REMOVING BRAINLESS SHORTCUTS IMPROVES NON-INVASIVE BRAIN-TO-TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

We find that major reported improvements in decoding words from non-invasive brain recordings are largely reproducible without any brain data. In the influential work of d'Ascoli et al. (2025), time series of brain activity from subjects perceiving continuous speech are segmented into fixed-length windows starting at each word. A neural network then generates predictions for all of the words in a sentence together. Neighbouring windows partially overlap, implicitly revealing the interval between words. Since these intervals indicate the duration of the words spoken, and different words tend to take longer—for example, "the" is much shorter than "supercalifragilisticexpialidocious"—the neural network can improve its predictions of words without relying on the underlying brain activity. Consistent with this, **the method reaches 22.0% balanced accuracy on synthetic signals containing no brain information, compared with 22.3% on real brain recordings**. To prevent the network from learning this shortcut, we make a single, simple change. Instead of jointly encoding all windows in a sentence, we process each independently. As a result, the neural network achieves better performance by learning underlying word-specific information from brain recordings. This makes two existing strategies become much more effective than before. Both aggregating predictions from distinct neural responses to the same word and using a pretrained LLM as a linguistic prior now substantially improve results. On our clinically motivated perceived speech benchmark, this simple recipe (*SimpleB2T*) achieves a **word error rate of 36.6%** with five observations per word, approaching past invasive speech decoding performance, albeit under different conditions. The results in this work expose an important shortcut in brain-to-text decoding and show that removing it leads to a simple and considerably more effective strategy.

1 INTRODUCTION

Restoring communication to people who have lost the ability to speak by decoding brain activity into speech is a central goal of brain–computer interface (BCI) research. Non-invasive approaches based on magnetoencephalography (MEG) or electroencephalography (EEG) offer a safe route towards this goal, but must recover speech information from weak and spatially blurred signals measured outside the brain. Nevertheless, the state of the art in the field has progressed from detecting the presence of speech (Dash et al., 2020), to matching audio with corresponding neural responses (Défossez et al., 2023), and more recently to decoding individual words (d'Ascoli et al., 2025).

In this paper, we study *word-aligned brain-to-text (B2T)* from perceived speech in non-invasive brain recordings. Here, a subject listens to speech (e.g. from an audiobook) or reads text while their brain activity is recorded with M/EEG and the task is to decode the words they perceived from their brain activity, knowing only when they perceived each word. Perceived speech is often used as a stepping stone towards the grander goal of decoding internal speech, such as inner monologues, because it provides stronger and more easily aligned neural responses (Martin et al., 2014). Thus, perceived speech acts as a common test-bed for developing non-invasive speech decoding methods (Défossez et al., 2023; Tang et al., 2023; Özdogan et al., 2025; Mantegna et al., 2026a).

One influential recent direction in word-aligned B2T, introduced by d'Ascoli et al. (2025), has been to jointly decode neural responses to all words in a sentence from a continuous M/EEG time series. In this setup, d'Ascoli et al. extract a fixed-length window from the continuous recording at each

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

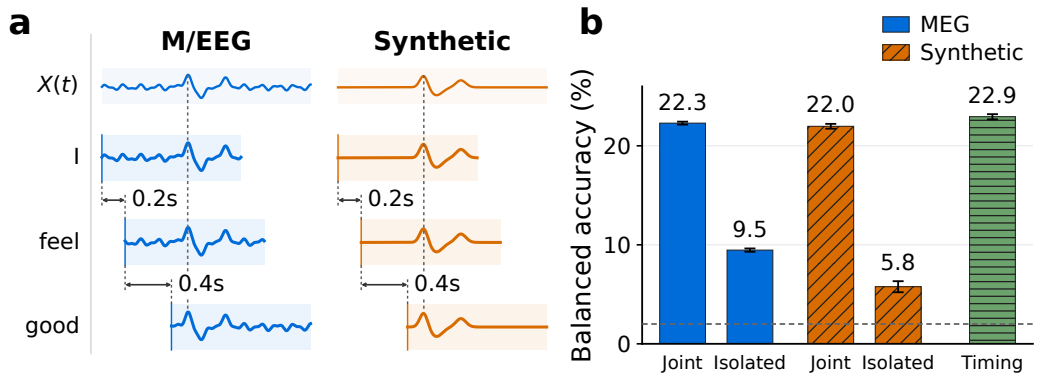


Figure 1: **Overlapping word-aligned windows leak word duration.** **(a)** Overlapping word-aligned windows contain shared samples whose relative offsets reveal the duration of words. **(b)** To test whether jointly decoding words uses neural information, we compare MEG with synthetic signals having no relation to brain activity. With MEG, *isolated* means we decode each window independently; with synthetic data, it means we make windows independent. *Timing* tests what happens when the same model receives an encoding of the interval between words directly. Jointly decoding words reaches nearly the same accuracy on synthetic and real inputs, whereas removing the shortcut substantially reduces performance. Error bars indicate $\pm 1\sigma$ across training seeds.

word onset and then jointly encode all of the windows in a sentence with a neural network, predicting all of the words in the sentence at once. Jointly decoding all of the words in the sentence this way improves word classification by an average of 50% compared with decoding each window independently (d'Ascoli et al., 2025). The setup has since been extended, evaluated, and incorporated into a series of subsequent non-invasive brain-to-text studies and benchmarks (Jayalath & Parker Jones, 2026; Wang et al., 2026; Li et al., 2026; Jayalath et al., 2025a; Özdogan et al., 2025; Mantegna et al., 2026a; Landau et al., 2026; Banville et al., 2026; Jayalath et al., 2026). However, we identify an important property of this setup. Neighbouring fixed-length windows typically overlap as the windows are longer than the individual words' durations. The same signal samples therefore appear at different relative positions in adjacent inputs, revealing the interval between word onsets and indicating the duration of words. Since typical word duration differs across words, the neural network can use this timing information to narrow the set of plausible words and improve its predictions, without using brain activity. This is an instance of *shortcut learning* (Geirhos et al., 2020), where a model can exploit an unintended decision rule that does not transfer to the intended use of the model.

We find that this effect is large enough to reproduce almost the entire gain from jointly decoding words without using brain activity at all (Figure 1). On real MEG, jointly decoding words reaches 22.3% balanced word accuracy, compared with 9.5% when words are decoded independently. When we replace the MEG with a synthetic continuous signal that contains no information about the stimulus but preserves the same window overlap, the decoder reaches 22.0%. Removing the overlap structure reduces accuracy to 5.8%. Thus, most of the apparent benefit of jointly decoding words can be recovered from the information exposed by overlapping inputs alone.

