Unlocking Non-Invasive Brain-to-Text

Dulhan Jayalath Gilad Landau Oiwi Parker Jones

Abstract

Despite major advances in surgical brain-to-text (B2T), i.e. transcribing speech from invasive brain recordings, non-invasive alternatives have yet to surpass even chance on standard metrics. This remains a barrier to building a non-invasive brain-computer interface (BCI) capable of restoring communication in paralysed individuals without surgery. Here, we present the first non-invasive B2T result that significantly exceeds these critical baselines, **raising BLEU $1.4-2.6\times$** over prior work. This result is driven by three contributions: (1) we extend recent word-classification models with LLM-based rescoring, transforming single-word predictors into closed-vocabulary B2T systems; (2) we introduce a predictive in-filling approach to handle out-of-vocabulary (OOV) words, substantially expanding the effective vocabulary; and (3) we demonstrate, for the first time, how to scale non-invasive B2T models across datasets, unlocking deep learning at scale and **improving accuracy $2.1-2.3\times$**. Through these contributions, we offer new insights into the roles of data quality and vocabulary size. Together, our results remove a major obstacle to realising practical non-invasive B2T systems.

1. Introduction

Transcribing natural language text directly from speech-related neural signals, known as brain-to-text decoding, remains one of neuroscience's most challenging and clinically significant frontiers. Recent breakthroughs showing remarkable accuracy, e.g. Moses et al. (2021); Metzger et al. (2023); Willett et al. (2023); Card et al. (2024); Littlejohn et al. (2025), use surgical methods to measure brain activity, implanting electrodes on the surface of, or even inside, the brain. These neurosurgical interventions come with significant inherent risks, including brain infection, haemorrhages, and cognitive side effects, limiting adoption outside of research trials. Instead, safe non-invasive imaging methods, such as M/EEG, which measure neural activity with sensors on or around the head, could unlock widespread use. These methods trade surgical risk for lower signal quality as skull and tissue conductivity as well as sensor distance introduce attenuation of the underlying neural signals (Ball et al., 2009). Despite advances in some classification tasks (Jayalath et al., 2025; Défossez et al., 2023; d'Ascoli et al., 2024), full B2T reconstruction from non-invasive electrophysiological signals has remained elusive, in spite of numerous attempts (Wang & Ji, 2022; Duan et al., 2023; Yang et al., 2024b;c;a), with no method convincingly surpassing chance across a set of standard metrics (Jo et al., 2024).

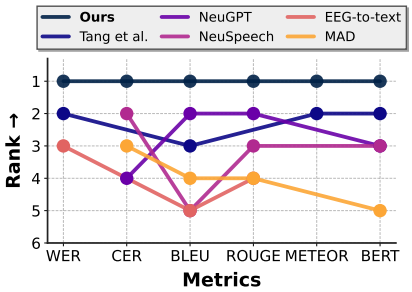

Figure 1: **Our approach outperforms all non-invasive B2T methods.** Ranks are calculated from absolute improvement over associated random baselines.

We present the first non-invasive B2T method to surpass every critical chance baseline across standard metrics ($p \ll .001$), comprehensively exceeding all prior work (Figure 1) and improving over previous BLEU scores by up to $2.6\times$.

In reaching this milestone, we address some fundamental challenges. Firstly, we extend individual word prediction (d'Ascoli et al., 2024) to contextual sequence decoding, rescoring word predictions according to their surrounding context using a large language model. This enables word predictors to become viable B2T decoders, overcoming the low performance of prior sequence-to-sequence speech decoding strategies (Yang et al., 2024b;c;a). Secondly, word predictors are bound to small closed vocabularies that are hard to scale up, limiting B2T performance. Hence, we propose an in-filling strategy to detect and predict out-of-vocabulary words, enabling open vocabulary transcription from closed-vocabulary decoders.

PNPL, University of Oxford . Correspondence to: Dulhan Jayalath <dulhan@robots.ox.ac.uk>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Although these techniques unlock non-invasive B2T, improving results further requires addressing a data bottleneck. Existing methods can not scale up data because individual speech datasets are small. Moreover, due to drastic differences across datasets, such as hardware, subject anatomy, and cognitive variation, they struggle to combine datasets to achieve meaningful performance improvements in complex tasks (e.g. (Jayalath et al., 2025; Gideoni et al., 2024; Ridge & Parker Jones, 2024; Zhang et al., 2023; Jiang et al., 2024; Wang et al., 2024)). We resolve this fundamental bottleneck by defining standalone dataset performance as a measure of quality. By observing that high quality datasets combine well with other datasets, we selectively pool data to more than double word classification accuracy.

In summary, our work makes three main contributions: (1) we advance non-invasive decoding from single-word prediction to full B2T through contextual LLM rescoring; (2) we transform closed-vocabulary B2T into open-vocabulary decoding with a predictive in-filling strategy; and (3) we enable scaling through our selective dataset pooling framework, improving word classification accuracy by up to $2.3\times$. These advances bring about the first system that surpasses the prerequisite chance baselines for establishing non-invasive B2T and dominates prior work. Additionally, our ablation studies rigorously validate each component of our approach against controls that invalidated prior methods, establishing a new standard for validating B2T systems. By demonstrating that non-invasive methods achieve meaningful results, our work challenges the assumption that speech decoding requires surgical imaging, *unlocking* the path for advances in the years to come.

2. Related Work

Speech decoding BCIs have progressed along divergent paths: surgical approaches with impressive performance but limited applicability, and non-invasive methods that are safe but must contend with lower signal to noise ratios. This tension between efficacy and accessibility defines the current landscape. In this section, we contextualise our non-invasive work within this rapidly evolving field.

Over the past few years, several breakthroughs have been made in surgical speech BCIs. In 2021, Moses et al. (2021) developed a B2T neuroprosthesis for a paralysed patient with a 50-word vocabulary. Two years later, Willett et al. (2023) showed similar performance but with a 125,000-word vocabulary. The main innovations of these papers were the recording of high-resolution brain signals and the use of deep neural networks. Following these milestones, subsequent papers have introduced rapid calibration to reduce degradation due to non-stationarities (Card et al., 2024) and real-time voice synthesis (Littlejohn et al., 2025). While these developments are remarkable, surgical approaches carry significant risk of complications and are consequently limited in use outside of controlled clinical trials.

Non-invasive approaches are safe but progress has been limited by noisier signal quality, constraining the most promising work to simpler tasks than B2T. Défossez et al. (2023) developed a contrastive learning method that matches pairs of audio and MEG data from speech perception. A preprint by d'Ascoli et al. (2024) extended this method to pairs of word embeddings and MEG data for word classification. While impressive, neither facilitate sequence-level brain-to-text as they either require paired audio (Défossez et al., 2023) or are bound to small vocabularies and limited by greedy word prediction (d'Ascoli et al., 2024).

Despite the difficulty, there have been attempts at non-invasive B2T. With fMRI, Tang et al. (2023) achieved significant results across three participants in speech perception. However, they are not able to decode at word-level granularity as the temporal resolution in fMRI is too low to resolve words, limiting it to a semantic paraphrasing method. With EEG, Wang & Ji (2022) and Duan et al. (2023) explored open-vocabulary B2T from participants reading text. Subsequent analysis by Jo et al. (2024) revealed that performance metrics in these studies were influenced by teacher-forcing during evaluation, and when addressed, performance was comparable to a baseline with random noise inputs. A series of recent MEG-based B2T preprints (Yang et al., 2024b;c;a) represent promising directions, though current evaluations lack noise baselines and have not yet surpassed random word selection. Finally, a neighbouring subfield has attempted non-invasive character decoding from brain or muscular activity due to typing or handwriting (Sivakumar et al., 2024; Lévy et al., 2025). While important, these are not B2T in the typical sense as they do not decode speech perception, but rather neural activity related to muscle movements. Thus, they require participants to move, making them infeasible for paralysed patients.

The apparent challenge of attaining significant results in non-invasive decoding is largely due to a data bottleneck. Datasets are typically very small, often just several hours, and collected across a few subjects. Thus, reaching deep learning scale necessitates combining datasets, which has yet to be convincingly achieved except for a few cases. These exceptions include brain foundation models (Zhang et al., 2023; Jiang et al., 2024; Wang et al., 2024; Yang et al., 2023; Kostas et al., 2021; Ye et al., 2023; Yi et al., 2023), particularly Jayalath et al. (2025) in speech decoding, as well as other approaches like Gideoni et al. (2024) using pooling in source space, and Ridge & Parker Jones (2024) through adversarial harmonisation. All of these methods show small improvements on simple tasks like classifying speech presence. In more sophisticated tasks, e.g. brain and audio segment matching (Défossez et al., 2023) or word

classification (d'Ascoli et al., 2024), no method has been able to improve performance by combining datasets.

To the best of our knowledge, ours is the first electrophysiological non-invasive B2T approach to: (a) decode sequences beyond chance and other baselines across a range of metrics and (b) show substantial improvements with dataset pooling. In the following section, we describe this B2T method.

3. Method

Our goal is to take continuous neural data and transcribe continuous text. This requires mapping a recording of brain activity $X \in \mathbb{R}^{s \times T}$, where s is the number of sensors and T is the number of time points in the signal, to a target text $Y \in \mathcal{V}^N$ where $\mathcal{V}$ is the vocabulary of the English language and N is the number of words constituting the stimulus text. Figure 2 provides an overview of our method which we describe in detail in the rest of this section.

3.1. Decoding Words

We begin by pooling brain recordings from multiple datasets. Then, we segment each brain recording into segments $x_i \in \mathbb{R}^{s \times t}$ of length t time points, where each segment starts at the onset of a word. We call this *aligning* the data. These segments then pass through a signal encoder.

The first stage of the signal encoder resolves differences between segments originating from different datasets and subjects. A spatial attention module (Défossez et al., 2023) projects data from different datasets with various sensor dimension s into a consistent latent dimension d_{pool}. Then, we apply a subject-specific linear layer to the latent embedding, maintaining its dimensions. The rest of the signal encoder is a series of dilated convolutions following the structure of standard brain signal encoders (Jayalath et al., 2025; Défossez et al., 2023; d'Ascoli et al., 2024). These convolutional layers learn a global representation of the neural data, transforming the latent spatial dimension from d_{pool} to d_{signal}. Finally, to get a time-invariant representation, we mean-pool over the temporal dimension t.

After the signal encoder, we follow d'Ascoli et al. (2024) in explicitly using a transformer to learn contextual relationships between word-aligned neural activity segments. The transformer encoder (Vaswani et al., 2017) learns to process a sequence of n consecutive latent embeddings and outputs n contextual embeddings. We use a contrastive loss (D-SigLIP (d'Ascoli et al., 2024)) to learn to match the outputs of the transformer to target word embeddings extracted from the middle layer of a T5 large language model (Raffel et al., 2020). We pre-compute the target embeddings from a vocabulary $\mathcal{V}_M$ made up of the top M most frequent words in the text of the dataset, which we call a retrieval set. We map the predicted embedding to a distribution of cosine similarities by measuring its distance to each of the target embeddings in the retrieval set. The model decodes words by predicting the word associated with the target embedding that maximises the cosine similarity to the prediction.

