhance task-relevant representations: $\mathbf{W}_{\mathbf{lang}}$ tuning tends to yield relatively stronger performance on phonetic sentence type prediction—the most linguistically demanding task—while $\mathbf{W}_{\mathbf{speech}}$ tuning shows advantages on phonemes prediction and emotion recognition, which rely more on acoustic and prosodic features. Future work with larger datasets will help clarify this relationship.

Metric	Window	Whisper	Wav2Vec 2.0	HuBERT
Emotion	$\mathbf{W}_{\mathbf{lang}}$	0.444	0.502	0.495
	$\mathbf{W}_{\mathbf{speech}}$	**0.458**	**0.510**	**0.499**
Phoneme	$\mathbf{W}_{\mathbf{lang}}$	0.399	0.462	0.466
	$\mathbf{W}_{\mathbf{speech}}$	**0.403**	**0.465**	**0.467**
Sentence Type	$\mathbf{W}_{\mathbf{lang}}$	**0.264**	**0.514**	**0.485**
	$\mathbf{W}_{\mathbf{speech}}$	0.251	0.512	0.481

Table 8: **Comparison of $\mathbf{W}_{\mathbf{lang}}$ vs $\mathbf{W}_{\mathbf{speech}}$ tuning**. Bold indicates higher value within each model.

D.7 Cross-Participant Validation

To examine whether ECoG-tuning benefits generalize beyond per-participant optimization, we conduct a cross-participant validation using Whisper. We pool data from all nine participants by concatenating electrodes into a unified target space ($N = 1,268$ electrodes) and train the encoder with a shared projection head. Figure 8 presents the results across both temporal windows. Cross-

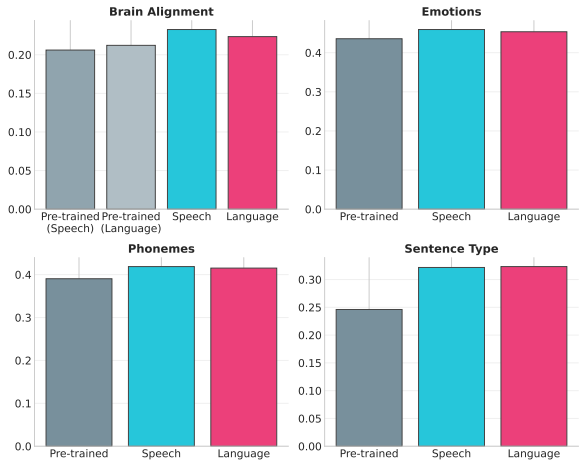

Figure 8: **Cross-participant ECoG-tuning results**.

participant ECoG-tuning consistently improves over pretrained baselines on all evaluation metrics.

Brain alignment improves by 5–13% relative to pretrained models. Downstream performance shows consistent gains of 4–7% in phoneme and emotion recognition, with particularly strong improvements in sentence type prediction (+31%), which requires higher-order linguistic understanding. These results suggest that ECoG-tuning benefits are not limited to per-participant optimization and can transfer across training configurations.

D.8 Comparison with fMRI-based Research

Direct comparison between modalities is inherently challenging due to differences in datasets, participants, and experimental paradigms. The following analysis is intended to contextualize rather than rank the two approaches.

D.8.1 Dataset Characteristics

We compare dataset characteristics with the established fMRI-based paradigm (Moussa et al., 2025). Table 9 summarizes the properties of the two datasets.

Property	fMRI	ECoG
Signal modality	BOLD	High-γ power
Participants	8	9
Stimulus material	27 stories	1 podcast
Total data duration	$\sim$51.2 h	$\sim$4.5 h
Temporal resolution	$\sim$2,000 ms	$\sim$2 ms
Sample granularity	TR-level	Word-level
Sample efficiency	$\sim$30/min	$\sim$170/min
Spatial coverage	30k–50k voxels	1,268 electrodes

Table 9: **Comparison of neural recording datasets for brain-tuning.** fMRI data from LeBel et al. (2023); ECoG data from the Podcast dataset (Zada et al., 2025).