We explore the simplest solution to this problem: decode words individually instead of jointly so that the neural network can never learn this shortcut. We otherwise retain the setup of d'Ascoli et al. (2025). Once words are decoded independently, two techniques that previously provided limited improvements become much more effective. Firstly, multiple occurrences of distinct neural responses to the same stimulus are commonly used in non-invasive BCIs to improve signal-to-noise ratio (Farwell & Donchin, 1988), and we apply the same principle here by combining predictions from multiple observations of the same word. Secondly, invasive speech BCIs routinely combine neural predictions with explicit linguistic priors by leveraging LLMs while non-invasive methods have so far struggled to do the same (Jayalath et al., 2025a). We find that predictions driven partly by timing are poorly suited to aggregation or combination with a linguistic prior, whereas isolated neural predictions provide information that can be combined effectively with both. Compared with jointly decoding words, both strategies now yield much larger benefits from only a few aggregated observations. On a core subset of clinically motivated sentences, this approach, which we refer to as *SimpleB2T*, achieves 65.6% WER with one observation per word and 36.6% with five observations.

Our work makes three points: **(1)** major improvements in non-invasive word-aligned B2T are largely reproducible without brain information; **(2)** we establish controls that expose the source of this shortcut as overlapping word-aligned windows; and **(3)** decoding words independently provides a simple fix that removes the shortcut, forcing the model to use information from brain activity and making aggregation of predictions and combination with a linguistic prior much more effective.

2 Word Duration Leakage in Word-Aligned Decoding

We revisit the setup of d'Ascoli et al. (2025) and show how neighbouring, overlapping word-aligned windows reveal the duration of words to the decoder. We then test whether this non-neural information is sufficient to explain the large improvements from jointly decoding words.

2.1 Decoding Words Jointly from Neural Responses

We describe the task and decoder of d'Ascoli et al. (2025). In word-aligned B2T from perceived speech, the task is to decode the words a subject perceives through an auditory or visual stimulus from their recorded brain activity. The word-aligned nature of the task assumes that the onset time of each perceived word in the stimulus is known, but no further information is assumed.

Let $X(t) \in \mathbb{R}^C$ denote a continuous M/EEG recording with C channels and t_i the known onset of word w_i. For each word, the model extracts a three-second window beginning at its onset,

$$x_i = X[t_i : t_i + 3\text{s}] \in \mathbb{R}^{C \times L}, \tag{1}$$

where L is the number of time samples in three seconds. Windows $x_1, \ldots, x_n$ from all words $w_1, \ldots, w_n$ in a sentence are then jointly encoded by a neural network into representations $\hat{z}_1, \ldots, \hat{z}_n$ where these representations are semantic word embeddings. For each word w, a target embedding e_w is obtained by encoding w with T5-large (Raffel et al., 2020). The neural network is trained with a contrastive objective which encourages the predicted representation $\hat{z}_i$ to be similar to the target embedding e_{w_i} and dissimilar to embeddings of other words in the batch.

At test time, the model predicts a word by nearest-neighbour retrieval. Given a candidate vocabulary $\mathcal{V}$, each predicted embedding is compared with the corresponding frozen T5 embeddings using cosine similarity, and the highest-scoring candidate is selected:

$$\hat{w}_i = \arg\max_{w \in \mathcal{V}} \cos(\hat{z}_i, e_w). \tag{2}$$

Therefore, the neural network predicts a point in the T5 embedding space and retrieves the nearest candidate word at evaluation. Jointly mapping all inputs $x_1, \ldots, x_n$ for a sentence to $\hat{z}_1, \ldots, \hat{z}_n$, instead of mapping each word-aligned window x_i to $\hat{z}_i$ individually, yields a 50% improvement on average in standard evaluations with continuous brain recordings (d'Ascoli et al., 2025).

2.2 Overlapping Windows Reveal Word Duration

An important property of this construction is that consecutive word windows are not independent. Natural speech contains several words within three seconds, so adjacent windows overlap:

$$x_i = X[t_i : t_i + 3\text{s}], \qquad x_{i+1} = X[t_{i+1} : t_{i+1} + 3\text{s}], \tag{3}$$

where t_{i+1} is often much less than $t_i + 3$s. The same underlying samples therefore occur in neighbouring inputs at a relative displacement of $t_{i+1} - t_i$. Applied to our data, 99.9996% of adjacent word pairs have overlapping windows and these share 90.6% of their samples on average. Moreover, the onset-to-onset interval correlates strongly with word duration ($r = 0.90$; Appendix A).

Indeed, this interval can be recovered directly from two adjacent inputs. In discrete time, let d denote the number of samples between their onsets. Because the windows are extracted from the same continuous recording, $x_i[n + d] = x_{i+1}[n]$ throughout their overlapping region. Hence

$$d = \arg\min_{\ell} \sum_{c} \sum_{n} \left(x_i^c[n + \ell] - x_{i+1}^c[n] \right)^2, \tag{4}$$

where c is the sensor channel, up to preprocessing and repeated signal patterns. Thus, recovering the interval between words (and thereby word duration) requires only matching the shared samples between neighbouring windows and is trivially solvable by minimising a sum of squared differences.

This raises a simple question: how much of the reported improvement from jointly decoding words comes from neural information, and how much can be recovered from the information exposed by overlapping windows?

2.3 Synthetic Signals Reproduce Decoding Improvements

We test whether improvements from jointly decoding words require neural information. The concern is that fixed-length windows for neighbouring words overlap, so decoding words jointly may use word durations revealed by shared signal samples. We therefore construct controls that preserve or remove this overlap, allowing us to test both whether overlap alone can reproduce this gain and whether jointly decoding words remains useful when it is removed.

Experimental setup. We train variants of d'Ascoli et al.'s decoder and evaluate them on subject 0 of LibriBrain100 (Mantegna et al., 2026b), the largest publicly available neural speech decoding dataset at the time of writing. In this dataset, the subject listened to around 80 hours of auditory stimuli from audiobooks, podcasts, and random spoken sentences while their brain activity was recorded with MEG. We split the dataset such that no stimuli overlap between splits and all examples from a session belong exclusively to training, validation, or test. We report balanced top-1 word accuracy over the fifty most frequent words in the dataset. Exact splits are in Appendix B and additional experimental details are in Appendix C.