3.2. Advancing From Words to Sequences

So far, we have shown how to predict a limited vocabulary from brain data. In this section, we propose a method for open-vocabulary B2T from closed word prediction.

Our model predicts n cosine similarity distributions simultaneously for a sequence of n neural data segments. We transform each cosine similarity distribution output by our model into a probability distribution $P_{\text{model}}(w_i|\mathbf{x})$ using softmax. If we model each time step independently, the probability of a complete word sequence $\mathbf{w}_{1:n}$ given all neural data segments $\mathbf{x}$ is the product of individual word probabilities, leading to the best sequence being given by

$$\hat{\mathbf{w}}_{1:n} = \arg\max_{\mathbf{w}_{1:n} \in \mathcal{V}_M^n} \prod_{i=1}^{n} P_{\text{model}}(w_i \mid \mathbf{x}), \tag{1}$$

where $\mathcal{V}_M$ is the retrieval set vocabulary. This method is commonly known as *greedy* decoding.

Greedy decoding does not take into account the likelihood of sequences appearing in natural English text. Therefore, we introduce a pre-trained large language model to rescore the sequences based on how representative they are of natural language using the probability the LLM assigns to the sequence. We rescore as part of a beam search where the score of a candidate beam at step i is

$$\begin{aligned} \text{score}(\mathbf{w}_{1:i}|\mathbf{x}, \mathbf{w}_{1:i-1}) &= \text{score}(\mathbf{w}_{1:i-1}|\mathbf{x}, \mathbf{w}_{1:i-2}) && (2)\\ &+ \log P_{\text{model}}(w_i \mid \mathbf{x}) && (3)\\ &+ \lambda \log P_{\text{rescorer}}(\mathbf{w}_{1:i}) && (4) \end{aligned}$$

and the final predicted sequence is chosen as

$$\hat{\mathbf{w}}_{1:n} = \arg\max_{\mathbf{w}_{1:n} \in \mathcal{B}} \left[\text{score}(\mathbf{w}_{1:n}|\mathbf{x}, \mathbf{w}_{1:n-1})\right], \tag{5}$$

where $\mathcal{B}$ is the set of beam candidates, λ is a weight balancing the encoding model and LLM contributions, and $P_{\text{rescorer}}(\mathbf{w}_{1:n})$ denotes the probability of the word sequence under the LLM. In our experiments, P_{rescorer} comes from Llama 3.2-1B (Meta). Related methods (Willett et al., 2023) have used phoneme-level language models to rescore predictions with a Viterbi search (Viterbi, 1967), an exact and optimal decoding algorithm. The computational cost of LLMs and word-level decoding over long sequences makes this computationally infeasible, so we use beam search as an approximate alternative.

Word-level decoding is limited by closed vocabularies. While increasing the vocabulary permits decoding more

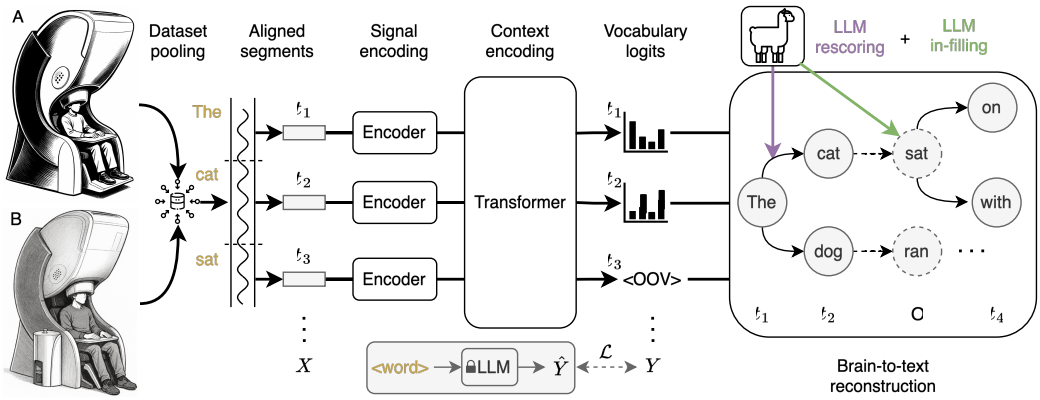

Figure 2: **Brain-to-text decoding method.** We pool data from multiple heterogeneous datasets (e.g. A and B) and align brain data segments to word onsets. Then, the segments are encoded by a signal encoder, handling dataset differences. A transformer learns the relationships between the encoded latents, embedding them with context. Its outputs are predictions of target word embeddings from a large language model. We map these predictions into logit distributions over the target vocabulary and a beam search with contextual rescoring constructs the highest probability sentence. If we detect out-of-vocabulary words in the stimulus, an in-filling model inserts words into these positions.

words, larger vocabularies reduce decoding performance (Appendix E). We propose taking advantage of performant vocabularies while compensating for out-of-vocabulary (OOV) words by detecting and in-filling them.

To detect OOV words, we hypothesise that the model will be less confident when presented with neural responses to words it has not seen. As the encoder's output probability distribution over words $\mathbf{p}_i \in \mathbb{R}^{|\mathcal{V}_M|}$ describes the model's confidence, we extract feature vectors from the trained encoder $\mathbf{f}_i = [\mathbf{p}_i, \phi(\mathbf{p}_i)]$, where $\phi(\mathbf{p}_i)$ is a set of distribution statistics, e.g. entropy. We fit a classifier to these features and estimate the probability a position is OOV, i.e. $P(w_i \notin \mathcal{V}_M | \mathbf{f}_i)$. This classifier selects positions to in-fill during inference. We provide further details in Appendix H. Next, we describe several possible in-filling strategies.

During beam decoding, we incrementally construct candidate word sequences $\mathbf{w}_{1:i}$. At each decoding step i, the model proposes the top K next-word candidates from its retrieval set vocabulary $\mathcal{V}_M$. For each beam hypothesis, if a proposed word $w_i \notin \mathcal{V}_M$, it is replaced in-place by an LLM completion conditioned on the current hypothesis.

Each beam is updated with the next word w_i chosen as

$$\begin{cases} \arg\max_{w_i \in \mathcal{V}_M} \text{score}(\mathbf{w}_{1:i} | \mathbf{x}, \mathbf{w}_{1:i-1}), & \text{if } w_i \in \mathcal{V}_M \\ \arg\max_{w_i \in \mathcal{V}_{\text{filler}}} P_{\text{filler}}(w_i \mid \mathbf{w}_{1:i-1}), & \text{if } w_i \notin \mathcal{V}_M \end{cases} \tag{6}$$

where P_{filler} is the LLM completion probability. Here, the full beam search proceeds with these substituted hypotheses and we select the best hypothesis from the beam set as before (Equation 5). Note that $P_{\text{filler}}(w_i \mid \mathbf{w}_{1:i-1})$ is not trivial to calculate from a typical LLM as they operate at the token level rather than the word level. To get word probabilities from token-level probabilities, we use a further token-level beam search over the LLM predictions, where the beam hypothesis is a word hypothesis. While the LLM used for in-filling could be different to the one used for rescoring, the weights are shared in our experiments.

In-filling during the beam search has two limitations. First, it requires another beam search to resolve token-level probability distributions into word hypotheses. Second, it uses only knowledge of prior words in the sequence and not future words. In response, we design an in-context LLM method that in-fills missing positions only after generating the entire best beam sequence. In this method, we perform the full beam search as in Equation 5 while inserting <UNK> in the positions of out-of-vocabulary words. We then prompt an LLM with the best beam sequence

$$W = \{w_0, w_1, \ldots, w_n\} \text{ where } w_i \in \mathcal{V}_M \cup \{\texttt{<UNK>}\}, \tag{7}$$

and instructions to fill in the missing positions and output the complete sequence. As the context now features all the known words in the sequence, in-filling makes use of both prior and future context. While this could equally be possible with a bidirectional LLM (e.g. BERT (Devlin et al., 2019)), doing this in-context allows direct in-filling of words rather than indirectly via tokens.

This method, too, has limitations. While in-filling during beam search does not use information from later in the candidate sequence, in-context in-filling does not rescore while taking into account words that could fill missing positions. These limitations can be addressed by both rescoring and in-filling at the same time. To do this, we prompt an LLM as before, but instead of beam rescoring first, we let the LLM

choose from the five highest probability words. Thus, rather than a beam sequence, our prompt contains a sequence of quintuples of words and their probabilities

$$P = \{p_0, p_1, \ldots, p_n\} \text{ where } p_i \in \mathcal{C}_5 \cup \{\texttt{<UNK>}\} \quad (8)$$

$$\mathcal{C}_5 = \{(w_{i1}, \rho_{i1}), \ldots) \mid w_{ij} \in V, \rho_{ij} \in [0, 1]\}. \quad (9)$$

Here, p_i is a quintuple consisting of pairs of words w_{ij} and their probabilities ρ_{ij}. The LLM selects the best word for known positions based on its notion of sentence coherency, and fills in the missing word for unknown positions, those marked with <UNK>, outputting a complete sequence.

4. Experiments

Datasets Our main dataset is LibriBrain (Landau et al., 2025; Özdogan et al., 2025), a 50-hour MEG dataset collected from a subject listening to audiobooks. We also use three auxiliary datasets for pooling and other experiments: Armeni et al. (2022), containing MEG recordings of 3 subjects listening to 10 hours each of stories, Gwilliams et al. (2023) with MEG from 27 subjects, each listening to 4 short stories, and Broderick et al. (2018), providing EEG from 19 subjects, each listening to 20 short speech segments.

Methods *Greedy* describes taking the highest probability words (Equation 1), *beam* refers to LLM rescoring (Equation 5), *+fill* indicates in-filling words while rescoring (Equation 6), *+IC fill* denotes in-filling words in-context after rescoring (Equation 7), and *IC transcribe* is both in-context rescoring and in-filling (Equation 8). We provide two random baselines (Appendix I). When specifying *OOV-D*, our OOV detector selects in-filling positions.

Evaluation We quote six brain-to-text metrics covering a broad range of sentence decoding facets. Word error rate (WER) and character error rate (CER) are coarse- and fine-grained measures of exact match decoding accuracy. As prior B2T work measure n-gram overlap, we also quote BLEU-1 (Papineni et al., 2002), ROUGE-1F (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) as these are standard (Tang et al., 2023; Jo et al., 2024; Yang et al., 2024a). Finally, we quote BERTScore (Zhang et al., 2020) as a measure of semantic similarity. When we do not decode sequences, we provide top-10 balanced word classification accuracy. This is a macro average equivalent to calculating the accuracy for each word in the vocabulary and taking the mean of these accuracies. In tables, we bold the best result and underline the second-best result in a column. Empty cells indicate metrics not reported in the original paper. We always quote the mean and standard error.

