Brain Translation: Multitask Sequence-to-Sequence Modeling for Intracortical Attempted Speech Decoding

Michał Olak[1], **Tommaso Boccato**[1], **Matteo Ferrante**[1]

[1]Tether Evo

Correspondence: michalrafalolak@gmail.com

Abstract

Decoding language from intracortical activity is a sequence transduction problem that stress-tests modern NLP methods: inputs are high-dimensional and nonstationary, alignment to linguistic units is unknown, and training data are orders of magnitude smaller than in standard translation or speech recognition. We study attempted-speech decoding from intracortical recordings as a neural-to-language translation task and ask what contextual sequence-to-sequence modeling contributes beyond the framewise connectionist temporal classification (CTC) pipelines that dominate this domain. We introduce a multitask Transformer encoder–decoder that jointly predicts phonemes, words (via a BART-initialized decoder), and auxiliary acoustic features, together with a day-conditioned calibration module (Neural Hammer & Scalpel, NHS) that combines a global affine transform with feature-wise linear modulation to handle session-level distribution shift. On the Willett et al. (2023a) dataset, our model achieves state-of-the-art phoneme error rate (14.3%); direct word decoding reaches 25.6% WER, falling to 19.4% with rescoring, however still trailing multi-stage systems with heavy language-model refinement. Ablations and held-out-day experiments quantify the contributions of multitask supervision and NHS and the cost of session drift, while attention analyses reveal recurring temporal chunking that phoneme and word decoders consume differently. Together, these results position seq2seq modeling as both a competitive decoder and an interpretable probe for how language evidence is organized in neural inputs.

1 Introduction

Decoding language from neural activity is a sequence transduction problem with several properties that stress-test modern NLP methods. Inputs are high-dimensional, continuous, and nonstationary across recording sessions; alignment between neural activity and linguistic units is unknown; and training data are orders of magnitude smaller than in standard machine translation or automatic speech recognition, typically a few thousand sentence pairs from a single participant. We study attempted-speech decoding from intracortical recordings as a neural-to-language translation task and ask what fully contextual sequence-to-sequence (seq2seq) modeling contributes in this regime.

Beyond its methodological interest, this problem has direct clinical motivation: speech brain–computer interfaces (BCIs) aim to restore communication for individuals with conditions such as amyotrophic lateral sclerosis or locked-in syndrome. Recent initial demonstrations of real-time speech neuroprostheses predominantly use recurrent networks trained with a connectionist temporal classification (CTC) objective (Moses et al., 2021; Willett et al., 2023b; Card et al., 2024). This hybrid paradigm has three properties that may cap performance. First, per-frame phoneme distributions are produced under a conditional-independence assumption, without conditioning on previously generated outputs. Second, locality is built in: each prediction is based on a narrow temporal window (80 ms in Willett et al., 2023b), potentially discarding mid-range cues such as slow articulatory dynamics. Third, the neural decoder is trained separately from the downstream language model, so reported word error rates (WER) reflect both components and recent gains are driven largely by stronger language-model refinement (Li et al., 2025). By contrast, phoneme error rate (PER) more directly measures how faithfully neural activity is mapped onto sublexical structure, isolating the neural decoder from the downstream language model. This matters for recordings from speech motor cortex, where population activity is organised around articulatory-gestural rather than lexical-semantic structure: activity is organised by pho-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 43901–43916
October 24-29, 2026 ©2026 Association for Computational Linguistics

netic/articulatory features (Bouchard et al., 2013) and encodes directly-measured articulator kinematics more strongly than higher-level identity features (Conant et al., 2018). The phoneme is the sublexical unit most directly tied to this articulatory code, making PER the linguistic metric closest to the neural signal itself. Because phoneme-level fidelity also upper-bounds what any downstream language model can recover—and is most decisive for out-of-distribution material such as names or novel strings, where lexical priors help least—we adopt PER as our primary measure of neural decoding quality and report WER separately.

Seq2seq encoder–decoder models (Vaswani et al., 2017) differ from CTC along two axes we expect to matter here. First, they relax CTC's conditional-independence assumption: the decoder generates autoregressively, conditioning each output on previously generated tokens as well as the encoded input. Second, the decoder accumulates evidence across multiple encoded timesteps to emit each output token, rather than forcing a prediction at every timestep as framewise CTC does, giving it greater flexibility in how neural evidence is mapped onto output units. This mirrors the setup that drove gains in machine translation (Vaswani et al., 2017; Devlin et al., 2019) and automatic speech recognition (Dong et al., 2018; Karita et al., 2019), and we exploit the analogy directly—"translating" neural activity into phoneme and word sequences. A contextual formulation is also motivated by the neural data itself: articulatory encoding in speech cortex is coarticulatory, with the representation of a phoneme depending on its neighbouring movements (Chartier et al., 2018). Despite this, seq2seq remains underexplored for speech BCIs: prior Transformer decoders have largely targeted closed vocabularies of ≤ 50 words (Komeiji et al., 2024; Makin et al., 2020), and seq2seq systems were absent among top entries in the Brain-to-Text '24 challenge (Willett et al., 2024). We instantiate this paradigm with a Transformer architecture, the standard attention-based realisation of contextual seq2seq. The attention mechanism also provides a qualitative window into how the model segments and aggregates neural evidence during decoding, allowing us to "decode the decoder" and generate hypotheses about the temporal organization of neural speech representations.

Our contributions are: **(i)** a multitask Transformer seq2seq architecture that jointly predicts phonemes, words (via a BART-initialized decoder (Lewis et al., 2020)), and auxiliary acoustic features, achieving state-of-the-art PER (14.3%) on the Willett et al. (2023a) dataset; **(ii)** the Neural Hammer & Scalpel (NHS) day-conditioned calibration module, combining a global affine transform with feature-wise linear modulation (Perez et al., 2018) to handle session-level distribution shift; **(iii)** an interpretability analysis showing that the encoder learns recurring temporal chunks that phoneme and word decoders consume at distinct timescales; and **(iv)** a power-law scaling analysis projecting error rates as a function of dataset size, contextualizing data requirements for future intracortical speech decoders.

2 Methods

2.1 Dataset and preprocessing

We evaluated all models on the intracortical speech BCI dataset released by Willett et al. (2023a). The dataset contains neural recordings from a participant with amyotrophic lateral sclerosis, implanted with four 64-channel Utah arrays in the speech motor cortex. To keep the modeling setting comparable to referenced baselines, we restricted our analyses to the 128 channels in area 6v (ventral premotor cortex), consistent with prior evidence that this region carries the strongest speech-decoding signal. Over 24 recording days, the participant attempted to produce 12,100 prompted sentences drawn from a large-vocabulary English language corpus. Each trial consisted of an instructed delay period (planning) followed by a go cue to initiate attempted speech; we used only neural activity after the go cue. The dataset provides two neural features per channel in 20 ms bins: multi-unit threshold crossings (spike counts) and high-frequency spike-band power (> 250 Hz). Each feature was z-scored within the recording block to reduce session-specific offsets. We used the standard train (8800 trials) and validation (880 trials) partitions, omitting the competition holdout.