We compare five conditions. *Joint (MEG)* uses the full decoder described by d'Ascoli et al. (2025) on real MEG data. *Isolated (MEG)* is similar, except it processes each word window independently in isolation, providing a reference for the gain from jointly decoding words. To test whether this gain requires neural information, in the *Joint (synthetic)* control, we replace MEG with a continuous synthetic signal that is unrelated to the stimulus, but extract windows at the same true word onsets. Adjacent synthetic windows therefore share the same underlying samples at offsets determined by the original word onsets. In the *Isolated (synthetic)* control, we generate a separate synthetic signal for each word window rather than extracting all windows from one continuous signal. Neighbouring inputs therefore share no samples, so their relative positions no longer reveal the true intervals between words. Comparing the synthetic joint and synthetic isolated conditions shows the information provided by overlapping windows. Lastly, in the *Timing* condition, we encode the log interval between words and supply only this to the same model.

Results. Figure 1 shows that the gain from jointly decoding words can be reproduced almost entirely without neural information. On MEG, jointly decoding words increases word accuracy from 9.5% to 22.3%. A synthetic signal with the same overlap structure reaches 22.0%, despite containing no brain activity. Both reach similar accuracy to the model trained directly on word intervals as inputs (22.9%). When this overlap is removed, accuracy falls to 5.8%. We reproduce this effect across a further 32 subjects (Appendix D). We also find that predictions from jointly decoding words align more strongly with duration-based word probabilities than isolated predictions (Appendix E). As an additional control, we retrain and evaluate on non-overlapping natural sentence constructions, preserving the original word sequences while removing shared samples. This does not recover an advantage over isolated word decoding (Appendix F).

These results do not imply that jointly decoding words is inherently unhelpful. Such decoding could still be useful in this setting when neighbouring responses can be modelled jointly without exposing word timing, for example when words are sufficiently separated in time to avoid overlap. In naturalistic speech, however, avoiding this timing information in word-aligned B2T is difficult because truncating or masking windows at word boundaries can itself reveal the same timing information. In settings where word onsets are unknown, this kind of leakage is not an issue. However, such a setting represents a substantially different and more difficult task, which we do not evaluate here. In the word-aligned B2T setting, decoding words individually presents a way to force a neural network to predict words from brain activity without using this shortcut.

3 Decoding Independently Improves Sentence Reconstruction

Having shown that much of the gain from jointly decoding words can arise from non-neural word duration information, we next ask what happens when we decode words independently to see how

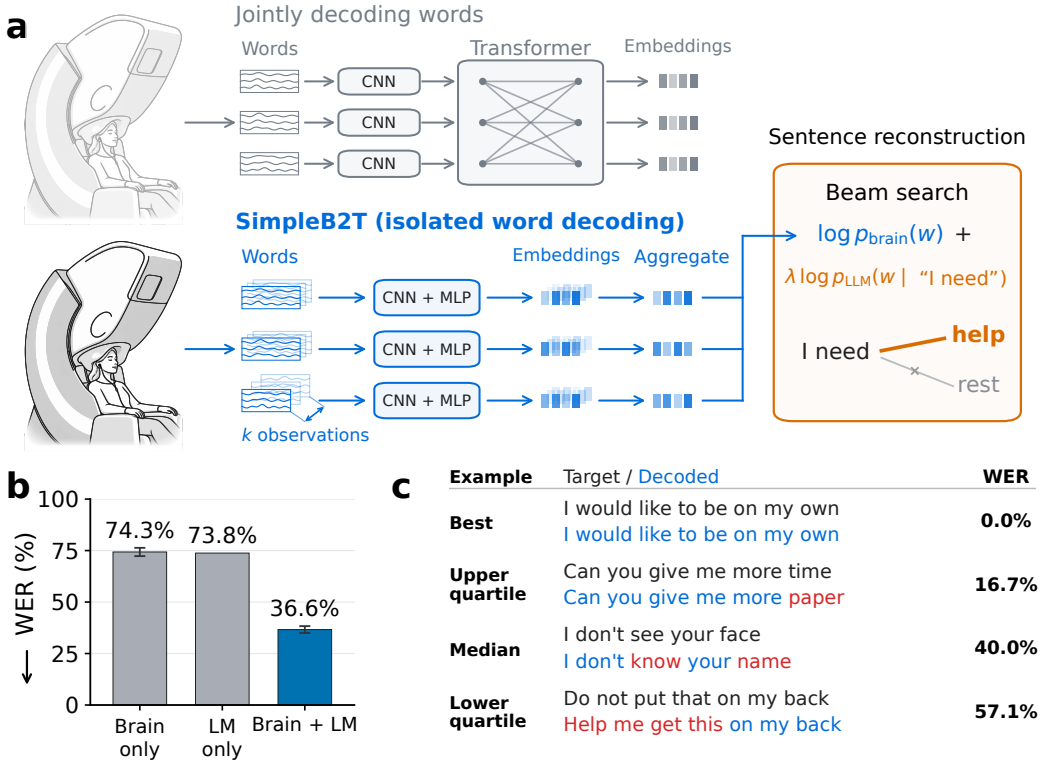


Figure 2: **SimpleB2T. (a)** Instead of jointly decoding words, we decode each neural response independently. We also test the effect of aggregating responses from distinct observations and using an LLM with beam search for sentence decoding. **(b)** Combining predictions with an LLM substantially reduces WER. **(c)** Decoded examples. Results are on our core set of 100 clinical sentences with $k = 5$ observations per word. Error bars indicate ± 1 standard deviation across five training seeds.

the neural network performs when it does not use the shortcut. We introduce a simple isolated word decoder and test whether its predictions can be more effectively used to reconstruct sentences, and whether these predictions can now benefit from established strategies (aggregating observations and using a language model prior). We test these questions on a clinically motivated communication benchmark and provide additional technical details of all experiments in Appendix C.

3.1 Decoding Words Independently With A Simple Neural Network

We retain the word-level encoder, semantic targets, and contrastive training framework of d'Ascoli et al. (2025), while decoding words independently.

Let x_i denote the three-second MEG window aligned to word w_i. A neural network f_θ maps this window to a unit-normalised embedding

$$\hat{z}_i = f_\theta(x_i). \tag{5}$$

Here, f_θ consists of a small CNN, whose output embeddings are temporally pooled, followed by a residual MLP that outputs $\hat{z}_i$. Effectively, f_θ is a simplification of the d'Ascoli et al. encoder, which processes all x_i in a sentence through the same small CNN with temporal pooling, and then jointly processes the pooled embeddings across a sentence with a large 16-layer, 16-head bidirectional transformer. The d'Ascoli et al. model has approximately 200 million parameters, while our simplification has about 20 million. Unlike d'Ascoli et al.'s encoder, f_θ has no access to the other windows in the sentence. A three-second window can nevertheless contain neural responses to subsequent words. We examine whether the predictability of this subsequent context is associated with decoding accuracy in Appendix G.