4.1. Realising Non-Invasive Brain-To-Text

On raw brain-to-text performance, we outperform all of the previously best reported non-invasive methods (Table 1). These results establish a new state-of-the-art in non-invasive speech decoding across a range of metrics. We provide annotated decoding examples and more in Appendix M.

As prior work uses different experimental protocols, the table also provides the difference between the best method and any quoted null or random baseline (*delta to random*). Here, our approach shows a substantially stronger improvement than any previous work. When compared to prior work in EEG (Wang & Ji, 2022; Duan et al., 2023), we achieve significantly better results across all metrics. This is unsurprising as the EEG-to-text methods are no better than chance (Jo et al., 2024). Compared to the state-of-the-art fMRI method (Tang et al., 2023), ours equals or outperforms it across nearly all metrics.

While there are no established works in B2T from MEG, there are a series of recent preprints (Yang et al., 2024b;c;a) which our methods also convincingly outperform (Table 2). While NeuSpeech (Yang et al., 2024b) shows promising character error rates, it does not surpass chance in BLEU or ROUGE. MAD (Yang et al., 2024c) is similar, but with worse character errors. NeuGPT (Yang et al., 2024a) is better on n-gram overlap metrics, but has a character error rate worse than random. Moreover, all of these methods show BERTScores only on par with their random selection baselines. Thus, none of these approaches convincingly surpass chance across all metrics. Our methods show superior performance across the board.

An ablation over our approach (Table 3) shows that it is highly significant, surpassing random selection and noise baselines across all metrics unlike prior attempts at B2T. We are careful to provide noise baselines as they verify that our model does not overfit to the target story and that the results are attributable to decoding brain data. Jo et al. (2024) showed that this is a critical failing of prior work. With MEG, we note that our beam method significantly improves n-gram metrics. Beam+fill improves almost everything. Beam+IC fill performs similarly and is excellent for semantics which we conjecture is because it has the entire sequence as context. Our method performs well even without known OOV positions. With 88% AUROC, the detector reliably distinguishes neural responses to in-vocabulary and OOV words. IC transcribe is also effective, though does not outperform true rescoring.

4.2. Overcoming the Data Bottleneck

On word classification, we show that selective pooling of datasets improves performance by more than double (Figure 3). This also improves WER on B2T (*pooled* in Table 2).

Speech decoding suffers from a data bottleneck as datasets are typically small and collected from few subjects. This bottleneck exists even for the largest MEG speech datasets

Table 1: **Our method improves non-invasive B2T over established alternatives.** We quote the mean over the three subjects of Tang et al. (2023) and the best results for Wang & Ji (2022) from a replication study without teacher-forcing (Jo et al., 2024). *Delta to random* shows the difference between a method and its associated random baseline.

Method	WER ↓	CER ↓	BLEU ↑	ROUGE ↑	METEOR ↑	BERT ↑
Tang et al. (2023)	0.93		.24		**.17**	**.81**
EEG-to-text (Wang & Ji, 2022; Duan et al., 2023)	1.00		.14	.12		
Ours (best)	**0.88**	**0.68**	**.25**	**.26**	.15	**.81**
Greedy	**0.88**	0.80	.21	.21	.12	.72
Beam+fill	0.91	**0.68**	**.25**	**.26**	.15	.80
Beam+IC fill	0.90	0.71	.24	.24	.15	**.81**
Delta to random						
Tang et al. (2023)	−.03		+.05		+.04	+.02
EEG-to-text (Wang & Ji, 2022; Duan et al., 2023)	+.00		+.00	+.00		
Ours (best)	**−.12**	**−.19**	**+.18**	**+.18**	**+.09**	**+.04**

Table 2: **Our method surpasses all prior MEG brain-to-text approaches.** As no code is available to reproduce NeuGPT (Yang et al., 2024a), we validate and test with the same data and splits and quote their results in this table, including their baselines (Yang et al., 2024b;c). We also provide two random selection baselines, one from Yang et al. (2024a) and our own. *Pooled* indicates the model was additionally trained jointly with LibriBrain and Armeni.

Method	WER ↓	CER ↓	BLEU ↑	ROUGE ↑	METEOR ↑	BERT ↑
Random (Yang et al., 2024a)		.87	.06	.07		.84
Random (ours)	$1.00_{\pm.0003}$	$.91_{\pm.004}$	$.04_{\pm.001}$	$.04_{\pm.002}$	$.04_{\pm.001}$	$.77_{\pm.0003}$
NeuSpeech (Yang et al., 2024b)		.77	.05	.08		**.84**
MAD (Yang et al., 2024c)		.90	.07	.07		.83
NeuGPT (Yang et al., 2024a)		1.0	.13	.13		**.84**
Ours						
Beam+fill	$0.97_{\pm.001}$	$\mathbf{.71}_{\pm.001}$	$\mathbf{.18}_{\pm.003}$	$\mathbf{.19}_{\pm.003}$	$\mathbf{.12}_{\pm.001}$	$.79_{\pm.001}$
Beam+IC fill	$0.98_{\pm.001}$	$.78_{\pm.003}$	$\mathbf{.18}_{\pm.003}$	$.18_{\pm.003}$	$\mathbf{.12}_{\pm.002}$	$.80_{\pm.001}$
Ours (pooled)						
Beam+fill	$\mathbf{0.95}_{\pm.001}$	$.72_{\pm.004}$	$\mathbf{.18}_{\pm.003}$	$\mathbf{.19}_{\pm.003}$	$.11_{\pm.002}$	$.78_{\pm.003}$
Beam+IC fill	$\mathbf{0.95}_{\pm.003}$	$.82_{\pm0.009}$	$.17_{\pm0.004}$	$.17_{\pm0.005}$	$.11_{\pm0.002}$	$.80_{\pm0.001}$

(Appendix C). Continued improvements from scaling indicate that more data could increase performance further.

One way to increase data is to combine datasets. This turns out not to be trivial. Since neural data come from heterogeneous sources with different numbers of sensors, we transform samples of various sensor dimensions to a consistent latent spatial dimension. We use a spatial attention (Défossez et al., 2023) which projects samples using attention scores derived from the (x, y) coordinates of each sensor. We note that there are many other plausible ways of doing this without using spatial information that perform equally well (Appendix A), mainly attributing improvements to selective pooling rather than any harmonisation method. We leave handling other sources of heterogeneity between datasets, such as anatomical, cognitive, and other hardware differences to the neural network.

While prior work has not seen significant improvements with pooled data in supervised decoding, we show for the first time that this is possible. Figure 3A demonstrates that training with higher quality data, as measured by standalone performance, improves accuracy on lower quality datasets. In fact, the standalone performance of a dataset correlates strongly and significantly with the average improvement from jointly training with it ($r = .95$, $p = .048$). Even when jointly training EEG data with MEG data, as in the case of Broderick, we see that MEG, which is higher in quality, raises accuracy on the EEG data to statistically significant performance. This suggests that jointly training noisier modalities with higher quality data may improve accuracy on these modalities.

Combining multiple higher quality datasets with a target dataset leads to even more significant gains (Figure 3B).

Table 3: **Our method is significant against all critical random baselines.** Random selection picks a word uniformly randomly from the story vocabulary. Noise inputs indicates when random noise of the same mean and standard deviation as true samples is input to the trained model. We use a retrieval set of 250 words and decode the LibriBrain dataset.

Method	WER ↓	CER ↓	BLEU ↑	ROUGE ↑	METEOR ↑	BERT ↑
Rand. selection	$1.00_{\pm.0002}$	$.87_{\pm.001}$	$.07_{\pm.002}$	$.08_{\pm.002}$	$.06_{\pm.001}$	$.77_{\pm.0004}$
Noise inputs						
Greedy	$1.00_{\pm.0003}$	$.87_{\pm.004}$	$.07_{\pm.0004}$	$.07_{\pm.001}$	$.05_{\pm.001}$	$.74_{\pm.0003}$
+rand. fill	$1.00_{\pm.0003}$	$.91_{\pm.004}$	$.07_{\pm.0004}$	$.08_{\pm.001}$	$.05_{\pm.001}$	$.77_{\pm.001}$
Beam	$0.99_{\pm.001}$	$.85_{\pm.003}$	$.10_{\pm.004}$	$.10_{\pm.004}$	$.06_{\pm.003}$	$.74_{\pm.001}$
+rand. fill	$0.99_{\pm.001}$	$.88_{\pm.003}$	$.10_{\pm.004}$	$.10_{\pm.004}$	$.06_{\pm.003}$	$.77_{\pm.001}$
+fill	$0.98_{\pm.001}$	$.75_{\pm.006}$	$.15_{\pm.008}$	$.16_{\pm.008}$	$.09_{\pm.006}$	$.78_{\pm.003}$
+IC fill	$0.99_{\pm.0004}$	$.77_{\pm.003}$	$.16_{\pm.005}$	$.16_{\pm.004}$	$.10_{\pm.003}$	$.79_{\pm.001}$
IC transcribe	$0.99_{\pm.001}$	$.84_{\pm.004}$	$.13_{\pm.001}$	$.13_{\pm.001}$	$.08_{\pm.002}$	$.79_{\pm.001}$
Real inputs						
Greedy	$\mathbf{0.90}_{\pm.002}$	$.77_{\pm.002}$	$.19_{\pm.003}$	$.20_{\pm.003}$	$.12_{\pm.002}$	$.74_{\pm.001}$
+rand. fill	$\mathbf{0.90}_{\pm.002}$	$.78_{\pm.002}$	$.19_{\pm.003}$	$.20_{\pm.003}$	$.13_{\pm.003}$	$.79_{\pm.001}$
Beam	$\mathbf{0.90}_{\pm.001}$	$.76_{\pm.001}$	$.22_{\pm.002}$	$.22_{\pm.002}$	$.13_{\pm.002}$	$.75_{\pm.0003}$
+rand. fill	$\mathbf{0.90}_{\pm.002}$	$.77_{\pm.002}$	$.22_{\pm.002}$	$.22_{\pm.002}$	$.13_{\pm.002}$	$.79_{\pm.0002}$
+fill	$\underline{0.91}_{\pm.002}$	$\mathbf{.69}_{\pm.001}$	$\mathbf{.25}_{\pm.002}$	$\mathbf{.25}_{\pm.002}$	$\mathbf{.15}_{\pm.002}$	$\underline{.80}_{\pm.001}$
+fill (OOV-D)	$\underline{0.91}_{\pm.002}$	$\mathbf{.69}_{\pm.001}$	$\underline{.24}_{\pm.002}$	$\mathbf{.25}_{\pm.002}$	$\mathbf{.15}_{\pm.001}$	$\underline{.80}_{\pm.001}$
+IC fill	$\mathbf{0.90}_{\pm.002}$	$\underline{.71}_{\pm.002}$	$\underline{.24}_{\pm.003}$	$\underline{.24}_{\pm.003}$	$\mathbf{.15}_{\pm.003}$	$\mathbf{.81}_{\pm.001}$
IC transcribe	$\mathbf{0.90}_{\pm.004}$	$.75_{\pm.004}$	$.21_{\pm.005}$	$.22_{\pm.005}$	$.14_{\pm.004}$	$\underline{.80}_{\pm.001}$