Targets were extracted at three levels: (1) 39-symbol ARPAbet phoneme sequences, (2) word-level targets tokenized via the BART tokenizer (Wolf et al., 2020), and (3) 14-dimensional Mel-frequency cepstral coefficients (MFCCs). The dataset includes time-synchronized microphone audio, but it is unintelligible due to the participant's condition; it nevertheless preserves coarse onset and envelope information. We therefore used MFCCs as auxiliary acoustic supervision to reg-

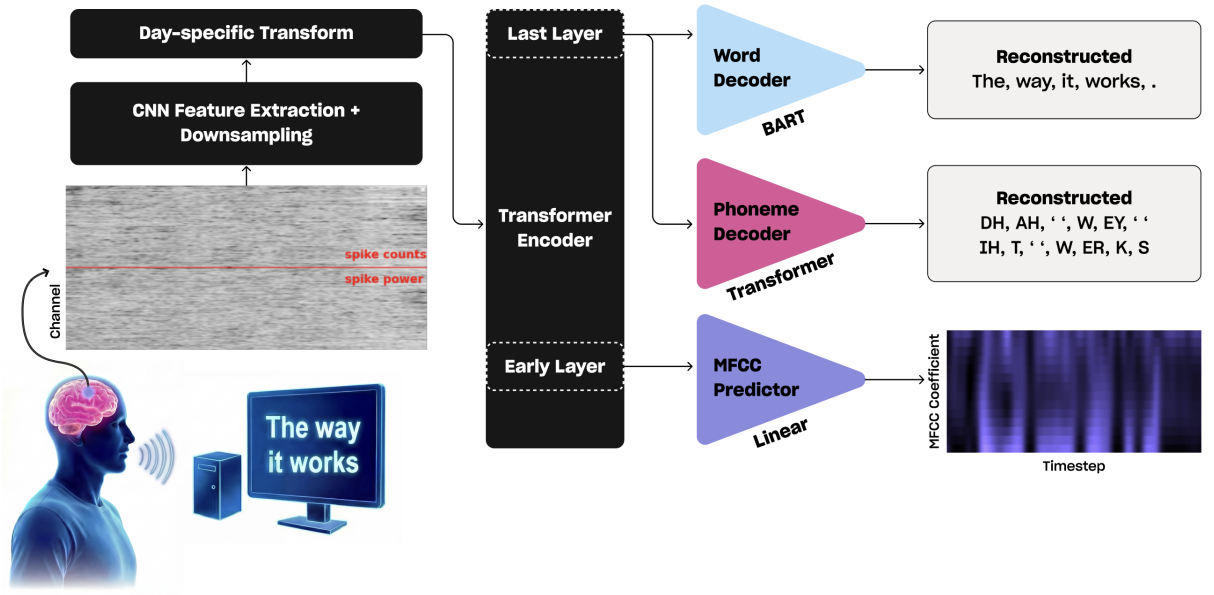


Figure 1: **Model overview.** Spike counts and spike-band power are processed in parallel convolutional branches, merged into a shared latent sequence, and reweighted by a lightweight content-dependent gate. The *NHS* day-specific transform then aligns the sequence by combining a global affine transformation (*hammer*) with FiLM-style feature-wise modulation (*scalpel*) through a learned gate. The calibrated representation is encoded by a 6-layer Transformer and decoded by three heads: an autoregressive phoneme decoder, a BART-based word decoder, and an auxiliary MFCC prediction head attached to an intermediate encoder layer.

ularize intermediate neural representations rather than as a decoding target.

2.2 Model overview

Our model is a multitask sequence-to-sequence architecture that maps intracortical activity to phoneme and word sequences while also predicting auxiliary acoustic features (Figure 1). It consists of four components: (1) a convolutional front end for feature extraction and temporal downsampling, (2) a day-specific calibration module, (3) a Transformer encoder, and (4) task-specific output heads.

Let the neural input be $X_0 \in \mathbb{R}^{T \times C}$, where $C = 256$ corresponds to 128 channels with two features per channel. Spike counts and spike-band power are first processed in separate one-dimensional convolutional branches and merged into a shared latent sequence. Temporal downsampling is performed by stride-4 convolutions, producing a representation at an effective 80 ms cadence:

$$X_{\text{feat}} \in \mathbb{R}^{L \times D},$$

where $L \approx T/4$ and $D = 512$. A lightweight content-dependent gate reweights latent channels frame by frame before calibration.

The calibrated representation is passed through a 6-layer Transformer encoder ($D_{model} = 512$, 8 attention heads, feedforward dimension 2048). Three output heads are attached at different depths. The *phoneme head* is an autoregressive Transformer decoder attached to the final encoder layer, trained with token-level cross-entropy against ARPAbet targets. The *word head* uses the decoder portion of `bart-base` (Lewis et al., 2020), also attached to the final encoder layer, with the first three layers frozen and the remaining layers fine-tuned for neural conditioning; it is trained with token-level cross-entropy against BART-tokenized word targets. The auxiliary *MFCC head* is a linear projection from an intermediate encoder layer (layer 2) trained with an L1 loss against the 14 MFCC coefficients. Loss weights for the active heads are equal across training. The overall training loss is therefore

$$\mathcal{L} = \mathcal{L}_{\text{phon}} + \lambda_{\text{mfcc}} \mathcal{L}_{\text{mfcc}} + \lambda_{\text{word}} \mathcal{L}_{\text{word}},$$

where λ_{mfcc} and λ_{word} are stage-dependent (see Section 2.6).

2.3 Day-specific calibration module

Chronic intracortical recordings exhibit substantial day-to-day nonstationarity. To address this within a known set of recording days, we introduce a day-specific calibration module termed Neural Hammer & Scalpel (NHS). The module operates on the convolutional feature sequence and is parameterized

by a learned day embedding, which means that *the recording day identity is assumed known at both training and inference time*. NHS therefore corrects for within-corpus session drift but does not, on its own, provide unseen-day generalization; we quantify the practical cost of this assumption in Section 3.1. NHS combines two complementary day-conditioned transformations: a global affine alignment that can remix latent features across dimensions (the *hammer*), which generalizes the linear day-specific transform of Willett et al. (2023b), and a FiLM-style feature-wise modulation (Perez et al., 2018) that provides milder per-feature adjustment (the *scalpel*). Let $X \in \mathbb{R}^{L \times D}$ denote the front-end output for one trial and let d index the recording day. The hammer branch applies

$$X_{\mathrm{h}} = XW_d + \mathbf{1}b_d^{\top},$$

where $W_d \in \mathbb{R}^{D \times D}$ and $b_d \in \mathbb{R}^D$ are learned day-specific parameters. The scalpel branch applies

$$X_{\mathrm{s}} = X \odot \gamma_d + \beta_d,$$

where $\gamma_d, \beta_d \in \mathbb{R}^D$ are produced from the day embedding through a small multilayer perceptron. The two branches are combined by a learned scalar gate,

$$\hat{X} = \phi(g_d X_{\mathrm{h}} + (1 - g_d)X_{\mathrm{s}}), \qquad g_d \in (0, 1),$$

where $\phi(\cdot)$ is a smooth nonlinearity. The design is motivated by two failure modes of chronic recordings: broad cross-session drift in the latent space, which calls for the higher-capacity affine branch, and more localized feature-specific variability, which is captured by the lighter-weight FiLM branch. We compare NHS against two baselines: a global affine day transform that matches the linear day transform used by Willett et al. (2023b), and a model without any day-specific transform. The affine baseline doubles as a scalpel ablation, isolating the contribution of the FiLM branch and the learned gate; results are reported in Section 3.1.