As before, f_θ is trained contrastively, following d'Ascoli et al. (2025), so that $\hat{z}_i$ is close to e_{w_i} and far from embeddings of other words. At test time, the cosine similarity $\hat{z}_i^\top e_w$ therefore measures the support for candidate word w.

For a candidate vocabulary of words $\mathcal{V}$, we convert these similarities into a distribution over words:

$$p_{\text{brain}}(w \mid x_i) = \frac{\exp(\hat{z}_i^\top e_w / T)}{\sum_{v \in \mathcal{V}} \exp(\hat{z}_i^\top e_v / T)}. \tag{6}$$

The temperature T is selected once on validation data and then fixed. This step turns the retrieval scores of the neural decoder into probabilities that can later be combined with a language model.

3.2 Strengthening Word Predictions

Since non-invasive neural responses have limited signal-to-noise ratio, we also test two simple ways to strengthen word predictions for sentence reconstruction. Firstly, when multiple responses to the same word are available, we aggregate their predictions. Secondly, we combine the resulting neural probabilities with an off-the-shelf language model, providing a linguistic prior to favour plausible word sequences. We refer to the recipe of independently decoding words, using aggregation, and leveraging an LLM prior as *SimpleB2T*. We provide further technical details in Appendix H.

Aggregating observations. When k distinct observations of the same word are available, we encode each independently and average their predicted embeddings,

$$m_i = \frac{1}{k} \sum_{j=1}^{k} \hat{z}_{i,j}, \qquad r_i = \|m_i\|_2, \qquad u_i = \frac{m_i}{r_i}. \tag{7}$$

As each predicted embedding is unit normalised, r_i measures agreement between observations and approaches one when their predictions point in similar directions and decreases when they disagree. For candidate word w, we define the aggregated neural score

$$A_i(w) = \frac{1 + \alpha r_i}{T} u_i^\top e_w, \tag{8}$$

where the agreement weight α controls how strongly agreement between observations influences the score. Normalising m_i to obtain the consensus direction u_i would discard this information, so we use r_i to modulate the sharpness of the neural scores.

LLM rescoring. For a candidate sentence $y = (w_1, \ldots, w_n)$, we combine the neural scores with the autoregressive likelihood assigned by a language model:

$$S(y) = \sum_{i=1}^{n} A_i(w_i) + \lambda \log p_{\text{LLM}}(y \mid q), \tag{9}$$

where q is a fixed prompt and λ controls the contribution of the language model. In this paper, p_{LLM} is the base model of Qwen3-8B (Qwen Team, 2025). We approximately maximise Equation (9) using left-to-right beam search over the candidate vocabulary. As in d'Ascoli et al. (2025), the number of word positions is known for each test sentence, so decoding terminates after that many words. This is part of the word-aligned B2T setting.

3.3 Experimental Setup

We largely follow the experimental setup described in Section 2.3. For this evaluation, however, validation and test decoding are performed over a vocabulary of 92 words on the sentences in our clinical communication benchmark (described next). We report WER and the proportion of sentences where all words are decoded correctly, which we call the *sentence match rate (SMR)*.

Clinical communication benchmark. Our primary evaluation is a benchmark of 200 clinically motivated communication sentences constructed from our held-out test recordings. The sentences are designed to represent simple patient communication needs such as requesting help, water, repositioning, environmental changes, or interaction with another person. They are divided into two

Method	k	Core WER (%) ↓	Core SMR (%) ↑	Expanded WER (%) ↓	Expanded SMR (%) ↑	Full WER (%) ↓	Full SMR (%) ↑
LM only	-	73.8	2.0	91.8	0.0	83.4	1.0
d'Ascoli	1	$98.9_{\pm 0.3}$	$0.0_{\pm 0.0}$	$98.8_{\pm 0.1}$	$0.0_{\pm 0.0}$	$98.8_{\pm 0.2}$	$0.0_{\pm 0.0}$
d'Ascoli + LM	1	$77.3_{\pm 1.1}$	$2.4_{\pm 0.3}$	$87.5_{\pm 0.8}$	$0.2_{\pm 0.1}$	$82.7_{\pm 0.2}$	$1.3_{\pm 0.2}$
SimpleB2T	1	$\mathbf{65.6_{\pm 0.8}}$	$\mathbf{6.4_{\pm 0.5}}$	$\mathbf{76.3_{\pm 0.5}}$	$\mathbf{4.2_{\pm 0.4}}$	$\mathbf{71.3_{\pm 0.5}}$	$\mathbf{5.3_{\pm 0.3}}$
d'Ascoli	5	$98.9_{\pm 0.3}$	$0.0_{\pm 0.0}$	$98.8_{\pm 0.4}$	$0.0_{\pm 0.0}$	$98.9_{\pm 0.1}$	$0.0_{\pm 0.0}$
d'Ascoli + LM	5	$79.4_{\pm 2.5}$	$2.4_{\pm 0.5}$	$85.5_{\pm 2.0}$	$0.2_{\pm 0.4}$	$82.7_{\pm 0.6}$	$1.3_{\pm 0.3}$
SimpleB2T	5	$\mathbf{36.6_{\pm 1.7}}$	$\mathbf{26.0_{\pm 2.5}}$	$\mathbf{44.8_{\pm 3.7}}$	$\mathbf{20.2_{\pm 4.3}}$	$\mathbf{41.0_{\pm 2.2}}$	$\mathbf{23.1_{\pm 3.0}}$

Table 1: **Effect of decoding words independently.** We compare the independent word decoder and d'Ascoli et al., with and without the same LLM rescoring. The table tests how using k observations and LLM rescoring behave when words are decoded jointly and when words are decoded independently. Results are means across five seeds and subscripts show sample standard deviations.

LM	WER (%)
Without	$74.3_{\pm 2.0}$
With	$\mathbf{36.6_{\pm 1.7}}$

(a) **LM contribution.** LM rescoring halves WER.

Input	WER (%)
LM only	73.8
Noise + LM	$97.1_{\pm 1.7}$
Shuff. brain + LM	$88.3_{\pm 1.9}$
Brain + LM	$\mathbf{36.6_{\pm 1.7}}$

(b) **Controls.** Results depend on example-specific information.