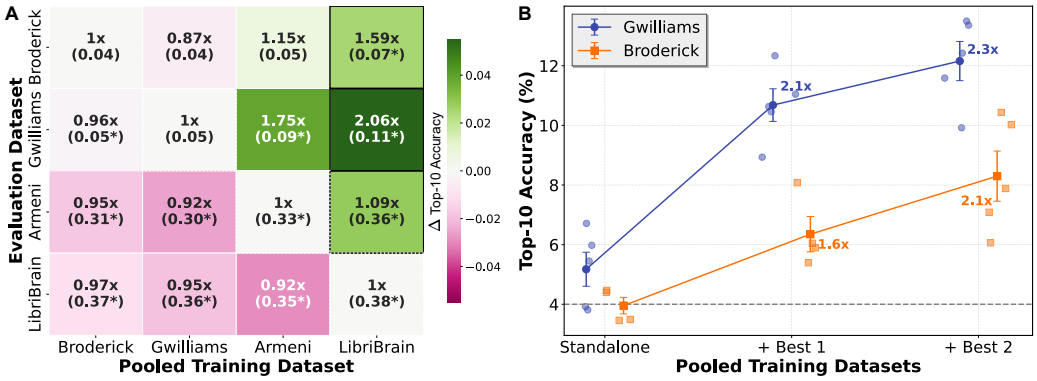

Figure 3: **Selectively pooling data improves accuracy by 2.3×. (A) Pairing datasets in training.** Numbers show accuracy improvement on the evaluation dataset when trained additionally with the training dataset. Numbers in brackets show raw accuracy (asterisks are statistically significant against chance). The diagonal shows no paired training, i.e. standalone. Shading shows the change in accuracy relative to standalone. **(B) Exploiting selective pooling doubles accuracy.** We combine the target data (Gwilliams and Broderick) with the best, and the best and second best, datasets by quality, i.e. LibriBrain and Armeni, leading to dramatic improvements on the target data.

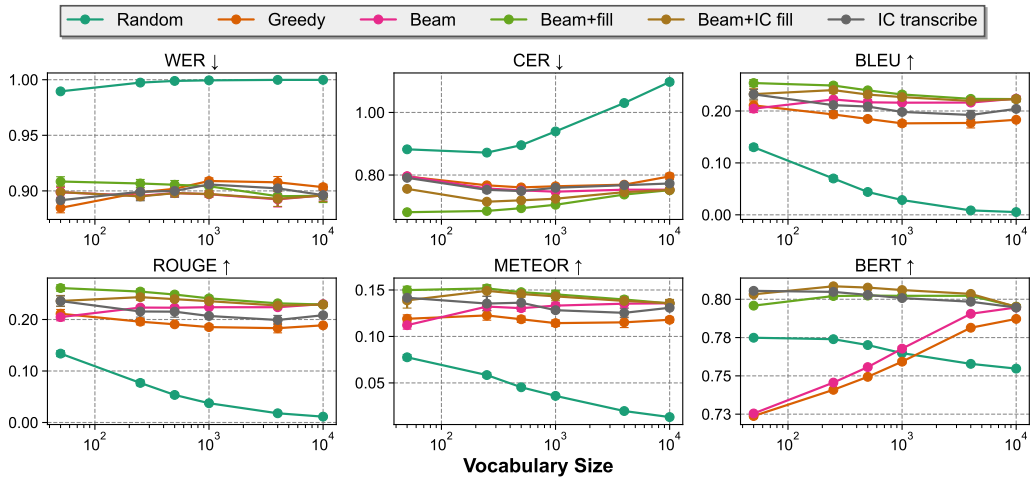

Figure 4: **Optimal vocabulary size differs by method and metric.** We train our model with increasingly large vocabularies and observe how this effects the performance of our decoding strategies across metrics. Methods with/without in-filling tend to converge with larger vocabularies as there are fewer words to in-fill.

Here, we again treat standalone performance as a measure of dataset quality and the metric by which to choose whether to pool a dataset with another. The factors that affect quality include the modality of the data, the number of subjects, the experimental protocol, the number of hours of recordings, and much more. We find that standalone performance is a useful overall proxy for quality as it takes into account all of these facets by being learned from the data.

4.3. Scaling Vocabulary Size

Increasing the vocabulary size lowers the number of OOV predictions but reduces the test accuracy within-vocabulary, even on a fixed set of words that are always in-vocabulary (Appendix E). When leveraging sentence decoding strategies, this trade-off implies an optimal vocabulary size which is small enough for high word accuracy, but large enough that we do not need to in-fill too many words. Figure 4 reveals that the optimal vocabulary size actually differs for different methods and evaluation metrics.

For WER, a smaller vocabulary is best, ensuring high word accuracy. Any signal on additional words with a larger vocabulary is offset by the general reduction in word classification accuracy. This is because WER is a hard measure where any difference to the ground truth word counts as an incorrect prediction. CER is softer as it operates at the character level where similar words, in terms of characters, will count as partially correct. Consequently, it can improve with the signal provided by larger vocabularies. Here, for most methods, the optimal vocabulary is 250 words.

On the remaining metrics, we see mixed effects. With BLEU, ROUGE, and METEOR, methods are optimal with either 50 or 250-word vocabularies, with minor differences. Generally, larger retrieval sets lead to worse performance. On BERT, in-filling methods are best with a 250-word vocabulary. Somewhat surprisingly, greedy and beam improve with vocabulary size. The additional signal from using a larger vocabulary, while reducing exact matches, may increase the overall number of semantically relevant words in the prediction, leading to this improvement in BERTScore.

5. Conclusion & Future Work

We have demonstrated the first non-invasive brain-to-text system that comprehensively outperforms all chance baselines critical for establishing non-invasive B2T. Moreover, it consistently surpasses existing methods across all standard metrics, improving BLEU scores by up to $2.6\times$. Our approach combines three key innovations: extending word-level decoding to sequence decoding with language model rescoring, transforming closed-vocabulary word-level classifiers into effectively open-vocabulary decoders with predictive in-filling, and unlocking scaling of B2T systems through a selective dataset pooling strategy that more than doubles word classification accuracy.

Building on these results and their limitations (Appendix J), we identify three future research directions: (1) cross-modal training by pooling invasive and non-invasive recordings to improve non-invasive performance through transfer from invasive data; (2) scaling up datasets to find the boundaries of non-invasive B2T; and (3) improving practical B2T without requiring exact matches, e.g. by boosting the semantics and coherence of decoded sequences.

Impact Statement

Although currently preliminary, maturity of research on non-invasive brain-to-text technology presents significant implications. Potential benefits include enabling communication for individuals with paralysis and advancing brain-computer interfaces generally. However, mature iterations of this technology raise concerns about brain data privacy, potential misuse of neural information, and accessibility disparities that may exacerbate existing inequalities. At present, we limit potential misuse by decoding heard speech rather than inner speech. As the field advances to inner speech, we encourage researchers to engage with ethical principles and set common ethical standards.

References

Anthropic. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-04-22.

Armeni, K., Güçlü, U., van Gerven, M., and Schoffelen, J.-M. A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data*, 9(1):278, 2022. Licensed under CC-BY-4.0.

Ball, T., Kern, M., Mutschler, I., Aertsen, A., and Schulze-Bonhage, A. Signal quality of simultaneously recorded invasive and non-invasive EEG. *Neuroimage*, 46(3):708–716, 2009.

Banerjee, S. and Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018. Licensed under CC0 1.0 Universal.

Card, N. S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F. R., Kunz, E. M., Fan, C., Vahdati Nia, M., Deo, D. R., et al. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

d'Ascoli, S., Bel, C., Rapin, J., Banville, H., Benchetrit, Y., Pallier, C., and King, J.-R. Decoding individual words from non-invasive brain recordings across 723 participants. *arXiv preprint arXiv:2412.17829*, 2024.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., and King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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*, pp. 4171–4186, 2019.

Duan, Y., Zhou, J., Wang, Z., Wang, Y.-K., and Lin, C.-t. Dewave: Discrete encoding of EEG waves for EEG to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918, 2023.

Gideoni, Y., Timms, R. C., and Parker Jones, O. Non-invasive neural decoding in source reconstructed brain space. *arXiv preprint arXiv:2410.19838*, 2024.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Gwilliams, L., Flick, G., Marantz, A., Pylkkänen, L., Poeppel, D., and King, J.-R. Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific data*, 10(1): 862, 2023. Licensed under CC0 1.0 Universal.

Harris, K. D. Nonsense correlations in neuroscience. *biorxiv*, pp. 2020–11, 2020.

Jayalath, D., Landau, G., Shillingford, B., Woolrich, M., and Parker Jones, O. The Brain's Bitter Lesson: Scaling speech decoding with self-supervised learning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, Canada, July 13-19, 2025*. PMLR, 2025.

Jiang, W., Zhao, L., and Lu, B. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Jo, H., Yang, Y., Han, J., Duan, Y., Xiong, H., and Lee, W. H. Are EEG-to-text models working? *The Fourth International Workshop on Human Brain and Artificial Intelligence at IJCAI*, 2024.

Kostas, D., Aroca-Ouellette, S., and Rudzicz, F. BENDR: Using transformers and a contrastive self-supervised

learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Landau, G., Özdogan, M., Elvers, G., Mantegna, F., Somaiya, P., Jayalath, D., Kurth, L., Kwon, T., Shillingford, B., Farquhar, G., Jiang, M., Jerbi, K., Abdelhedi, H., Mantilla Ramos, Y., Gulcehre, C., Woolrich, M., Voets, N., and Parker Jones, O. The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. *arXiv preprint*, 2025.

Lévy, J., Zhang, M., Pinet, S., Rapin, J., Banville, H., d'Ascoli, S., and King, J.-R. Brain-to-text decoding: A non-invasive approach via typing. *arXiv preprint arXiv:2502.17480*, 2025.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Littlejohn, K. T., Cho, C. J., Liu, J. R., Silva, A. B., Yu, B., Anderson, V. R., Kurtz-Miott, C. M., Brosler, S., Kashyap, A. P., Hallinan, I. P., et al. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature Neuroscience*, pp. 1–11, 2025.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR*, 2019.

Meta. meta-llama/Llama-3.2-1B. https://huggingface.co/meta-llama/Llama-3.2-1B. Accessed: 2025-04-26.