2.4 Controlled CTC baseline

To attribute differences to the decoding pipeline rather than to the front end or calibration, we constructed a controlled hybrid baseline built on the RNN-CTC backbone of Willett et al. (2023b). It shares our convolutional front end, NHS calibration module, MFCC auxiliary head (where applicable), and BART-based word head (where applicable), replacing only the Transformer encoder and autoregressive phoneme decoder with an RNN encoder trained with a CTC objective; its parameter count and training-time augmentations are matched to the reference implementation of Willett et al. (2023b). Encoder architecture and decoding objective change together, however, so the comparison does not on its own separate their individual contributions.

2.5 Word-level rescoring

To improve word-level decoding beyond greedy generation, we use HuggingFace's `generate()` interface to produce a hypothesis set ($N = 148$) per sentence (see Appx. C for the exact configuration). Candidates are rescored using three complementary signals: **(1) PER-score:** the phoneme error rate between the model's own phoneme prediction and the phoneme sequence obtained by applying a G2P converter (Park and Kim, 2019) to the candidate sentence; **(2) Phoneme-score:** the mean log-likelihood of the candidate's phoneme sequence under the model's autoregressive phoneme decoder in teacher-forcing mode, conditioned on the trial's neural encoder outputs; and **(3) LLM-score:** the sentence log-likelihood under an external language model (`Mistral-7B-v0.3`). A linear blend of these scores, with coefficients tuned on a held-out portion of the training set disjoint from the validation split used for model selection, selected the final prediction for each trial.

2.6 Training procedure

Models that include the word head are trained in two stages, an empirical choice we found gave the best phoneme- and word-level performance. In Stage 1, the encoder, phoneme decoder, and MFCC auxiliary head are trained jointly ($\lambda_{\mathrm{mfcc}} > 0$, $\lambda_{\mathrm{word}} = 0$). In Stage 2, the MFCC head is deactivated ($\lambda_{\mathrm{mfcc}} = 0$) and the BART word head is introduced ($\lambda_{\mathrm{word}} > 0$), with the encoder and phoneme decoder continuing to train jointly with the word decoder. We did not formally ablate joint training of all three objectives in a single stage. Phoneme-only and phoneme+MFCC variants are trained in a single stage.

Optimization used AdamW with learning-rate reduction on plateau based on validation phoneme error rate. Regularization included dropout, time masking, and channel masking. Model selection was based on validation PER. The same

train/validation partitions were used across ablations and baseline comparisons.

2.7 Evaluation metrics and reporting

We report phoneme error rate (PER) and word error rate (WER), both computed as normalized Levenshtein distance. PER was computed after removing sequence boundary tokens; WER was computed after lowercasing and removing punctuation. Direct decoding and rescored decoding are reported separately for word-level results. Unless otherwise stated, results are aggregated over five random seeds, reported as mean ± standard deviation.

3 Results

Table 1 summarizes decoding performance across published reference systems, controlled internal baselines, and day-transform ablations. The strongest result of our model is at the phoneme level. The best multitask seq2seq variant (Transformer + MFCC + BART) achieved a phoneme error rate of 14.3 ± 0.3%, improving over our controlled RNN–CTC baseline (17.4 ± 0.8%), the previously reported Willett et al. hybrid baseline (19.7%), and the recent context-aware diphoneme system of Li et al. (15.34%). These results indicate that fully contextual sequence-to-sequence decoding improves sublexical neural readout beyond both monophone-based framewise CTC decoding and diphoneme-enriched CTC pipelines. A concurrent foundation-model approach (Zhang et al., 2026) pairs a Transformer encoder with a CTC phoneme-decoding objective and reaches 14.4% PER on the same participant when the encoder is pretrained with self-supervised learning on ∼367 hours of cross-species (human and monkey) recordings, and 17.3% PER without pretraining.[1] Our single-participant model thus performs on par with their pretrained encoder despite using no large-scale pretraining, and ahead of from-scratch counterpart at a comparable data scale.

At the word level, the picture is different. Our direct seq2seq model reached 25.6 ± 0.2% WER, which improved to 19.4 ± 0.3% when candidate generation and rescoring were added. This substantially narrows the gap to the Willett et al. two-stage hybrid pipeline (17.4% WER), but remains above the strongest published multi-stage systems, including Li et al.'s DCoND-LIFT result (5.77%), which relies on heavier multistage language-model-assisted refinement and ensembling. The tradeoff is computational: on the full validation split (880 trials, 5,166 words), our direct Transformer + MFCC + BART model decodes in 46 s on a single H100, rising to 282 s with rescoring, compared to 562 s for the Willett-style WFST + LM pipeline. Overall, we interpret our main performance contribution as improved phoneme-level fidelity rather than state-leading final text accuracy, achieved within a lightweight, end-to-end pipeline closer to real-time deployment constraints.

Within our controlled ablations, multitask supervision improved performance consistently. Adding MFCC supervision lowered PER in both the recurrent and seq2seq model families, while adding the word-level decoder further improved phoneme decoding and enabled direct word generation. The full Transformer + MFCC + BART configuration provided the best performance without rescoring. We additionally compared pretrained against randomly initialized BART decoders (Appx. B); the two yielded similar performance, suggesting that in this low-resource regime the useful word-decoder structure is learned primarily from the supervised neural-decoding task rather than inherited from large-text pretraining.

3.1 Day-specific calibration and held-out-day generalization

Replacing the proposed NHS transform with either the global affine day transform of Willett et al. (2023b) (which corresponds to a hammer-only configuration of NHS, ablating the FiLM scalpel and the learned gate) or no day-specific transform degraded both PER and WER (Table 1). Relative to the full model, the affine transform increased error by 2.7 PER points and 2.6 WER points, while removing day-specific calibration altogether increased error by 3.6 PER points and 4.4 WER points.

To assess robustness to out-of-distribution neural data, we evaluated held-out-day generalization using models trained either on all 24 recording days or on the first 21 days only. As a reference, the 24-day model was first evaluated on days 22–24 using the correct day-specific transforms. We then reused the transform from the last training day (day 21) in two conditions: (i) the 24-day model re-evaluated

[1] PER figures for Zhang et al. (2026) are validation-set numbers from their Table 4; their word-level results are reported on the Brain-to-Text '24 competition holdout split rather than the validation partition we use, so WER figures are not directly comparable. Published numbers are taken as reported and are not seed-matched to our runs.

Table 1: Comparison of phoneme- and word-level decoding performance. Published reference systems are shown separately from controlled internal baselines and day-specific transform ablations. Our results are reported as mean ± SD across 5 random seeds. Published results are taken from the respective papers and are not directly seed-matched to our runs.