Prompt q	WER (%)
Generic	$55.2_{\pm 1.2}$
Task-oriented	$\mathbf{36.6_{\pm 1.7}}$

(c) **LM prompt.** Describing the task and domain reduces error.

Beam width	WER (%)
1	$67.7_{\pm 1.6}$
25	$37.9_{\pm 2.4}$
50	$\mathbf{36.6_{\pm 1.7}}$
200	$35.9_{\pm 2.5}$
500	$36.0_{\pm 2.7}$

(d) **Beam width.** WER plateaus beyond width 200.

LM weight λ	WER (%)
0	$74.3_{\pm 2.0}$
0.25	$40.8_{\pm 2.6}$
0.5	$\mathbf{36.6_{\pm 1.7}}$
1	$47.7_{\pm 1.3}$
2	$59.7_{\pm 0.8}$

(e) **LM weight.** LM evidence is most useful at half-weight.

Agreement wt. α	WER (%)
0	$49.3_{\pm 1.5}$
0.5	$42.7_{\pm 1.0}$
1	$40.7_{\pm 1.5}$
2	$\mathbf{36.6_{\pm 1.7}}$
4	$38.1_{\pm 2.6}$

(f) **Agreement weight.** Weight 2 gives the lowest WER.

Table 2: **Ablations.** Values selected on a development set of 50 sentences, except for (c) which is retrospective and did not inform q. Results on the core set of test sentences with $k = 5$. **Bold** indicates the selected value. Results are means across training seeds, with sample standard deviations.

100-sentence subsets: *Core*, containing short care-related utterances, and *Expanded*, containing a broader range of questions, statements, temporal expressions, and social interactions. These clinically motivated sentences are inspired by the task design of the surgical decoding work from Moses et al. (2021), who used a smaller 50-word vocabulary and 50 total sentences. This constrained communication domain also provides a stronger linguistic prior, helping an LLM resolve uncertain neural predictions into plausible patient messages. Each sentence position is constructed using a recorded occurrence of the corresponding word from the held-out test sessions. Consequently, neighbouring sentence positions come from distinct recording events and do not share overlapping signal samples, so the shortcut exploited by jointly decoding words is unavailable. To allow for aggregating neural responses, each sentence position is assigned up to five distinct occurrences of the same word. Occurrences are sampled without replacement and are not reused across benchmark positions. Thus, five occurrences correspond to five separately recorded MEG responses to the same word. A list of all sentences is provided in Appendix I.

3.4 Results

When decoding words jointly, aggregation provides little benefit, and LLM rescoring yields performance close to or worse than the LM-only baseline (Table 1). After decoding words independently, both become substantially more effective. A likely explanation on our benchmark is that the shortcut can no longer be exploited, so predictions from d'Ascoli et al.'s model provide no useful information. More generally, predictions from jointly decoding words may not have

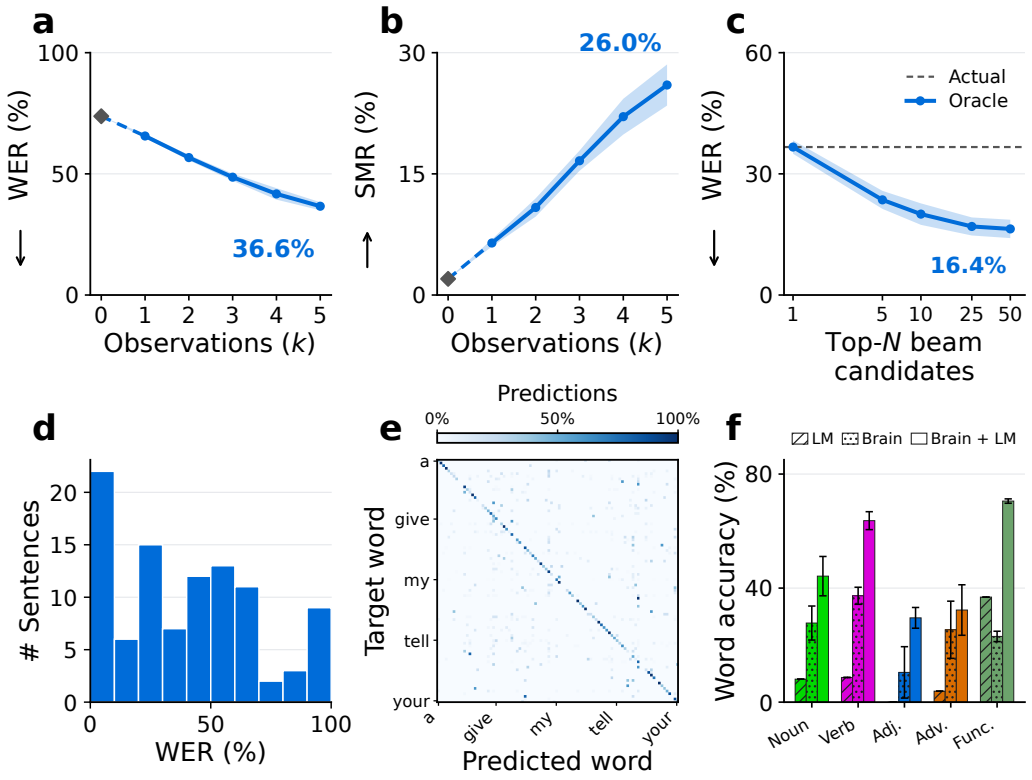


Figure 3: **SimpleB2T results. (a, b)** WER and exact sentence match as the number of observations increases. **(c)** Oracle selection among the top-N beam candidates. **(d)** Sentence WER distribution. **(e)** Confusion matrix across the 92-word vocabulary. **(f)** Word accuracy by part of speech for the LM, neural decoder, and their combination. Panels c–f use $k = 5$; all results are on the Core test set. Shading and error bars indicate ± 1 standard deviation across five training seeds.

benefitted from linguistic priors because the information output by a decoder that learns the shortcut is based on language statistics (i.e. the duration of the word), which is less complementary to an external LLM prior. We analyse a version of the model retrained without overlapping windows in Appendix J. By contrast, isolated word decoding must make each prediction from a single neural response, producing information from brain activity that can be combined more effectively across observations and with an LLM. At $k = 5$, LLM rescoring reduces WER from 74.3% to 36.6%, while increasing the number of observations from one to five reduces WER from 65.6% to 36.6%.

Aggregating observations provides a trade-off between measurement and accuracy. Figures 3a–b show that sentence decoding improves steadily as additional responses to the same word are combined, without retraining the neural decoder. Thus, collecting more observations provides a simple way to trade additional recording burden for more reliable predictions. We also note that the language-model weight selected on the development set decreases with k, consistent with the neural predictions becoming more informative as observations are combined (Appendix H).