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3): 217–227, 2021.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Ridge, J. and Parker Jones, O. Resolving domain shift for representations of speech in non-invasive brain recordings. *arXiv preprint arXiv:2410.19986*, 2024.

Sivakumar, V., Seely, J., Du, A., Bittner, S., Berenzweig, A., Bolarinwa, A., Gramfort, A., and Mandel, M. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. *Advances in Neural Information Processing Systems*, 37:91373–91389, 2024.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967. doi: 10.1109/TIT.1967.1054010.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., and Li, H. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

Wang, Z. and Ji, H. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Yang, C., Westover, M., and Sun, J. BIOT: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Yang, Y., Duan, Y., Jo, H., Zhang, Q., Xu, R., Parker Jones, O., Hu, X., Lin, C.-t., and Xiong, H. NeuGPT: Unified multi-modal neural GPT. *arXiv preprint arXiv:2410.20916*, 2024a.

Yang, Y., Duan, Y., Zhang, Q., Jo, H., Zhou, J., Lee, W. H., Xu, R., and Xiong, H. NeuSpeech: Decode neural signal as speech. *arXiv preprint arXiv:2403.01748*, 2024b.

Yang, Y., Jo, H., Duan, Y., Zhang, Q., Zhou, J., Lee, W. H., Xu, R., and Xiong, H. MAD: Multi-alignment MEG-to-text decoding. *arXiv preprint arXiv:2406.01512*, 2024c.

Ye, J., Collinger, J., Wehbe, L., and Gaunt, R. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36:80352–80374, 2023.

Yi, K., Wang, Y., Ren, K., and Li, D. Learning topology-agnostic EEG representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Zhang, D., Yuan, Z., Yang, Y., Chen, J., Wang, J., and Li, Y. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36: 26304–26321, 2023.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*, 2020.

Özdogan, M., Landau, G., Elvers, G., Jayalath, D., Somaiya, P., Mantegna, F., Woolrich, M., and Parker Jones, O. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. *arXiv preprint*, 2025. Licensed under CC-BY-NC.

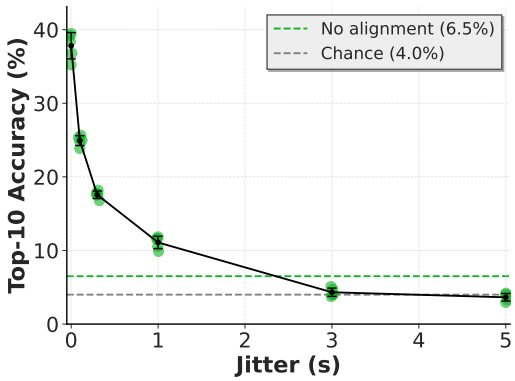

Figure 5: **Alignment is critical.** We add a random jitter in the range $[0, \text{jitter}]$ to the aligned input samples. We train on LibriBrain with a vocabulary size of 250 and quote word classification accuracy.

A. Sensor Positions and Dataset Pooling

A key concern in pooling data is that of the differences in sensor geometry between scanners. For example, LibriBrain uses 306 sensors, the Armeni scanner uses 269 sensors, and the Gwilliams scanner has 208. In each dataset, the sensor geometry is different, with sensors placed in different configurations and therefore picking up signals from different parts of the brain. Explicitly leveraging sensor position information in the neural network to provide useful inductive biases is one way in which datasets could be pooled more effectively.

In Table 4, we compare the spatial attention (SA) method (Défossez et al., 2023) to inserting zeroes to match spatial dimensions between datasets (*padding*), dataset-conditional linear or convolutional projections to a shared dimension (*gating*) (Jayalath et al., 2025), as well as spatial attention followed by gating. In all of these methods, once the data is in the same space, we allow the neural network to resolve other data differences during training.

We note that padding and gating are not explicitly provided spatial information via sensor positions and in all cases, the performance differences are not significant. While spatial attention can improve individual dataset performance (Défossez et al., 2023, Table A.2), our results indicate that it does not benefit aggregated performance. This means that either spatial differences between data can be harmonised through learning without explicit knowledge of sensor positions, or that the networks are not yet able to make proper use of this spatial information.

Table 4: **Pooled dataset performance with different pooling methods.** We evaluate Gwilliams while jointly training it alongside Armeni and LibriBrain. We conduct a t-test against the spatial attention method and find that none of the alternatives perform differently to a statistically significant degree with five seeds. Uncertainty is standard error.

Method	Top-10 Accuracy	Different to SA? (p-value)
SA + Gating	$.127_{\pm.007}$	No (.62)
Gating	$.124_{\pm.005}$	No (.75)
Padding	$.119_{\pm.005}$	No (.79)
SA	$.122_{\pm.007}$	—

B. Relaxing Alignment and Decoding Without Alignment

Although aligning brain data to word onsets is possible by getting patients to space their attempted speech (e.g. Moses et al. (2021)), decoding without alignment is useful as it permits natural speaking patterns. Here, we investigate reducing alignment and decoding without alignment altogether.

Figure 5 shows that decoding is highly sensitive to alignment. Even a small jitter drastically reduces word classification accuracy. Still, removing alignment by evenly distributing the input samples in a sequence leads to statistically significant results.

Table 5: **CTC Decoding Results.**

Method	WER ↓	CER ↓	BLEU ↑	ROUGE ↑	METEOR ↑	BERT ↑
Rand. selection	$\underline{1.00}_{\pm.0002}$	$.87_{\pm.001}$	$\underline{.07}_{\pm.002}$	$\mathbf{.08}_{\pm.002}$	$\mathbf{.06}_{\pm.001}$	$\mathbf{.77}_{\pm.0004}$
CTC-Greedy	$\mathbf{0.99}_{\pm.001}$	$\mathbf{.73}_{\pm.01}$	$\underline{.07}_{\pm.003}$	$.07_{\pm.002}$	$.05_{\pm.002}$	$.76_{\pm001}$
CTC-Beam	$\underline{1.00}_{\pm.001}$	$\underline{.74}_{\pm.011}$	$\mathbf{.08}_{\pm.004}$	$\mathbf{.08}_{\pm.005}$	$\mathbf{.06}_{\pm.003}$	$\mathbf{.77}_{\pm.004}$

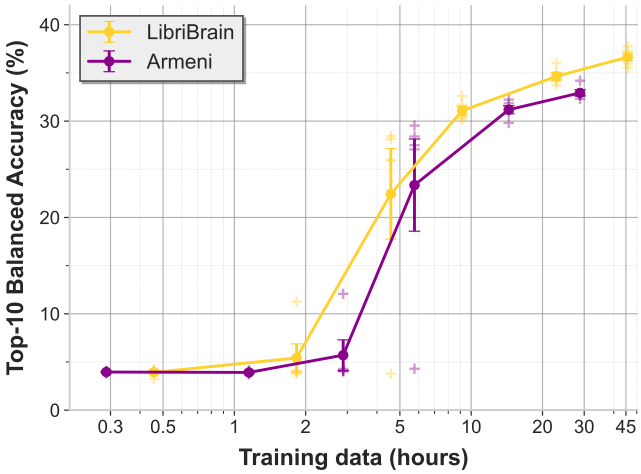

Figure 6: **Speech decoding is data-limited.** The plot shows top-10 word classification accuracy on the test set as we scale up the volume of data in the train set of LibriBrain and Armeni. To scale up, we randomly sample, without replacement, increasing numbers of samples from all sessions and subjects in the training set.

However, evenly distributing inputs requires knowing the length of the target sequence. Thus, we use CTC-style merging, i.e. duplicate merging and alignment-free prediction, when even this is not known. This gives significantly better-than-chance CER, suggesting that this predicts meaningful acoustic signals (Table 5).

Our merging method (a variant of Graves et al. (2006)) decodes sequences from a trained model without known word onsets or alignments. In this setup, we space all the MEG windows at a fixed interval of 0.3s which is approximately the average time between words. We then make a set of predictions and slide all the segments until we reach the end of the sentence. The slide interval is half the MEG window spacing to make sure that we can reasonably resolve most words.

Once we have all the predictions, we average the probabilities for positions where there are multiple predictions and apply a 1D average pooling operation on the probabilities in the time dimension with a kernel size of 5 and stride of 3. Then, we collapse repeats in any section of consecutive repeats to a single prediction. CTC-greedy is then the argmax word predictions while CTC-beam emerges from applying LLM rescoring to reconstruct the sentence. All CTC results use a model trained on the LibriBrain dataset with a 250-word retrieval set.

C. Scaling Data Is Not Enough

Among MEG speech decoding datasets, there are none with more than a hundred subjects, and none with more than 50 hours of within-subject recordings. As a result, speech decoding is data-limited, which we show in Figure 6. Extrapolating the performance on LibriBrain suggests that scaling up to 80% top-10 accuracy would require several orders of magnitude more data. Evidently, further work needs to be done on methods that scale in addition to collecting more data.

D. Data Splits and Nonsense Correlations

We provide details on our data splits in Table 6. In designing our data splits, we were very careful to avoid nonsense correlations (Harris, 2020) by using independent sessions collected at different times. We also took care to avoid stimulus

leakage within-dataset as well as across datasets when jointly training. This allows us to accurately evaluate content generalisation and ensures that all the information in the test set, including both the data and stimuli, are unseen. Taking random splits over these datasets, as is common in machine learning, would lead to leakage of stimuli as the same stimulus can be presented to multiple subjects. Even when stimuli are guaranteed to be non-overlapping, taking test set samples from the same recording as train set samples risks nonsense correlations as neural data and target text can be autocorrelated over time. For example, the semantics of two consecutive sentences in a session are likely to be very similar. Slow drift artifacts in the data at the time of these sentences can therefore be correlated with their semantics, thus leading to a nonsense correlation.

Table 6: **Data splits.** When training Armeni and LibriBrain jointly, we additionally exclude Sherlock3 session 9 and 10 from LibriBrain in the train set as the stories (The Adventure of the Engineer's Thumb and The Adventure of The Noble Bachelor) overlap with the validation and test set stories in the Armeni dataset. When comparing to the prior MEG brain-to-text work (Yang et al., 2024b;c;a), we instead validate on Gwilliams task 1 and test on Gwilliams task 0 to match the experimental setup for the results quoted in Yang et al. (2024a). We use all available subjects for train, validation, and test.

Dataset	Train	Validation	Test
LibriBrain (Özdogan et al., 2025)	Sherlock1-6 (excl. val and test)	Sherlock1 Sess. 9	Sherlock1 Sess. 10-12
Armeni (Armeni et al., 2022)	Sess. 1-8	Sess. 9	Sess. 10
Gwilliams (Gwilliams et al., 2023)	Task 2-3	Task 0	Task 1
Broderick (Broderick et al., 2018)	All sess. (excl. val and test)	Sess. 3, 9, & 19	Sess. 7, 8, & 14

In practice, we find that differences in results due to potential nonsense correlations are not statistically significant (Table 7). If there were nonsense correlations, we would expect the test set in the overlapping case to outperform the test set in the independent case due to overfitting to nonsense correlations. Similarly, we would expect the holdout set to perform worse than the test set in the overlapping case due to this overfitting. Here, and in Figure 7, we use the same experimental setup. We hypothesise that high-pass filtering, standardisation of samples, and baseline correction all help to avoid more significant nonsense correlations due to slow drift artifacts. Other sources of nonsense correlations may be insignificant in this experimental setup.