Model	**PER (%)**	**WER (%)**
Willett et al. (2023) RNN–CTC + LM	19.7	17.4
Li et al. (2025) DCoND-LIFT	15.34	**5.77**
RNN–CTC (ours)	17.4 ± 0.8	–
+ MFCC	17.0 ± 0.3	–
+ BART	21.0 ± 0.7	30.9 ± 0.8
+ MFCC + BART	17.0 ± 0.5	29.0 ± 0.4
Transformer Seq2Seq	14.8 ± 0.3	–
+ MFCC	14.6 ± 0.1	–
+ BART	14.5 ± 0.3	26.0 ± 0.4
+ MFCC + BART	**14.3 ± 0.3**	25.6 ± 0.2
+ MFCC + BART + rescoring	14.3 ± 0.3	19.4 ± 0.3
Transformer Seq2Seq + BART + Affine DT	17.2 ± 0.5	28.6 ± 0.4
Transformer Seq2Seq + BART + No DT	18.1 ± 0.6	30.4 ± 0.5

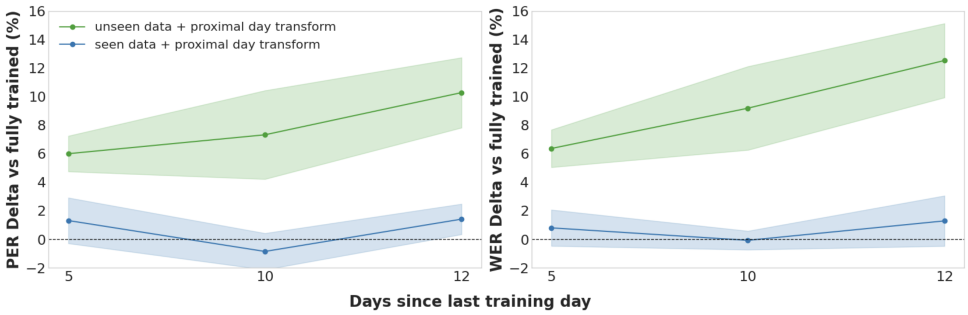


Figure 2: **Held-out-day generalization.** PER/WER deltas on held-out days 22–24 relative to a 24-day model evaluated with day-specific transforms (0%). Blue curves (*seen data + proximal*) show the same model re-evaluated on these days using the transform from day 21. Green curves (*unseen data + proximal*) show a 21-day model evaluated on days 22–24 with the day-21 transform, after subtracting the mean 21- vs. 24-day performance gap on days 1–21. Mean over 5 seeds, shaded regions denote 95% bootstrap confidence intervals.

on days 22–24 (*seen days + proximal transform*: tests transform substitution only); and (ii) the 21-day model evaluated on days 22–24 with the day-21 transform (*unseen days + proximal transform*: tests generalization to recording sessions never seen during training).

As shown in Figure 2, substituting the proximal transform on days included in training changed PER and WER by at most approximately 2 percentage points, indicating that neighboring-day NHS transforms are often similar. For unseen days, however, the performance penalty started at roughly 6 percentage points and increased to approximately 10–13 percentage points as the temporal gap grew. These results show that day-specific calibration substantially improves decoding, but does not fully remove across-day distribution shift. In its current form, NHS reduces rather than eliminates the calibration burden, leaving meaningful residual drift for the shared encoder–decoder to handle.

3.2 Attention structure and emergent temporal organization

Because the dataset does not provide explicit temporal alignments between neural activity and phoneme or word units, the seq2seq model must learn its own internal organization of time. We therefore analyzed encoder self-attention, decoder cross-attention, and decoder self-attention to examine how neural evidence is segmented, retrieved,

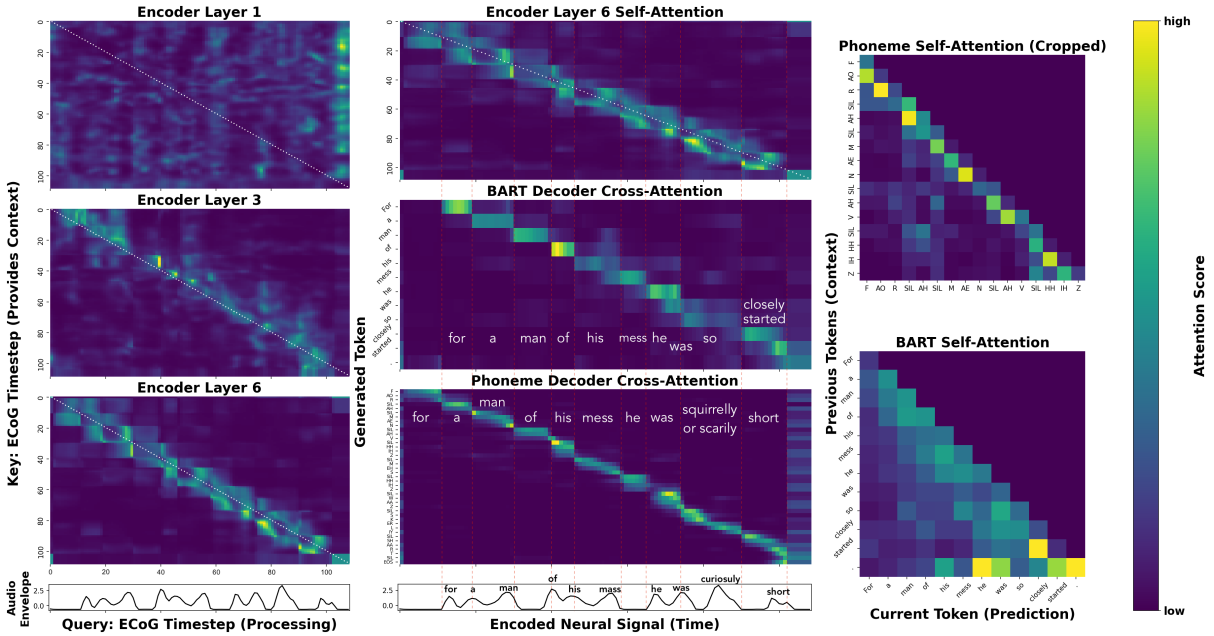


Figure 3: **Attention analysis.** Left: encoder self-attention for layers 1, 3, and 6. Middle: cross-attention for the BART and phoneme decoders. Right: decoder self-attention. Dotted vertical lines align with salient envelope events. Late encoder layers exhibit clean diagonals and "boxed" segments that decoders repeatedly attend to; phoneme self-attention is local, while BART self-attention is broadly triangular. The trial text is *"for a man of his mass he was curiously short"*; target words are approximately aligned to envelope features, as no ground-truth timing annotations are available in this dataset.

and accumulated during decoding. Figure 3 shows a representative trial from the validation set.

In the encoder, early layers distribute attention broadly, whereas later layers develop clearer local structure. By layer 3, diagonal patterns and compact local clusters begin to emerge. By layer 6, these patterns often take the form of short boxed segments separated by sharper transitions, suggesting that the encoder organizes the trial into temporally localized chunks. These boundaries qualitatively track salient events in the neural envelope, such as onsets and offsets, although we emphasize that the dataset lacks precise speech-unit timing annotations.

The two decoders re-use these encoder-derived segments in different ways. The phoneme decoder tends to focus locally within a segment and often progresses through a short temporal cascade, consistent with finer-grained sublexical resolution. The word decoder usually concentrates its evidence more coarsely, often over one dominant segment or a small number of adjacent segments, consistent with broader evidence aggregation. Decoder self-attention mirrors these roles, with the phoneme decoder remaining highly local and the BART decoder exhibiting a broader triangular pattern consistent with longer-context conditioning. We also observed a mild temporal asymmetry: phoneme cross-attention often peaks slightly before the envelope peak, suggesting anticipatory behaviour, whereas word cross-attention tends to lag behind it, consistent with evidence accumulation.

These observations are qualitative and should not be interpreted as a definitive mechanistic account. Rather, they provide a plausible picture of how the model often segments time and allocates evidence during decoding. Although the analysis is qualitative, manual inspection of all 880 validation trials indicated that similar temporal chunking patterns recur throughout the validation set, suggesting that the example in Fig. 3 is representative rather than exceptional. Additional examples are provided in Appx. D: Fig. 5 shows encoder-decoder chunking and segment reuse across 20 randomly sampled trials, and Fig. 6 shows related encoder-derived chunking patterns across 4 multitask variants and 15 random trials.