Neural and linguistic information are complementary. At $k = 5$, the neural decoder obtains 74.3% WER and the LLM alone 73.8%. Combining them reduces WER to 36.6% (Table 2). This improvement depends on information specific to the corresponding brain response. Noise inputs matching the statistics of MEG give 97.1% WER and shuffling the predicted embeddings across test occurrences gives 88.3%. The complementarity also shows across parts of speech (Figure 3f), where neural evidence helps verbs while the LM contributes strongly to function words. Additional analyses are provided in Appendix K.

Significant remaining error is in candidate selection. At $k = 5$, the top-ranked sentence has 36.6% WER, but the lowest WER among the top N candidate sentences can be much lower, even

when N is small (Figure 3c). This suggests that improving candidate selection even a little could significantly improve WER.

A task-specific prompt strengthens the linguistic prior. Explicitly describing the patient communication setting in the prompt substantially improves decoding, indicating that tuning the prior to the domain is an important part of using linguistic priors effectively (Table 2c; Appendix L).

4 DISCUSSION

Much of the gain from jointly decoding words reported by d'Ascoli et al. (2025) can be reproduced without brain information. Improvements arise because overlapping word-aligned windows expose word durations, which are informative about word identity. Therefore, when providing external information, one should be careful not to enable artificial shortcuts. To avoid the shortcut here, we decode words independently instead of jointly and find that two established neural decoding techniques—combining multiple observations and using an explicit linguistic prior—become substantially more effective. When predictions are driven by timing, aggregating observations gives a better estimate of a word's typical duration, improving predictions through the same shortcut. Without the shortcut, aggregation instead improves the estimate of the word-specific neural signal. Similarly, this signal is more complementary to LLM priors than information from the shortcut.

Scope of the timing-leakage result. Since its publication in late 2025, the approach introduced by d'Ascoli et al. has already been extended by several later brain-to-text methods, e.g. Jayalath et al. (2025a), Wang et al. (2026), and Li et al. (2026), broadening the issue beyond the original work. This method also been evaluated in several subsequent studies (Jayalath & Parker Jones, 2026; Özdogan et al., 2025; Mantegna et al., 2026a; Landau et al., 2026), included in benchmarks (Banville et al., 2026; Jayalath et al., 2026), and discussed in recent reviews (Yang et al., 2026). Accordingly, the timing shortcut is relevant to a growing line of non-invasive brain-to-text work that builds on or evaluates this approach.

Repeated observations and word alignment. Our evaluation assumes that the temporal location of each word is known (word-aligned B2T), following d'Ascoli et al. (2025). This differs from solving speech segmentation jointly with word decoding, but is not uncommon among existing BCIs: for example, Moses et al. (2021) used a separate neural speech-detection model to identify the onset and offset of attempted words before classification. A non-invasive system could similarly use a paced interface in which individual word attempts are made separately, making segmentation substantially easier with non-invasive speech detection models (Dash et al., 2020; Jayalath et al., 2025b). Such a setting would also make repeated observations natural. Repetition is already widely used in simple non-invasive BCIs such as P300 spellers, where multiple responses are combined to improve signal-to-noise ratio (Farwell & Donchin, 1988). The scale of this effect here is notable relative to comparable speech decoding work. In the 2025 PNPL competition (Landau et al., 2025), averaging 100 observations per phoneme was used to improve decoding reliability, whereas SimpleB2T obtains large sentence-level gains from only five observations per word.

Towards useful non-invasive speech decoding. In 2021, Moses et al. provided a landmark demonstration that attempted speech could be decoded into sentences in a person with paralysis, achieving a median WER of 25.6% using a 50-word vocabulary and a set of 50 test sentences. While we do not decode speech from paralysed patients, and have not yet solved the problem of segmenting words without knowing their onsets, SimpleB2T shows that in decoding perceived speech from non-invasive recordings, it is possible to achieve WERs much closer to that landmark invasive result. This is possible while using repeated observations and operating over a larger 92-word vocabulary within a broader communication benchmark. Like their work, we focus on a constrained communication setting which allows for a stronger linguistic prior than unrestricted language decoding.

Once independent word decoding removes the timing shortcut introduced by the word-aligned input construction, the remaining challenges become clearer. SimpleB2T contains many better sentences among the top beam search candidates, suggesting opportunities for further improvement. More importantly, extending these results to decoding internally generated speech, such as imagined speech, will be a key step towards practical non-invasive communication interfaces.

AI USE STATEMENT

We used generative AI tools to assist with construction of the clinically motivated communication benchmark and to assist with drafting and editing the manuscript. All generated benchmark sentences were reviewed by the authors before use. Generative AI was also used to assist with code generation and debugging. All resulting code was inspected and its outputs were verified by the authors. All experimental designs, analyses, results, and scientific claims were reviewed and verified by the authors. We take responsibility for the final content of the paper, including text and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no new human-subject data collection. All neural recordings come from the publicly released LibriBrain100 dataset where ethical approval and participant consent for the original data collection are described by Mantegna et al. (2026b). The communication benchmark we constructed from this data is not clinically validated. Thus, we encourage user-centred evaluation before any real-world use. More broadly, progress in neural speech decoding raises important questions about privacy and consent. Practical systems should only decode deliberately provided signals with the informed consent of their users. However, we do not interpret the reported method as yet demonstrating a practical BCI since it operates on perceived and not imagined speech.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the experiments throughout the paper and appendices. Section 2 describes the protocol for our controls testing the overlap-window timing leakage; Section 3 defines SimpleB2T and its scoring procedure; Appendix H gives the preprocessing, architecture, training, and language-model decoding settings; Appendix C specifies the experimental controls; Appendix B lists the exact train, validation, and test splits; and Appendix I gives the 200-sentence communication benchmark and its construction procedure. We report the random seeds used for training and evaluate models across up to five seeds.

REFERENCES

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Elisa Cascardi, et al. NeuralBench: A unifying framework to benchmark NeuroAI models. *arXiv preprint arXiv:2605.08495*, 2026.

Debadatta Dash, Paul Ferrari, Satwik Dutta, and Jun Wang. NeuroVAD: Real-time voice activity detection from non-invasive neuromagnetic signals. *Sensors (Basel, Switzerland)*, 20, 2020.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5: 1097 – 1107, 2023.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert J. Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16:10521, 2025.