E. Vocabulary Scaling and Coverage

In Figure 8, on the left we show the percentage of the LibriBrain story text that can be covered by vocabularies of various sizes constructed from the N most frequent words. Notably, a 50-word retrieval set covers 48% of all words in the text and a 250-word dataset covers 68%. Scaling up to 1000 words brings this up to only 82%. On the right of the figure, we show that scaling up the retrieval set size reduces the decoding accuracy, first across the retrieval set, and second across a fixed set of the top 50 most frequent words, which are always in the vocabulary.

Table 7: **Generalisation with data splits of overlapping vs independent sessions.** We use subject 1 of the Armeni dataset for this experiment to avoid stimuli leakage. For independent session splitting, we randomly sample from the first nine sessions, seven sessions as the train set, and one session each as the validation and test set. For overlapping session splitting, we randomly split (without replacement) $\frac{7}{9}$ths of the data in the first nine sessions as train, another $\frac{1}{9}$th as validation, and the last $\frac{1}{9}$th as test. For both methods, we use session 10 as the holdout set to test generalisation beyond the potentially nonsensically correlated test set. We determine the significance of the difference between the data splitting approaches using 10 seeds. Note: n.s. = not significant.

Evaluation set	Overlapping	Independent	Significance
Test	$.282_{\pm.004}$	$.282_{\pm.003}$	n.s. ($p = .87$)
Holdout	$.278_{\pm.002}$	$.282_{\pm.002}$	n.s. ($p = .21$)
Significance	n.s. ($p = .15$)	n.s. ($p = .77$)	

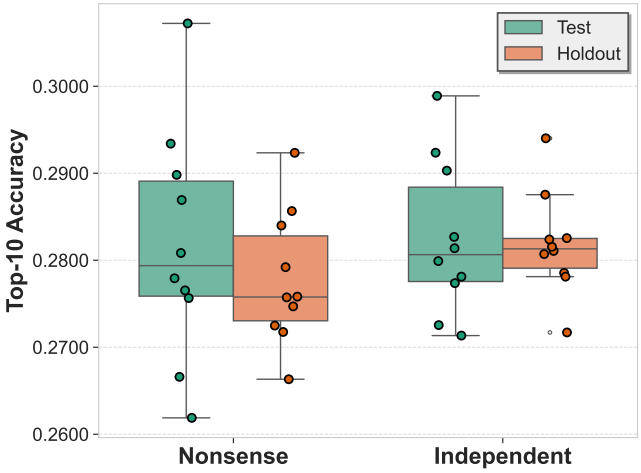

Figure 7: **Generalization with data splits containing overlapping vs independent sessions.** Refer to Table 7 for complete details.

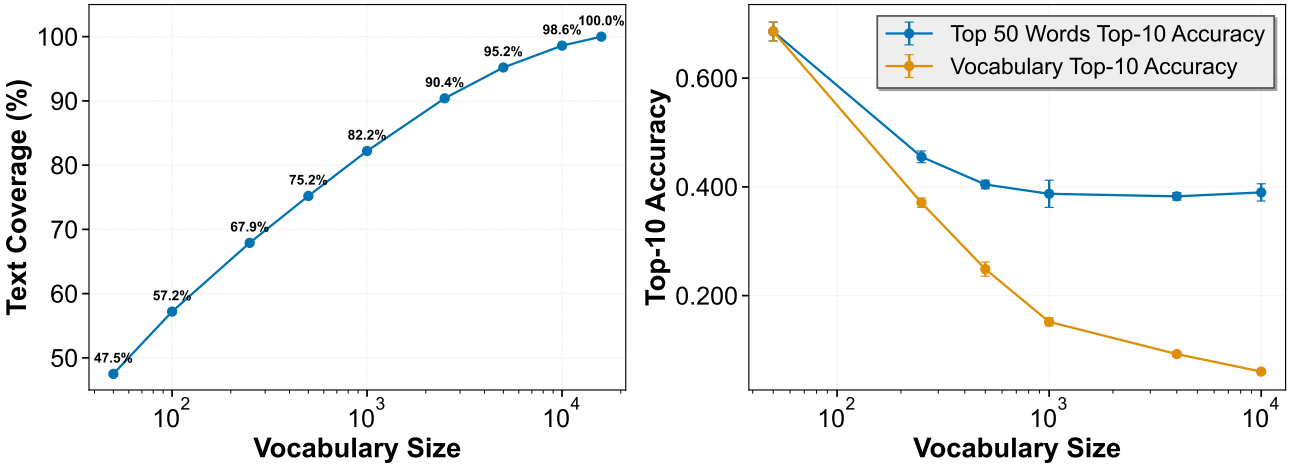

Figure 8: **Scaling vocabulary size on the LibriBrain dataset.** The left plot shows how much of the story a vocabulary of a certain size will cover. The right plot shows the top-10 word classification accuracy as we scale the vocabulary. The plot on the right supports a similar result seen by d'Ascoli et al. (2024, Figure 7) in which scaling vocabulary size decreases accuracy.

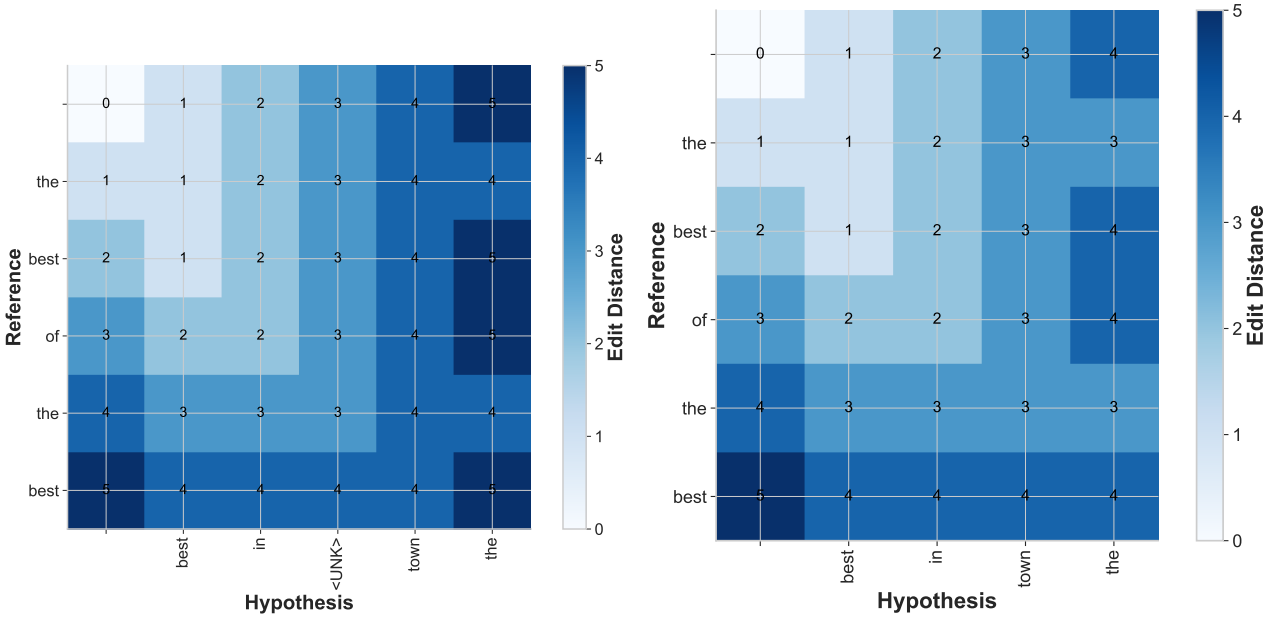

Figure 9: **Edit distance matrix with and without <UNK> for a toy example.**

F. Improving Evaluation With Missing Words

Excluding any text in place of out-of-vocabulary words can lead to artificial improvements in word error rates. This is caused by coincidental alignments when words are intentionally excluded. The following is a toy example where we include or exclude <UNK> at the position of an OOV word.

True sequence:
the best of the best

Prediction:
best in <UNK> town the
Required edits: 5

Prediction (no <UNK>):
best in town the
Required edits: 4

In the example, the same prediction but without an <UNK> token in place of the OOV position leads to a lower word error rate. This is because the lack of in-filling leads to a coincidental alignment of the word "the" (Figure 9). This implies an undesirable property in our evaluation: sentences with words filled in will not necessarily be *at least as good* as sentences without any words filled in.

To address this, we always insert <UNK> tokens in OOV positions when no in-filling method is used. However, this can have its own side-effects, potentially affecting metrics other than WER. To ensure we have a fair evaluation, we also include in-filling with randomly selected words as a baseline in all of our method ablations.

G. Additional Experiment Details and Hyperparameters

Statistical testing For all experiments, unless otherwise stated, we run 5 random seeds. Statistical tests are independent Welch's t-tests with the standard threshold for statistical significance ($p < .05$).

Preprocessing We first notch filter the data around 50Hz to remove powerline artifacts. Then, we band-pass filter in

the range 0.1-40Hz and resample the data to 50Hz. Following this preprocessing, we scale each recording such that the amplitude range $[-1, 1]$ covers the interquartile range and then clamp outliers above and below an amplitude of 5. Lastly, we apply a baseline correction to each individual sample, subtracting the mean of the sample calculated from the first 0.5s.

Extracting targets We optimise the transformer output embeddings for the target embeddings from the 12th layer of a pre-trained large T5 LLM. When words consist of multiple tokens, they are averaged into a single target embedding.

In-filling during rescoring Although a nested beam search allows the in-filling method to guarantee complete word predictions, in practice, the highest probability predictions are generally singular tokens. Thus, to reduce inference time, we discard the nested beam search and use a single-token prediction instead.

In-context transcription In our prompt (Appendix L Box 2), we provide the probabilities of the top five predictions. Given a large retrieval set, these probabilities can be rather similar. To make these differences easier for an LLM to reason about, we sharpen the probabilities we provide by applying the softmax operator

$$\text{softmax}(z_i) = \frac{e^{z_i/0.01}}{\sum_{j=1}^{n} e^{z_j/0.01}}, \tag{10}$$

over the top five probabilities with a temperature of 0.01. We selected this temperature as it provided much better separation amongst the top probabilities.