3.3 Scaling with dataset size

Finally, we examined whether classical scaling-law behavior also appears in intracortical speech decoding. Using day-stratified subsampling of the 8,800 training trials at fractions

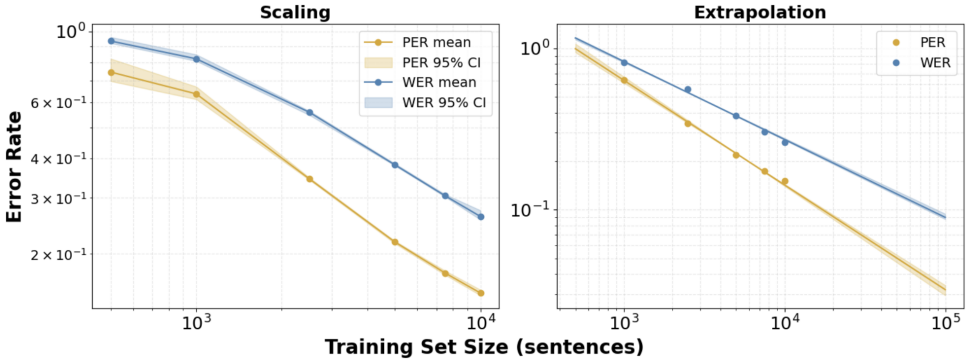


Figure 4: **Scaling behavior and power-law extrapolation. Left:** Phoneme (PER) and word (WER) error rates for six dataset fractions $\{0.05,\ 0.10,\ 0.25,\ 0.50,\ 0.75,\ 1.00\}$, plotted as corresponding training-set sizes, with 95% bootstrap confidence intervals across seeds. **Right:** Power-law models fitted to the multi-seed mean curve (excluding the 0.05 fraction) with bootstrap 95% confidence bands and extrapolated performance up to 100,000 training trials.

$\{0.05, 0.10, 0.25, 0.50, 0.75, 1.00\}$, we retrained the best-performing seq2seq variant (Transformer + MFCC + BART, without rescoring) with 5 seeds per fraction while holding other training factors fixed. Figure 4 shows that both PER and WER follow an approximately linear trend in log–log space, consistent with scaling behavior reported in other machine learning domains (Hestness et al., 2017; Kaplan et al., 2020).

We fitted a power law of the form $e(N) = aN^b$ to the multi-seed mean curve and used bootstrap resampling to estimate parameter uncertainty. Because the 0.05 fraction appears to lie in a pre-asymptotic regime, the extrapolation shown in Figure 4 excludes this point. Under that fit, increasing the dataset from approximately 10k to 20k sentences is projected to reduce PER from 14.3% to 8.4% and WER from 25.6% to 18.3%. A 100k-sentence dataset is further projected to reach approximately 2.9% PER and 8.4% WER.

These extrapolations should be interpreted cautiously. Power-law fits assume that the data distribution remains sufficiently stable as dataset size grows, whereas chronic intracortical recordings are subject to electrode drift, recalibration, and other nonstationarities. We therefore view these projections as illustrative estimates under relatively stable recording conditions rather than as guaranteed practical outcomes.

4 Discussion

Our results show that fully contextual seq2seq decoding is valuable not only as a speech decoder but also as an analysis framework: it improves phoneme-level readout from intracortical activity, exposes the importance of session-specific calibration, and makes it possible to qualitatively "decode the decoder" through attention analyses of how neural evidence is segmented and reused over time.

A central takeaway is that PER and WER capture different aspects of performance and should be interpreted as complementary rather than interchangeable. WER reflects both the neural decoder's accuracy and the strength of downstream linguistic correction, while PER provides a more direct view of the neural-to-speech mapping. This distinction is particularly relevant for attempted-speech decoding from ventral premotor cortex, where the dominant signal is expected to reflect articulatory and motor-planning variables rather than high-level semantic content (Bouchard et al., 2013). This picture is borne out in the model's errors (Appendix A): at the median and tail of the WER distribution, mistakes are dominated by high-acoustic-similarity phonetic confusions (e.g. way→why, fifty→fifteen, cases→taxes) that preserve utterance rhythm while shifting lexical identity. The decoder thus tends to recover phonetically faithful sequences that nonetheless resolve to the wrong words—the signature of strong sublexical readout (low PER) paired with limited lexical disambiguation (higher WER). Viewed through this lens, the main strength of our model is improved sublexical fidelity: by conditioning each output on the full neural input and the generated history, the contextual seq2seq decoding yields a lower PER than both monophone CTC baselines (Willett et al., 2023b) and the diphoneme-enriched CTC pipeline of Li et al. (2025). The strongest reported WERs, by contrast, are obtained by heavier multi-stage systems with strong post hoc linguistic refinement, which can obscure whether gains arise

from better neural readout or from language correction. Because those two axes are complementary, they compose rather than trade off: a stronger neural-to-phoneme mapping is a better substrate for any downstream language-model stage layered on top of it. This is most consequential for open-vocabulary or out-of-distribution material such as names and novel strings, where lexical priors offer little help and accuracy is limited by the fidelity of the sublexical readout itself — there, improving PER raises the effective ceiling on achievable text accuracy, whereas heavier LM refinement is an orthogonal, stackable gain concentrated on in-vocabulary text. External results are consistent with this: systems that pair a similar-quality neural readout with heavier LM-based decoding and ensembling achieve substantially lower WER (Zhang et al., 2026).

Our architectural ablations highlight that representation shaping matters even more than head choice. Auxiliary MFCC supervision improved PER in both recurrent and seq2seq families, indicating that coarse acoustic structure can regularize early neural representations. Adding the word-level decoder further improved phoneme decoding, suggesting that higher-level linguistic supervision can beneficially shape sublexical representations. In contrast, pretrained BART initialization provided only modest benefit (Appx. B), implying that much of the useful structure is learned from the supervised neural-decoding task itself rather than inherited from large-text pretraining—a noteworthy point for low-resource sequence transduction more broadly.

Session variability emerged as a major practical factor. NHS outperformed both the affine day transform (Willett et al., 2023b) and no day-specific transform, indicating that effective calibration requires more flexibility than global alignment alone. However, held-out-day experiments showed that reusing a proximal-day transform still incurs nontrivial penalties that grow with temporal distance, so NHS reduces but does not eliminate the calibration burden.

Finally, the attention analyses provide a qualitative window onto how the model solves the decoding problem. Encoder layers repeatedly developed temporally localized chunks, which the phoneme and word decoders reused at different scales: the phoneme decoder retrieved evidence locally, while the word decoder aggregated more broadly across a small number of segments and its own output history. Consistent with this local, chunk-wise structure, we found no evidence that the model exploits long-range dependencies over the neural input; its gains are concentrated at the sublexical (phoneme) level rather than in long-range linguistic modelling. We read these observations as a structured hierarchy of evidence aggregation rather than a definitive mechanistic account, and the locality of the learned solution in turn motivates streaming or limited-lookahead seq2seq variants that exploit the chunked structure.

Overall, our results position seq2seq modeling as both a competitive decoder and an interpretable probe for how language-relevant information is organized in noisy, low-resource neural inputs, while identifying session variability as the most consequential remaining barrier to robust deployment.