Lawrence A. Farwell and Emanuel Donchin. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography & Clinical Neurophysiology*, 70(6):510–23, 1988.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673, 2020.

Dulhan Jayalath and Oiwi Parker Jones. MEG-XL: Data-efficient brain-to-text via long-context pre-training. *International Conference on Machine Learning (ICML)*, 2026. arXiv preprint arXiv:2602.02494.

Dulhan Jayalath, Gilad Landau, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text. *International Conference on Machine Learning (ICML), Workshop on Generative AI and Biology*, 2025a. arXiv preprint arXiv:2505.13446.

Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark W. Woolrich, and Oiwi Parker Jones. The Brain's Bitter Lesson: Scaling speech decoding with self-supervised learning. *International Conference on Machine Learning (ICML)*, 2025b. arXiv preprint arXiv:2406.04328.

Dulhan Jayalath, Benjamin Ballyk, and Oiwi Parker Jones. A common measure of communication for speech brain-computer interfaces. *arXiv preprint arXiv:2609.02887*, 2026.

Gilad Landau, Miran Özdogan, Gereon Elvers, Francesco Mantegna, Pratik Somaiya, Dulhan Jayalath, Luisa Kurth, Teyun Kwon, Brendan Shillingford, Greg Farquhar, Minqi Jiang, Karim Jerbi, Hamza Abdelhedi, Yorguin Mantilla Ramos, Caglar Gulcehre, Mark Woolrich, Natalie Voets, and Oiwi Parker Jones. The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. *Advances in Neural Information Processing Systems (NeurIPS), Competition Track*, 2025. arXiv preprint arXiv:2506.10165.

Gilad D Landau, Dulhan Jayalath, and Oiwi Parker Jones. The semantic bottleneck: Leveraging semantic representations for non-invasive speech decoding. *arXiv preprint arXiv:2609.10296*, 2026.

Yueyang Li, Shuran Chen, Wai Ting Siok, and Nizhuan Wang. HDND: Hierarchical dynamic neural decoding for multilingual word/character retrieval from non-invasive brain recordings. *arXiv preprint arXiv:2609.24095*, 2026.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. *International Conference on Learning Representations (ICLR)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

Francesco Mantegna, Gereon Elvers, Dulhan Jayalath, Gilad Landau, Tasha Kim, Miran Özdogan, Luisa Kurth, Teyun Kwon, SungJun Cho, Benjamin Ballyk, Alex Fung, Anna Greer, Pratik Somaiya, Christian Herff, Jean-Rémi King, Yorguin Mantilla Ramos, Hamza Abdelhedi, Karim Jerbi, Greg Farquhar, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The 2026 PNPL competition: Word-classification and efficient cross-subject generalisation in LibriBrain100. *arXiv preprint arXiv:2609.03231*, 2026a.

Francesco Mantegna, Dulhan Jayalath, Gereon Elvers, Tasha Kim, Benjamin Ballyk, Alex Fung, SungJun Cho, Teyun Kwon, Luisa Kurth, Miran Özdogan, Gilad Landau, Pratik Somaiya, Natalie Voets, Mark Woolrich, and Oiwi Parker Jones. LibriBrain100: One hundred hours of broad and deep MEG data for neural speech decoding at scale. *arXiv preprint arXiv:2608.25204*, 2026b.

Stéphanie Martin, Peter Brunner, Chris Holdgraf, Hans-Jochen Heinze, Nathan E. Crone, Jochem W. Rieger, Gerwin Schalk, Robert T. Knight, and Brian N. Pasley. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, 7, 2014.

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. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

OpenAI. GPT-6 Astra: A New Generation of Intelligence. `https://openai.com/index/gpt-6-astra/`, 2026. Accessed 2026-09-19.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. RobustScaler available from `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html`.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

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

Jiaqi Wang, Huawen Hu, and Shu Zhang. Margin-regularized structured semantic alignment for brain-language correspondence. *arXiv preprint arXiv:2608.16975*, 2026.

Yiqian Yang, Yiqun Duan, Chenyu Liu, Yiqi Wang, Xinliang Zhou, Chin-Teng Lin, and Yu Zhang. Brain-to-language decoding: Tasks, signals, methods, evaluation, practical use and beyond. *arXiv preprint arXiv:2609.27650*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023.

Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. arXiv preprint arXiv:2506.02098.

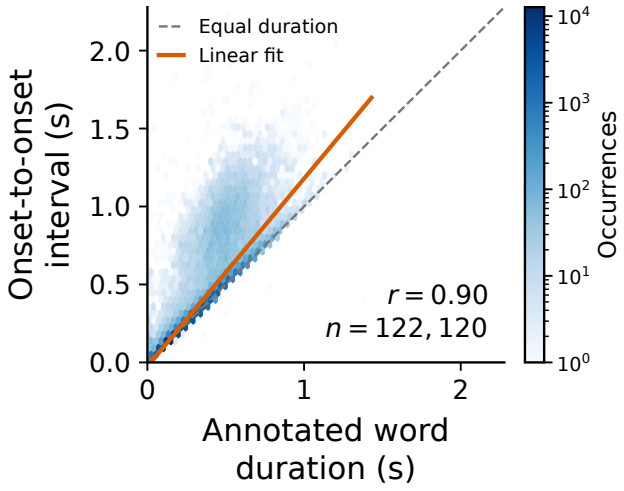


Figure 4: **Intervals between word onsets closely track word duration.** Relationship between the annotated duration of a word and the interval from its onset to the onset of the following word across adjacent word pairs. The two quantities are strongly correlated ($r = 0.90$). The dashed line indicates equality and the solid line shows a linear fit.

A Relationship Between Word Timing and Duration

The timing shortcut we identify relies on two properties of the word-aligned construction, namely that neighbouring windows must overlap, and the resulting onset-to-onset intervals must contain information about word duration. We find that both hold strongly in the perceived speech data we study. Three-second windows overlap for 99.9996% of adjacent word pairs. Across 122,120 adjacent word pairs, the onset-to-onset interval is strongly correlated with the annotated duration of the preceding word ($r = 0.90$; Figure 4).

B Data Splits

We partition subject-0 recordings into training, validation, and test sets as detailed in Table 3, assigning all recordings sharing a session label to the same partition across tasks. Test sessions were selected using word counts to supply at least five distinct occurrences for every word position in the core 100-sentence clinical benchmark while limiting the amount of data reserved from training. The additional expanded 100 sentences use previously unassigned occurrences from the same test sessions. Session 11 was reserved for validation, with all remaining sessions used for training.