In both LLM-based methods, we instruct the LLM to output an enumerated dictionary in the format $\{\texttt{1:}\ \ w_1\texttt{,}\ \texttt{2:}\ \ w_2\texttt{,}\ \ldots\texttt{,}\ \ n\texttt{:}\ \ w_n\}$. We find that this constrained format leads to more robust responses with the correct number of outputs n, which we attribute to the enumerated keys. Additionally, LLM methods can use modes that allow self-reflective chain-of-thought to deeply reason. We leverage this in our experiments using Claude Sonnet 3.7 thinking (Anthropic). Our prompts are in Appendix L.

Compute resources Training a model on a single dataset takes up to 12 hours on an NVIDIA V100 GPU with 32GiB of GPU memory. We use 64GiB of system memory and require up to 200GiB of local storage. Generating sequences using beam search and other post-processing methods takes up to 4 hours. We ran all experiments on an internal cluster and estimate that overall, we used on the order of 1000 GPU hours. This includes preliminary experiments that did not make it into the paper.

H. OOV Prediction

After training our encoding model, we extract a feature vector for each encoder output from the training set containing: (1) the output probability distribution, computed via a softmax over the cosine similarities and (2) additional statistics calculated from the probability distribution (Table 9). We record whether each feature vector is associated with an OOV position or not and fit a binary classification XGBoost model (Chen & Guestrin, 2016) to the features. At test time, we use this model to select which words should be in-filled. This OOV detection method achieves 88% AUROC on LibriBrain. Including the additional statistics as features leads to a small but meaningful improvement of 1-2%.

An alternative method would be to train a binary OOV classification head on top of the transformer outputs. While also straightforward, this method entails re-training word classification models with OOV detection. Separating these concerns simplifies the process and allows OOV detectors to be reused with models trained on different seeds. Moreover, fitting an XGBoost model is much faster than training a word classifier from scratch.

I. Random Baselines

For completeness, we define two kinds of random baselines. The first, which we call *random selection*, is based on random chance given the vocabulary of the dataset. We generate our random selection baseline by uniformly randomly sampling from the decoder vocabulary for every in-vocabulary position and from the rest of the unique words in the dataset for every out-of-vocabulary position. This ensures a fair comparison for every in-vocabulary prediction and for every out-of-vocabulary in-filling step. The second random baseline we use is *random noise* and inspired by the findings in Jo et al. (2024). In this baseline, to test whether brain-to-text systems truly generalise or overfit to the stories in the dataset, we train a model as usual and then test it with random noise inputs of the same scale and variance as the MEG data. Jo et al. (2024) noticed that the models they tested this way failed to be any better than this baseline. Therefore, these models were

Table 8: **Hyperparameters.**

Hyperparameter	Value
Training	
Batch size / seq. length	64
Learning rate	1e-5
Optimiser	AdamW (Loshchilov & Hutter, 2019)
Annealing schedule	Cosine (min. 1e-6 after 50 epochs)
Early stopping patience	5 epochs
Early stopping metric	Val. top-10 word class. accuracy
Pooling	Spatial attention (Défossez et al., 2023)
Pooling output dim.	270
Signal encoder	Brain model (Défossez et al., 2023)
Init. conv.	1x1 (in: 270, out: 270)
Subject layer	in: 270, out: 270
Main conv. channels	(270, 320, 320, ... x10)
Kernel	3
Stride	1
Dilation growth	2
Groups	1
Dilation period	5
Context encoder	d'Ascoli et al. (2024)
Transformer depth	16
Transformer heads	16
Transformer dimension	1024
Transformer attn. dropout	0.1
Transformer pos. emb.	Rotary
Rescoring	
Beam search rescorer	Llama 3.2 1B
Rescorer weight (λ)	1.5
Beam search in-filler	Llama 3.2 1B
Max. words in context	8 words
Beam width	5 beams
OOV position detector	XGBoost (Chen & Guestrin, 2016)
Features	Probs. + Table 9
Learning rate	0.05
No. of estimators	200
Max. depth	4
Min. child weight	2
Subsample	0.8
Subsample ratio columns by tree	0.8
Gamma	1
α (L1)	0.1
λ (L2)	1
LLM APIs	
In-context in-filler	Claude 3.7 Sonnet w/ 4096 thinking tokens
In-context transcriber	Claude 3.7 Sonnet w/ 4096 thinking tokens

Table 9: **Additional descriptive features for OOV prediction.**

Feature	Description
Entropy	Information-theoretic uncertainty measure: $-\sum(p_i \times \log_2(p_i))$
Variance	Statistical measure of dispersion in the data
Mean	Average of all values
Median	Middle value when sorted
Max	Maximum value
Min	Minimum value
Skew	Measure of distribution asymmetry
Kurtosis	Measure of peakedness/tail heaviness
Gini	$1 - \sum p_i^2$; measures impurity/diversity
Top 1 prob.	Highest probability value
Top 2 prob.	Second highest probability value
Top 1 ratio	Ratio between highest and second highest: top1/top2
Peaks	Count of values above the mean
Zeros	Count of near-zero values ($< 10^{-10}$)
Nonzeros	Count of non-zero values ($> 10^{-10}$)
90th percentile	90th percentile value
10th percentile	10th percentile value
90th / 10th percentile ratio	Ratio between 90th and 10th percentiles
Top 5 sum	Sum of five highest probability values

not learning to use the brain data.

J. Limitations

As we establish the first above-chance electrophysiological non-invasive brain-to-text method, there remain some limitations which we hope future work will resolve to drive progress towards clinical application. Like the rest of the community, we use data from heard speech stimuli and not other forms of speech. The leap from perceived to inner or attempted speech is one that remains to be made among non-invasive approaches. Nevertheless, we expect the methods here to generalise to other varieties of speech.

Our dataset pooling framework improves performance through joint training with high-quality datasets. However, we have not shown the ability to improve overall performance across all datasets. Presently, this requires the collection of more and varied speech data. Still, jointly training EEG alongside MEG may make the decoding of speech from EEG, which does not require magnetic shielding, more feasible, opening the possibility of future mobile and low-cost speech BCIs.

Additionally, we rely on aligning neural data to word onsets. Although we make strides towards alignment-free decoding (Appendix B), further work is needed. However, decoding without alignment may not be necessary in a speech BCI as patients can space out the words that they attempt to speak (e.g. Moses et al. (2021)) to naturally segment neural data by word onset.

The last and most important limitation is overall decoding performance. We outperform all existing methods and establish a significant baseline, but we do not see the present results as strong enough for a clinical speech BCI. Our view is that this is firstly a data problem, requiring scaling up data and collecting more datasets. Even though our scaling laws suggest that to achieve high word accuracy, much more data must be collected (Appendix C), we encourage finding methods that improve scaling efficiency and also propose re-evaluating what it means to achieve a realistic non-invasive speech BCI. The focus should be on practical necessity over exact decoding accuracy. While scaling up data $100\times$ may not bring non-invasive performance near invasive speech decoding word error rates, it could be enough to enable semantic decoding to a level where non-invasive speech BCIs are useful. As the barriers to surgical BCIs are so large, non-invasive BCIs can be clinically practical without achieving similar levels of performance.

K. Broader Impacts

Although currently preliminary, maturity of research on non-invasive brain-to-text technology presents significant implications. Potential benefits include enabling communication for individuals with paralysis and advancing brain-computer interfaces generally. However, mature iterations of this technology raise concerns about brain data privacy, potential misuse of neural information, and accessibility disparities that may exacerbate existing inequalities. At present, we limit potential misuse by decoding heard speech rather than inner speech. As the field advances to inner speech, we encourage researchers to engage with ethical principles and set common ethical standards.

L. Prompts

Box 1: In-Context In-Filling

Prompt
I have a noisy speech recognition system which predicts 64 words at a time. I am going to give you its predictions in an ordered list. [UNK] indicates that the target word for that position is out-of-vocabulary.

I want you to fill in any [UNK] positions with words that you think fit well in the sequence. Do not replace anything that is not [UNK]. Your output should be formatted as a Python dictionary mapping all 64 positions (0-indexed) to words, preserving the system's predictions and replacing any [UNK] with your suggestions. Do not output anything else.

Output example:
{0: "don't", 1: "the", 2: "scowl", ..., 63: "if"}

Predictions from speech recognition system:

0: the
1: ¡UNK¿
2: sat
3: on
...
63: built

Answer
{0: "the", 1: "cat", 2: "sat", 3: "on", ..., 63: "build"}

Box 2: In-Context Transcription

Prompt
I have a noisy speech recognition system which predicts 64 words at a time. I am going to give you its predictions in an ordered list of pairs (word, probability). For each position, I give you the top-5 word predictions ordered from most likely to least likely along with their probabilities. [UNK] indicates that the target word for that position is out-of-vocabulary.

I want you to predict the most likely sequence from this information, picking the words from the predictions for each position that go best together. Where there is an [UNK] I want you to replace it with your own prediction for a word that fits well. In places with no [UNK], your job is to just pick the best fitting word from the predictions list (do not use any other word). Your output should be formatted as a Python dictionary mapping all 64 positions (0-indexed) to words. Do not output anything else.

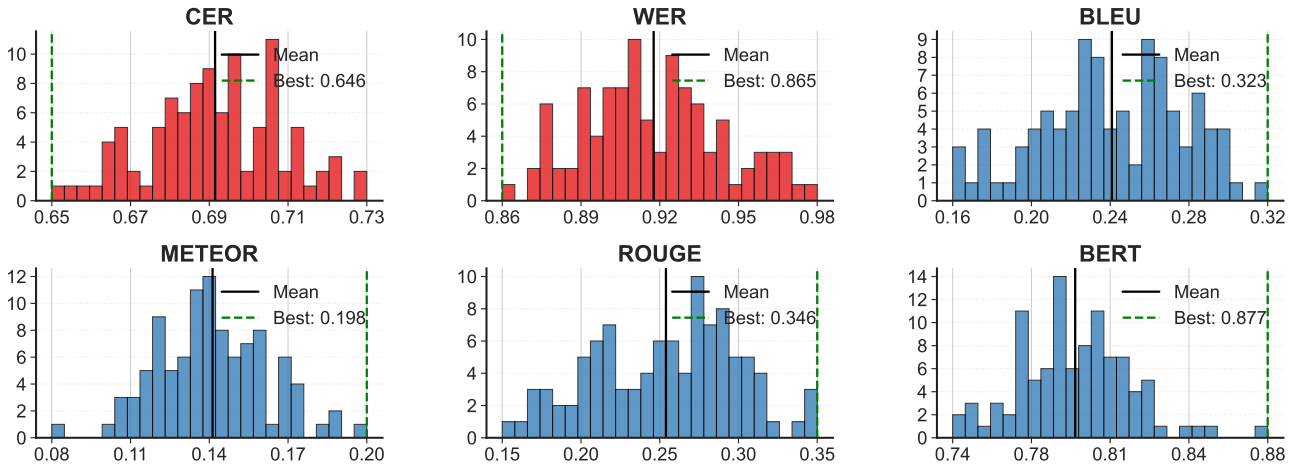

Figure 10: **Score histograms.** The scores are specific to the test set of LibriBrain for the experimental setup described in Appendix M.