The two modalities offer complementary strengths for neural supervision. ECoG provides millisecond-level temporal resolution that captures rapid lexical processing dynamics temporally smoothed by hemodynamic responses in fMRI. This precision enables word-level sample construction ($\sim$170 samples per minute), well-suited for investigating sub-second encoding processes.

However, ECoG acquisition depends on clinical opportunities from invasive monitoring; to our knowledge, the Podcast dataset is currently the only publicly available resource suitable for ECoG-tuning. While fMRI also requires substantial infrastructure, its non-invasive nature permits broader recruitment and extended sessions, facilitating larger-scale studies. The fMRI dataset spans 27 narratives with varied linguistic structures, whereas the Podcast offers more limited diversity. These data constraints shape training dynamics: our ECoG-tuning employs early stopping, terminating at 17.4 ± 4.0 epochs—typically before reaching default limits (30 epochs) in fMRI-based approaches.

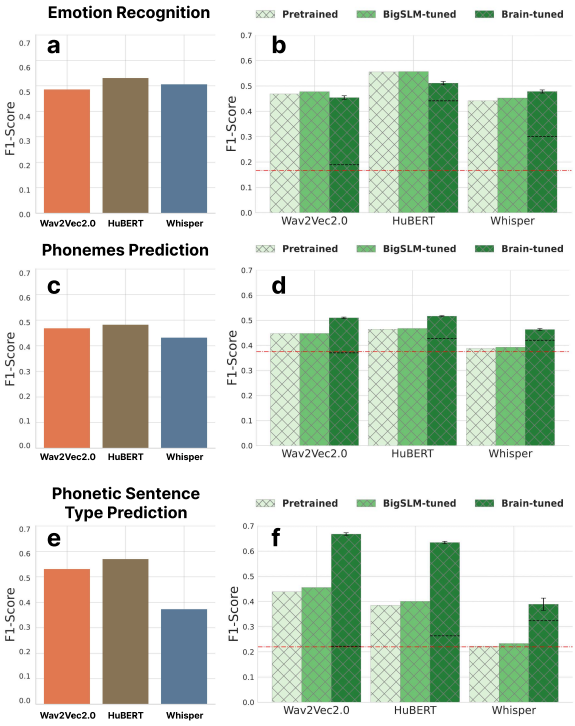

Figure 9: **Comparison of downstream task performance.** (a, c, e) ECoG-tuning results. (b, d, f) fMRI-based brain-tuning results (reproduced from Moussa et al. (2025)). Axes are scaled for direct comparison.

D.8.2 Downstream Task Performance

We compare downstream performance between fMRI-based brain-tuning and ECoG-tuning from the best-performing participant (Figure 9).

On emotion recognition, the best ECoG-tuned model achieves higher performance than the fMRI-based result, though direct comparisons are limited by differences in datasets and experimental conditions. This pattern is consistent with ECoG's capacity to capture prosodic and affective dynamics on sub-second timescales. On phonetic sentence type prediction, ECoG-tuning shows smaller gains, consistent with the limited syntactic diversity of a single podcast compared to 27 narratives in the fMRI dataset. Notably, these results are achieved with ∼4.5 hours of total recording time, suggesting that expanded datasets may further enhance efficacy. Combining fMRI's spatial coverage with ECoG's temporal precision is a promising direction for future work.

E Data Availability and Ethics

E.1 Data Access

The Podcast ECoG dataset is publicly available on OpenNeuro (https://openneuro.org/datasets/ds005574) under the CC0 license, following BIDS-iEEG standards.

E.2 Ethical Considerations

All participants provided informed consent under IRB approval (NYU Langone Medical Center, protocol s14-02101). The dataset contains de-identified data with anonymized anatomical scans (processed with pydeface). Clinical diagnoses are included in the original dataset for transparency, but are not used in our analyses.