Limitations

Our study has several limitations. First, all experiments were performed on data from a single participant and a single cortical region (area 6v), and decoding was evaluated offline rather than in a streaming setting; we therefore cannot assess how the method generalizes across participants, array locations, or real-time constraints. Second, the NHS day-specific calibration module assumes that recording day identity is known at inference time, which limits direct deployment realism: replacing explicit day labels with inferred session embeddings, calibration from limited adaptation data, or continuous test-time adaptation are natural next steps that we did not evaluate here. Third, while the held-out-day experiment characterizes the practical limit of proximal-transform substitution, we did not run finer-grained NHS sub-ablations (e.g., scalpel-only or no-gate variants), so the relative contributions of the FiLM branch and the learned gate beyond the affine "hammer" baseline remain partially confounded. Fourth, although the proposed model achieves state-of-the-art PER, its best WER remains above that of the strongest hybrid and LLM-assisted pipelines, and we do not attempt heavy multi-stage refinement. Fifth, the attention analyses are qualitative; chunking patterns recurred broadly across the validation set on manual inspection, but we did not quantify them with entropy, segment-consistency, or diagonal-mass metrics. Sixth, our controlled RNN–CTC baseline reproduces the field-standard pipeline under a matched front end, calibration, and auxiliary heads, establishing that our

improvement over framewise CTC reflects the contextual seq2seq paradigm as a whole rather than differences in preprocessing. It does not isolate the architectural (Transformer vs. RNN) from the objective (seq2seq vs. CTC) contribution; we therefore attribute our gains to the paradigm as deployed, not to self-attention or the autoregressive objective individually. Decomposing these effects would require a factorial architecture$\times$objective comparison (Transformer+CTC, RNN+seq2seq), which our controlled design does not attempt. Finally, the power-law scaling extrapolations assume that data distribution remains sufficiently stable as dataset size grows, whereas chronic intracortical recordings are subject to electrode drift and recalibration; we view these projections as illustrative under relatively stable recording conditions rather than as guarantees. We also characterized scaling only for our best seq2seq configuration; how the seq2seq and CTC paradigms scale relative to one another—whether the seq2seq advantage widens, narrows, or persists with more data—is an informative direction we leave to future work.

Ethical Considerations

This study uses the publicly available intracortical dataset of Willett et al. (2023a); no new human data were collected. The original dataset was gathered under IRB oversight with informed consent, and all analyses were performed on de-identified recordings. We complied with the dataset license and did not attempt re-identification or linkage with external sources.

Speech decoding from neural signals raises distinctive ethical considerations. First, decoding errors can misrepresent user intent—particularly for out-of-distribution phrases such as names or novel strings—motivating interfaces that expose calibrated uncertainty, enable rapid user correction, and provide conservative fallbacks. Second, because neural signals are sensitive, deployments should support intention-based operation (e.g., go signals, reliable "off" states) and ensure decoding occurs only with explicit consent and intentional engagement. Third, reducing reliance on cloud-based computation is desirable for privacy: lightweight end-to-end decoding pipelines like the one studied here provide a practical foundation for on-device, privacy-preserving operation, even when higher-accuracy multi-stage refinement remains useful in offline settings. Our models are intended to advance scientific understanding of neural speech decoding and are not suitable for clinical, legal, or surveillance use; any future deployment should include explicit consent, human oversight, and protections against non-consensual decoding.

References

Kristofer E. Bouchard, Nima Mesgarani, Keith Johnson, and Edward F. Chang. 2013. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332.

Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. 2024. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618.

Josh Chartier, Gopala K. Anumanchipalli, Keith Johnson, and Edward F. Chang. 2018. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron*, 98(5):1042–1054.e4.

David F. Conant, Kristofer E. Bouchard, Matthew K. Leonard, and Edward F. Chang. 2018. Human sensorimotor cortex control of directly measured vocal tract movements during vowel production. *Journal of Neuroscience*, 38(12):2955–2966.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint*.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryo Yamamoto, Xiong Wang, and Shinji Watanabe. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Shuji Komeiji, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano, Koichi Shinoda, and Toshihisa Tanaka. 2024. Feasibility of decoding covert speech in ecog with a transformer trained on overt speech. *Scientific Reports*, 14(1).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Transactions of the Association for Computational Linguistics*, 8:807–828.

Jingyuan Li, Trung Le, Chaofei Fan, Mingfei Chen, and Eli Shlizerman. 2025. Brain-to-text decoding with context-aware neural representations and large language models. *Journal of Neural Engineering*, 22(5):056026.

Joseph G. Makin, David A. Moses, and Edward F. Chang. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4):575–582.

David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227.

Kyubyong Park and Jongseok Kim. 2019. g2pE: A simple python module for english grapheme-to-phoneme conversion. `https://github.com/Kyubyong/g2p`.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Francis R. Willett, David T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. 2023a. Data from: A high-performance speech neuroprosthesis. Dryad Digital Repository. Intracortical attempted-speech decodings (12,100 sentences) with RNN baseline and language model.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. 2023b. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

Francis R. Willett, Jingyuan Li, Trung Le, Chaofei Fan, Mingfei Chen, Eli Shlizerman, Yue Chen, Xin Zheng, Tatsuo S. Okubo, Tyler Benster, Hyun Dong Lee, Maxwell Kounga, E. Kelly Buchanan, David Zoltowski, Scott W. Linderman, and Jaimie M. Henderson. 2024. Brain-to-text benchmark '24: Lessons learned. *arXiv preprint*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Yizi Zhang, Linyang He, Chaofei Fan, Tingkai Liu, Han Yu, Trung Le, Jingyuan Li, Scott Linderman, Lea Duncker, Francis R. Willett, Nima Mesgarani, and Liam Paninski. 2026. A cross-species neural foundation model for end-to-end speech decoding. In *International Conference on Learning Representations (ICLR)*.

Reproducibility Statement

We took multiple steps to ensure reproducibility. The dataset source and all preprocessing are documented in Section 2.1. The full model architecture, the NHS day transform, training objectives, training protocol, and evaluation metrics are specified in Section 2.2 and Section 2.6. Hyperparameters are listed in Appendix E. The scaling-law protocol (fractions and fitting) appears in Section 3.3. Attention extraction procedures are described in Appendix D. We report means and standard deviations over five seeds in Table 1 and provide exact splits (stratified by day) in our code release.

A Qualitative error analysis

Table 2 shows representative predictions of Transformer + MFCC + BART model (no rescoring) at the 10th, 50th (median), and 90th WER percentiles. Errors at the median and tail are dominated by phonetic confusions with high acoustic similarity (e.g., *way*→*why*, *fifty*→*fifteen*; *cases*→*taxes*), often preserving sentence rhythm but shifting semantics. This suggests that the neural signal supplies strong sub-lexical acoustic evidence while semantic disambiguation remains limited without explicit rescoring.

B BART initialization and training: Linguistic prior impact

We examined whether word-level linguistic priors encoded in the pretrained BART decoder influence decoding performance. In Stage 2, we compared three configurations: (i) initializing the BART decoder from the `bart-base` checkpoint while freezing its first three layers to preserve pretrained linguistic structure, (ii) initializing from `bart-base` and finetuning all layers, and (iii) replacing the decoder with an identical BART architecture initialized with random weights. In all conditions, the same pretrained BART tokenizer was used, so the subword vocabulary and segmentation were held fixed.

As shown in Table 3, all strategies achieve highly similar phoneme and word error rates. The fully finetuned pretrained model yields the best WER, but the difference relative to random initialization is small. Notably, the randomly initialized decoder performs on par with the pretrained variants, despite lacking any large-corpus language modeling priors in its weights. This suggests that the linguistic structure exploited during decoding arises primarily from patterns in the supervised dataset and from the encoder's neural representations, rather than from the pretrained BART language model itself.