Output example:
{0: ”don’t”, 1: ”the”, 2: ”scowl”, ..., 63: ”if”}

Predictions from speech recognition system:

0: (the, 0.45), (he, 0.23), (she, 0.15), (i, 0.11), (car, 0.09)
1: ¡UNK¿
2: (house, 0.15), (inn, 0.13), (in, 0.09), (new, 0.05), (nought, 0.01)
...
63: (built, 0.35), (with, 0.20), (love, 0.15), (of, 0.17), (my, 0.05)

Answer
{0: ”the”, 1: ”cat”, 2: ”in”, ..., 63: ”love”}

M. Decoding Examples

Examples in the table that follows are all generated using the rescoring and in-filling during beam search approach (beam+fill) from a single random seed on the LibriBrain dataset with a 250-word retrieval set. The table shows the best, median, and worst sequence for each metric. We also show the distributions of the scores for these predictions in Figure 10.

Annotations have been automatically added to the examples table. **Exact matches** are computed from a short exact match window where words are in similar locations between the stimulus and prediction. Similar n-grams are computed over words longer than two letters and where the `SequenceMatcher` from Python’s `difflib` gives a word sequence similarity ratio greater than 0.9. Finally, *semantically similar* words are matched with SpaCy similarity with a threshold of 0.9. Words that match this way are annotated with a number to show their correspondence. We did not find any clear way to highlight semantic similarity between phrases when this is present.

Stimulus	Prediction	Score
	Metric: WER	
	Best sequence	

at the thought of his own impotence what was that in the silence he *heard*[1] a gentle scratching sound low but very distinct in the quiet **of the** night it *came*[2] from **the door of the house** ferrier crept into the hall and *listened*[3] intently there was a pause for a few moments and then the low insidious sound was repeated someone was evidently	could be seen with our own eyes it is gone from the face that is the face **of the** earth face *turned*[1] towards to the north **of the house** and *turned*[1] out **the door of the house** to find out his father *said*[3] he was he was the son of a young man and that he had been in their employ for three years	0.7969
Median sequence		
the saints worn and exhausted he leaned upon his rifle and shook his gaunt hand fiercely at the silent widespread city beneath him as he looked at it he observed that there were flags in some of the *principal*[1] streets and other signs of festivity he was still speculating as to what this might mean when he heard the clatter of horse's hoofs and	of all the good people i know through my work with the his panic *business*[1] and with the his panic *business*[1] of us when we thought to us that we were we were going to get in the car and and so on for the it is not possible why many about these little things that we take a lot of for granted until	0.9062
Worst sequence		
led you safe to the chosen valley gave you a goodly share of land and allowed you to wax rich under our protection is not this so it is so *answered*[1] john ferrier in return for all this we asked but one condition that was that you should embrace the true faith and conform in every way to its usages this you promised to	question too broad in this case as the *question*[1] is to find who is in the best position over their opponents while a few time on my head without any problem and i had my mind being away from long time now yes i have been trying to get this since the whole end of this month and we are off	1.0000
Metric: CER		
Best sequence		
to keep to **the right track** for the moon had not yet risen **and the** high cliffs on either side made the obscurity more profound weighed down with his burden and weary from his exertions he stumbled along keeping up his heart by the reflection that every step brought him nearer to lucy and that he carried with him enough to ensure them food	to be on **the right track** or the wrong was not so bad **and the** right way of doing it about an hour or so ago out in my garden she was by my side but she was not in the mood to be nice to any one which will be of help and it will be of them before the lord from the	0.5959
Median sequence		
with bated breath lest something which *fell*[1] from their lips might be misconstrued and bring down a swift retribution upon them the victims of persecution had now turned persecutors on their own account and persecutors **of the** most terrible description not the inquisition of seville nor the german vehmgericht nor the secret societies of italy were ever able to put a more formidable machinery	*cried*[1] out in the morning having been out of bed about this hour and half into the night and after her it was a dream that it would be to make my life and my house a great place for both their children and their parents at the time **of the** birth **of the** child being far more than one of two or three	0.6862
Worst sequence		

led you safe to the chosen valley gave you a goodly share of land and allowed you to wax rich under our protection is not this so it is so *answered*[1] john ferrier in return for all this we asked but one condition that was that you should embrace the true faith and conform in every way to its usages this you promised to	question too broad in this case as the *question*[1] is to find who is in the best position over their opponents while a few time on my head without any problem and i had my mind being away from long time now yes i have been trying to get this since the whole end of this month and we are off	0.7547
Metric: BLEU		
Best sequence		
at the thought of his own impotence what was that in the silence he *heard*[1] a gentle scratching sound low but very distinct in the quiet **of the** night it *came*[2] from **the door of the house** ferrier crept into the hall and *listened*[3] intently there was a pause for a few moments and then the low insidious sound was repeated someone was evidently	could be seen with our own eyes it is gone from the face that is the face **of the** earth face *turned*[1] towards to the north **of the house** and *turned*[1] out **the door of the house** to find out his father *said*[3] he was he was the son of a young man and that he had been in their employ for three years	0.4062
Median sequence		
puzzled john ferrier sorely for his servants slept in an outhouse and the doors and windows had all been secured he crumpled the paper up and said **nothing to** his daughter but the incident struck a chill into his heart the twenty nine days were evidently the balance **of the** month which young had promised what strength or courage could avail against an enemy	question **of the** day there was a time in my life where i really had **nothing to** make any money these days the first thing is come clear from my mind and i can see the world before or after i go to sleep i have a dream that the world should open their eyes so that the whole **of the** back **of the**	0.2344
Worst sequence		
eked out by such employment as he could *pick*[1] up he travelled from town to town through the united states in quest of his enemies year passed into year his black hair turned grizzled but still he wandered on a human bloodhound with his mind wholly set upon the one object upon which he had devoted his life at last his perseverance was rewarded	the thing so far is that we do not doubt that the will of the people under the constitution of any state to be a state of their own there will be nothing for your course of action end of the *day*[1] in this case about the same time which seemed to have been and again to him this *day*[1] by the	0.1249
Metric: METEOR		
Best sequence		
i have not married ferrier answered but women were few and there were many who had better claims than i i was not a lonely man i had my daughter to *attend*[1] to my wants it is of that daughter that i would speak to you said the leader **of the** mormons she has *grown*[2] **to be the** flower of utah and has found	one of my best friends eyes my eyes my mind and i can see that in my mind of him i don't like her thing but when he was the one in my life i thought was young and over the whole thing with me from the beginning **of the** *year*[1] then this is **to be the** *year*[1] for you and many *years*[2]	0.2389
Median sequence		

suspense was unnerving he concealed his fears from his *daughter*[1] however and affected to make light of the whole *matter*[2] though she **with the** keen *eye*[3] of love saw plainly that he was ill at ease he expected that he would receive some message or remonstrance from young as to his conduct and he was not mistaken though it *came*[4] in an unlooked for	question of whether the patient will be **with the** same *face*[3] and body to go out **with the** young *man*[1] she saw over the counter at the pharmacy *said*[4] that we can get our medication this *way*[2] is the most effective without causing or worsening and also that in other cases where it can be done while i find a few more for	0.1408
Worst sequence		
sacred council of four the *girl*[1] is young and we would not have her wed grey hairs neither would we deprive her of all choice we elders have many heifers1 but our children must also be provided stangerson has a son and drebber has a son and either of them would gladly welcome your daughter to their house let her choose between them they	question to myself i have no doubt that i shall go back to it again and again through the years over the same period since is quite large and can be all found in one place from the comfort and safety that will be so important to me any time i had to over his head it is not *black*[1] enough he	0.0862
Metric: ROUGE		
Best sequence		
at the thought of his own impotence what was that in the silence he *heard*[1] a gentle scratching sound low but very distinct in the quiet **of the** night it *came*[2] from **the door of the house** ferrier crept into the hall and *listened*[3] intently there was a pause for a few moments and then the low insidious sound was repeated someone was evidently	could be seen with our own eyes it is gone from the face that is the face **of the** earth face *turned*[1] towards to the north **of the house** and *turned*[1] out **the door of the house** to find out his father *said*[3] he was he was the son of a young man and that he had been in their employ for three years	0.4062
Median sequence		
when he *heard*[1] the click of the latch and looking through the window saw a stout sandy haired middle aged *man*[2] coming up the pathway his heart leapt to his mouth for this was none other than the great brigham young himself full of trepidation for he knew that such a visit boded him little *good*[3] ferrier ran to **the door** to greet the	when i *took*[1] the test in the morning after my out **the door** without a shower and a change of face cloth in the morning the first *thing*[3] to be done no course is a doubt and the same is between us and our children and we must have made a mess of this last point of mine on the *death*[2] of my own	0.2500
Worst sequence		
as keen as on that memorable *night*[1] when he had stood by john ferrier's grave disguised and under an assumed name he *returned*[2] to salt lake city careless what became of his own *life*[3] as long as he obtained what he knew to be justice there he found evil tidings awaiting him there had been a schism among the chosen people a few months	own ing that they are in business but they have done some of the best work here under the sun here we are in the middle of summer by now as we my mind and *brought*[2] up the subject with my head in my hands until i still don 't have them i will find a *way*[1] of making it in this *world*[3]	0.1406
Metric: BERT		
Best sequence		

that his wealth and position would be of no avail to him others as well known and as rich as himself had been spirited away before now and their goods given over **to the** church he was a brave man but he trembled at the vague shadowy terrors which hung over him any known danger he could face with a *firm*[1] lip but this	you can use the same as what have been used for us to in any way to get in the door but those are about between them and they are not open **to the** *public*[1] there is a small man who has been at my side since i came **to the**se hands so long ago i am seen as a young man and i	0.8256
Median sequence		
vindictiveness far from doing so it had if anything augmented it the hunter's mind was of a hard unyielding nature **and the** predominant *idea*[1] of revenge had taken such complete possession of it that there was no room for any other emotion he was however above all *things*[2] practical he soon *realized*[3] that even his iron constitution could not st**and the** incessant strain which	question here from the *point*[1] we *saw*[3] in most of case to be case to be the case to be **and the** be the *like*[4] of which is such as is in him that he can not see any light before him the whole room was his own and all he had to *end*[2] his own life and went to the police station where	0.8024
Worst sequence		
polygamy without a female population on which to draw was a barren doctrine indeed strange rumours began to be bandied about rumours of murdered immigrants and rifled camps in regions where indians had never been seen fresh women appeared **in the** harems **of the** elders women who pined and wept and bore upon their faces the traces of an unextinguishable horror belated wanderers upon	question about the use of i don't have time from the beginning **of the** day to the **in the** end end to an end to and end to with and end to his own end **of the** day the same day **of the** week as the previous business day has closed out its business for today with an average of is	0.7696