F Notation Reference

Table 10 summarizes key notation used throughout this paper.

Symbol	Description
Temporal	
t_w	Onset time of word w
t_s	Neural window start
t_n	Neural window endpoint
t_e	Audio context endpoint
$W(t_w, \tau)$	Time window function returning $[t_s, t_n]$
τ	Window type $\in$ {speech, lang}
$\mathbf{W}_{\text{lang}}$	Language window for ECoG-tuning
$\mathbf{W}_{\text{speech}}$	Speech window for ECoG-tuning
Neural	
$\mathbf{E}_w$	ECoG target tensor $\in \mathbb{R}^{N \times K}$
$\hat{\mathbf{E}}_w$	Predicted ECoG tensor
N	Number of electrodes
K	Number of time points (= 102)
r_n	Brain alignment for electrode n
B	Aggregate brain alignment
Model	
$\mathbf{A}_w$	Audio input for word w
$\mathbf{H}^{(l)}$	Hidden states from layer l
L	Number of encoder layers (= 12)
D	Hidden dimension (= 768)
M	Number of output tokens
$\mathbf{z}$	Aggregated representation $\in \mathbb{R}^{LD}$
Training	
$\mathcal{B}$	Training batch
$\mathcal{T}$	Test set
$\boldsymbol{\theta}_{\text{enc}}$	Encoder parameters
$\boldsymbol{\theta}_{\text{proj}}$	Projection head parameters
$\boldsymbol{\theta}_{\text{feat}}$	Feature extractor parameters (frozen)

Table 10: Notation reference.

G Model-Brain Alignment Research at ACL/EMNLP/NAACL (2024–2025)

G.1 Directly Related Work

Papers in Table 11 directly inform our methodology or findings and are cited in the main text.

G.2 Broader NLP–Neuro Research

Papers in Table 12 represent the broader research landscape connecting language models with human brain processing.

Venue	Reference	Key Contribution
ACL 2025 Findings	Yin et al. (2025b)	Introduces the Association dataset and shows that simulating associative memory during language model inference improves brain alignment.
ACL 2025 Findings	He et al. (2025a)	Reveals dissociation between LLM linguistic performance and competence using neurolinguistic paradigms, finding that form competence exceeds meaning competence.
ACL 2024	Oota et al. (2024)	Demonstrates that speech models' alignment with high-level language regions is primarily driven by low-level acoustic features, lacking brain-relevant semantics—a gap our ECoG-tuning approach addresses.
NAACL 2025	AlKhamissi et al. (2025a)	Identifies language-selective units in LLMs using neuroscience-inspired localization methods and establishes their **causal role** in downstream language task performance.
EMNLP 2025	AlKhamissi et al. (2025b)	Shows that brain alignment in LLMs primarily tracks **formal linguistic competence** (grammar, syntax, phonetics) rather than functional competence (world knowledge, reasoning).
EMNLP 2025	Srijith et al. (2025)	Investigates cross-modal alignment between text/speech representations and MEG brain activity, finding asymmetric knowledge transfer across modalities.
EMNLP 2024	Nakagi et al. (2024)	Demonstrates that LLMs outperform traditional language models in predicting brain activity for high-level narrative content using 8.3 hours of fMRI data.

Table 11: Directly related work on brain-language model alignment cited in the main text.

Venue	Reference	Key Contribution
ACL 2024	Gao et al. (2024)	Proposes a Composition Score based on Transformer feed-forward networks to quantify meaning composition, correlating with fMRI activity in multiple brain regions.
ACL 2024	Franzluebbers et al. (2024)	Uses LLM-enhanced dependency parsers with fMRI data to demonstrate that human syntactic parsing maintains multiple analyses simultaneously (multipath parsing).
EMNLP 2025	Yin et al. (2025a)	Identifies data leakage issues in cross-subject brain-to-text decoding and proposes correct data splitting criteria for the field.

Table 12: Broader NLP–neuroscience research at major venues (2024–2025).