Comparison to Table 1 shows that adding the word-level BART decoder—regardless of how its weights are initialized—reduces phoneme error rates relative to a phoneme-only baseline. This indicates that the encoder representations are shaped by word-level supervision even when the decoder itself carries no pretrained linguistic knowledge. Taken together, these results imply that our transformer encoder can learn higher-level, linguistically structured representations directly from neural activity and task data, with pretrained BART weights providing at most a modest additional benefit in this regime.

C Generation Strategies and Scoring Analysis

We employed two complementary decoding configurations to construct a rich hypothesis space for the rescoring stage:

- **Beam-20 (deterministic).** A standard beam search with `num_beams=20`, producing 20 high-likelihood candidates per trial. This method explores the most probable regions of the model's output distribution in a deterministic manner.
- **Nucleus-128 (stochastic).** A high-diversity sampling setup that generates 128 candidates using nucleus sampling (`top_p=0.95`, `temperature=1.0`, `top_k=0`). Unlike beam search, this strategy introduces stochasticity and explores a broader portion of the distribution, often producing phrasings that would not appear under purely probability-maximizing generation.

Table 4 summarizes the effect of different candidate selection approaches. The default greedy decoding provides a strong baseline, while beam search offers only a modest improvement when simply the highest overall-probability beam output is chosen. More substantial gains arise when candidates are rescored using phoneme-level evidence (PER-score and phoneme-score), especially compared to LLM-only scoring–highlighting the weakness of BART decoder to differentiate in the acoustic space as shown in A. The best overall performance is obtained when all scoring components

Table 2: Example word-level predictions of Transformer+MFCC+BART model (no rescoring) from 10th, 50th (median) and 90th WER percentiles.

WER	%ile	Target sentence	Predicted text
0.0	0	Do you know where it might have gone?	Do you know where it might have gone.
0.0	0	I am an artist, lost in my own vision.	I am an artist, lost in my own vision.
0.0	0	Read the decision below.	Read the decision below.
0.2	50	I don't think so anymore.	I don't think so many.
0.2	50	Just **way** in the back.	Just **why** in the back?
0.2	50	Sometimes they're not very **open**.	Sometimes they're not very **hard**.
0.6	90	**Fifty** nine **kilometers per** gallon.	**Fifteen** nine **automobiles for** gallon.
0.6	90	We were **promised civil liberties**.	We were **most several places**.
0.6	90	**Special rules** for employment **cases**.	**Paces** for employment **taxes**.

Table 3: Performance comparison between different strategies of BART initialization and training. Results reported as mean ± SD across 5 global seeds

Model	PER (%)	WER (%)
Transformer Seq2Seq + MFCC + BART		
BART pretrained + Freeze first 3 layers	14.3 ± 0.3	25.6 ± 0.2
BART pretrained + No freezing	14.6 ± 0.4	$\mathbf{25.2 \pm 0.1}$
BART random init + No freezing	$\mathbf{14.2 \pm 0.2}$	25.8 ± 0.4

Table 4: Word-level performance comparison between strategies of candidate selection. Score-based results are obtained by selection of candidates generated using beam search and stochastic sampling strategies. All results reported as mean ± SD across 5 global seeds

Model	WER (%)
Transformer Seq2Seq + MFCC + BART	
Greedy decoding	25.6 ± 0.2
Beam search (top result)	24.8 ± 0.4
PER score only	23.2 ± 0.4
Phoneme head score only	22.6 ± 0.2
LLM score only	27.2 ± 0.4
All scores combined	19.4 ± 0.3
Oracle	14.5 ± 0.2

are combined with ratio PhonemeHead:PER:LLM 9:4:5, lowering the final WER from 25.6% to 19.4% (delta 6.2 pp.). The word-level performance could theoretically be improved further as the oracle (always choose the best) WER is 14.5%.

D Attention extraction and further examples

To analyze model behavior, we extracted attention maps from the encoder and both decoders. The PyTorch TransformerEncoder and phoneme TransformerDecoder were modified setting nn.MultiheadAttention class attribute `return_weights=True` to return attention weights for each layer. For the word decoder, we used the Hugging Face BART implementation with `output_attentions=True` to visualize attention heads across tokens. These attention maps allow us to interpret which neural timesteps contribute to specific phoneme or word predictions, shedding light on model alignment and representation dynamics. Figures 5 and 6 provide further examples of the observed patterns among randomly sampled 20 and 15 validation trials respectively.

E Transformer model parameters

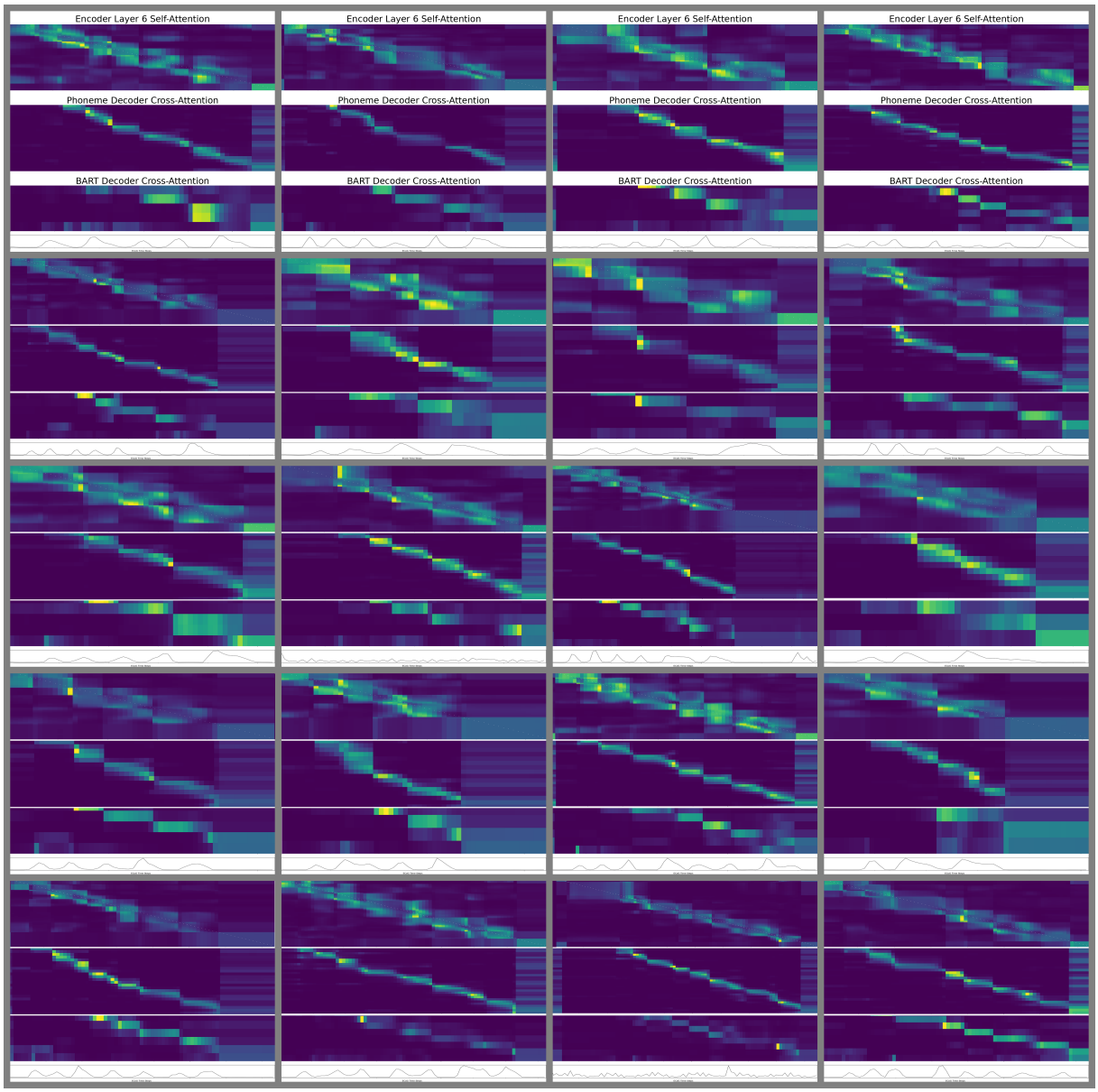


Figure 5: **Self and cross attention: Usage of temporal chunking.** Each cell shows vertically stacked 1) Encoder self-attention in the last layer, 2) Phoneme Decoder cross-attention, 3) BART word decoder cross-attention, 4) Audio envelope, obtained from Transformer Seq2Seq + MFCC + BART model, visualized for 20 randomly chosen validation trials. The pattern of temporal chunking of neural signal in the encoder and the downstream usage of those chunks by both decoders is prevalent across all samples - as well as the temporal lag of word decoder relative to the phoneme decoder, indicating evidence accumulation.

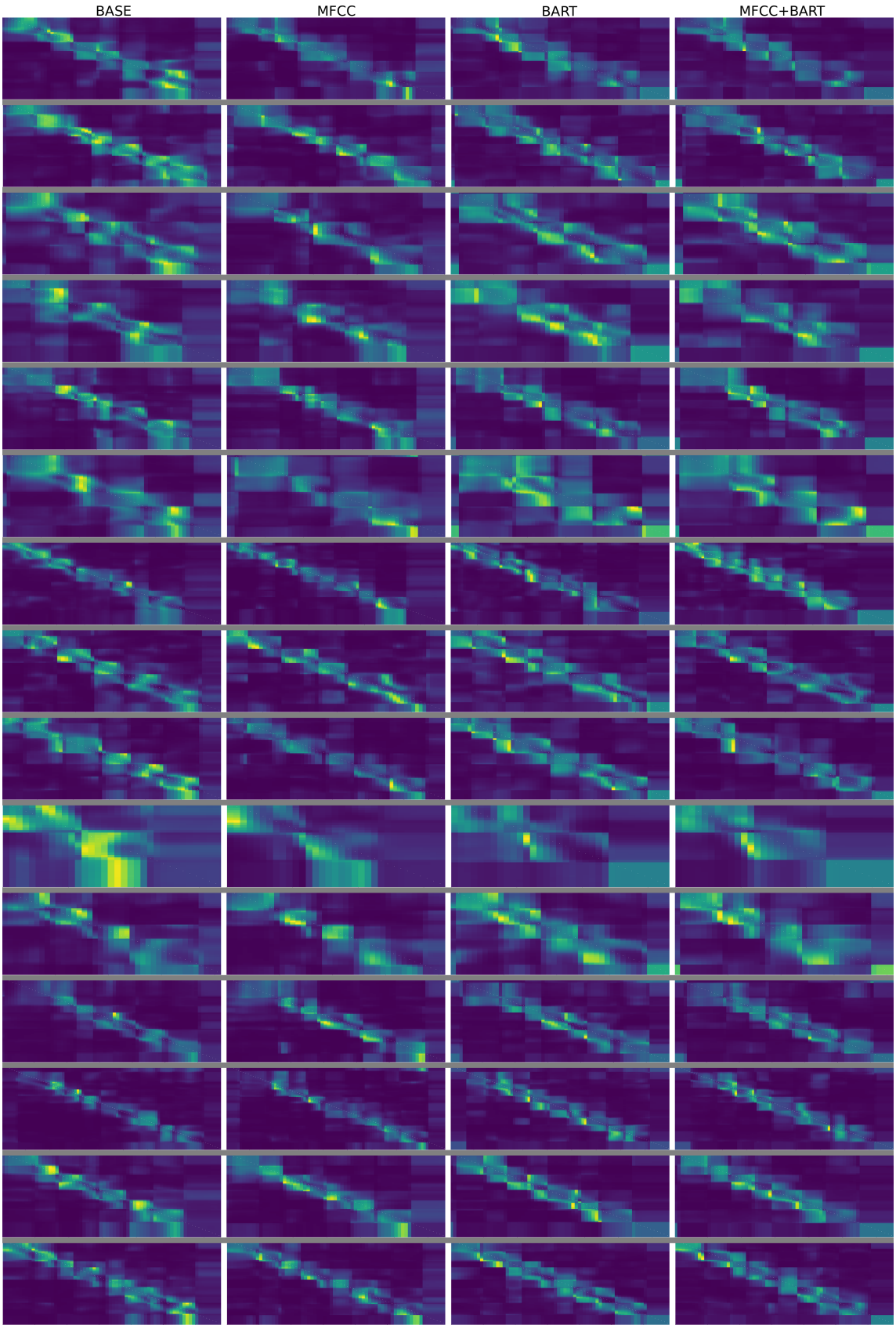


Figure 6: **Self-attention: Temporal chunking.** Self-attention map of the last (6th) layer of encoder across model variants (columns) visualized for random 15 trials (rows). The representation pattern of temporal chunking of the neural signal into "boxes" is prevalent across the trials regardless of the auxiliary objectives

Table 5: Training and model parameters for the two-stage training process of Transformer+MFCC+BART model. Stage 1 focuses on pre-training the ECoG encoder with an auxiliary MFCC prediction task alongside the main phoneme decoding task. Stage 2 introduces a BART-based word decoder head and fine-tunes the entire model jointly on phoneme and text decoding. The Stage 2 column only lists values in Multi-Task Setup section, as only then they differ from Stage 1.

Parameter	**Stage 1 Value**	**Stage 2 Value**
GPU	Nvidia H100 80GB of HBM3 memory	
	Core Transformer Architecture	
Model dimension (`d_model`)	512	
Attention heads (`n_head`)	8	
Encoder, Decoder layers	6, 6	
Feed-forward dimension	2048	
Activation function	GELU	
	Input Processing & Feature Extraction	
ECoG features	256 (128 channels x 2 features per-channel)	
Conv kernel size	5	
Downsampling factor	4	
Number of days	24 (all)	
	Regularization & Augmentation	
Dropout	0.4	
Time masking probability	0.3	
Max time mask length	25 steps	
Max time mask proportion	20%	
Channel masking probability	0.25	
Max channels masked	15 electrodes	
	Multi-Task Setup	
Training objective	Joint phoneme decoding & MFCC prediction	Joint phoneme & text decoding
Phoneme decoder loss weight	1.0	1.0
MFCC auxiliary task	Enabled	Disabled
MFCC head type	Linear	N/A
MFCC aux. loss weight	0.001	N/A
Aux. head input layer index	1	N/A
BART text decoder task	Disabled	Enabled
BART model type	N/A	`bart-base`
BART loss weight	N/A	1.0
BART freezing strategy	N/A	Freeze first 3 decoder layers
	Optimizer & Scheduler	
Optimizer	AdamW	
Learning rate	1e-4	
Weight decay	1e-3	
Scheduler	ReduceLROnPlateau	
	Training Details	
Batch size	4	
Gradient accumulation	5 steps	
Epochs	200	
Precision	32-bit float	