C Experiment Details

LM-only baseline. To measure how much of the benchmark can be solved from linguistic priors alone, we run the same sentence-decoding procedure with the neural evidence removed. The baseline uses the frozen Qwen3-8B-Base model with the same clinical prompt as the corresponding Brain + LM condition. Beam search generates each candidate prefix autoregressively, with no access to MEG signals or reference words. The only difference from Brain + LM is therefore the absence of the neural score. Because decoding is deterministic once the prompt and sentence length are fixed, we report a single LM-only result. This baseline quantifies how much performance is explained by the language model under the vocabulary and length constraints of our benchmark.

Noise control. The language model may generate plausible clinical sentences even when the neural input contains no useful information. We therefore replace each MEG window with Gaussian noise matched to the first two moments of the preprocessed training data. We sample $M = 10{,}000$ training windows without replacement and estimate the mean and population variance separately for

Session ID	Split	Tasks	Recordings	Hours
1	Train	MOCHATIMIT, Sherlock1–9, TIMIT, TheMoth	12	6.96
2	Train	MOCHATIMIT, Sherlock1–9, TIMIT, TheMoth	12	6.85
3	Train	MOCHATIMIT, Sherlock1–9, TIMIT, TheMoth	12	6.45
4	Train	MOCHATIMIT, Sherlock1–9, TIMIT, TheMoth	12	6.47
5	Train	Sherlock1–9, TIMIT, TheMoth	11	5.51
6	Train	Sherlock1–9, TIMIT, TheMoth	11	6.61
8	Train	Sherlock1–9, TIMIT, TheMoth	11	6.39
9	Train	Sherlock1–9, TIMIT, TheMoth	11	7.02
10	Train	Sherlock1–9, TIMIT, TheMoth	11	5.77
12	Train	Sherlock1–7, Sherlock9, TIMIT, TheMoth	10	6.43
17	Train	TheMoth	1	0.25
19	Train	TheMoth	1	0.18
23	Train	TheMoth	1	0.24
24	Train	TheMoth	1	0.24
25	Train	TheMoth	1	0.29
26	Train	TheMoth	1	0.24
27	Train	TheMoth	1	0.13
28	Train	TheMoth	1	0.24
11	Validation	Sherlock1–7, Sherlock9, TIMIT, TheMoth	10	5.57
0	Test	Sherlock9	1	0.08
7	Test	Sherlock1–9, TIMIT, TheMoth	11	7.00
13	Test	Sherlock5–7, TIMIT, TheMoth	5	2.63
14	Test	Sherlock5–7, TIMIT, TheMoth	5	2.84
15	Test	Sherlock5, TheMoth	2	0.70
16	Test	TheMoth	1	0.22
18	Test	TheMoth	1	0.18
20	Test	TheMoth	1	0.23
21	Test	TheMoth	1	0.21
22	Test	TheMoth	1	0.21
29	Test	TheMoth	1	0.22
30	Test	TheMoth	1	0.19
Total: 31 session labels			162	86.57

Table 3: **LibriBrain100 subject-0 session split.** Session IDs and task names are taken from recording filenames. Task ranges are inclusive: for example, Sherlock1–9 denotes the nine tasks Sherlock1 through Sherlock9. Each listed task contributes one recording for that session ID. All recordings sharing a session label are assigned to the same split, including across tasks. The split contains 121 training recordings (66.29 hours), 10 validation recordings (5.57 hours), and 31 test recordings (14.71 hours). Durations refer to complete recordings before windowing. Core and Expanded use the test partition; development sentences use the validation partition.

each sensor c and onset-relative time point t:

$$\mu_{c,t} = \frac{1}{M}\sum_{j=1}^{M} x_{j,c,t}, \qquad \sigma^2_{c,t} = \frac{1}{M}\sum_{j=1}^{M} \left(x_{j,c,t} - \mu_{c,t}\right)^2. \tag{10}$$

Each test occurrence is then replaced by

$$\tilde{x}_{c,t} = \mu_{c,t} + \sigma_{c,t}\epsilon_{c,t}, \qquad \epsilon_{c,t} \sim \mathcal{N}(0,1), \tag{11}$$

with independent draws across sensors, time points, and occurrences. The resulting inputs therefore match the sensor- and time-specific mean and variance of real MEG in expectation, but contain no information about the perceived word.

Shuffled control. The Gaussian-noise control also removes the spatial and temporal structure of real MEG. We therefore use a second control that preserves the neural representations produced from real brain recordings while removing their correspondence to the correct words. We apply a single random permutation, chosen independently of the labels and without replacement, to the predicted embeddings before sentence decoding. This preserves the distribution and structure of the neural representations, but assigns them to the wrong test occurrences. A substantial advantage for correctly aligned embeddings over this shuffled control therefore shows that decoding depends on example-specific evidence decoded from the brain.

Joint synthetic signal. To test whether a contextual decoder can use window overlap without any neural information, we replace the MEG with a synthetic continuous signal. A simple Gaussian-noise signal would preserve identical samples across overlapping windows, but provides little structured variation for the CNN to encode. We therefore generate a sparse, smoothly varying signal with recognisable local features whose positions can be recovered after a temporal shift.

Specifically, we generate a 306-channel signal sampled at 50 Hz from three independent pulse processes, smoothed with Gaussian kernels of standard deviation 60, 120, and 240 ms. Pulses occur at a combined rate of 3 Hz. Pulse locations are shared across channels, while their amplitudes are sampled independently for each channel. The three smoothed components are summed and normalised to unit population variance. Importantly, pulse generation is entirely independent of the words and their onset times.

For each sentence, we generate one continuous synthetic signal and extract 3-second windows at the *true* word onsets, using the same windowing procedure as for the real MEG. Neighbouring windows therefore contain the same synthetic features at relative offsets determined by the true inter-word timing. Each window is baseline-corrected using its first 0.5 s and clipped to $[-5, 5]$.

Isolated synthetic signal. The independent control uses the same synthetic signal generator, but samples a separate signal for every word window. Consequently, neighbouring windows share no underlying samples and the true relative word timings are no longer encoded in the inputs.

During training, synthetic signals are regenerated on every forward pass. Validation and test signals are generated deterministically using sentence-specific random seeds. Comparing the shared and independent conditions therefore isolates the information provided by overlap between word-aligned windows.

Timing. To test whether word intervals alone support decoding, we train a timing-only control that replaces each MEG window with the interval to the next word onset within the same sentence. Intervals are log-transformed and standardised using training-set statistics. A two-layer MLP maps each scalar to a 1024-dimensional, ℓ_2-normalised representation, which is passed to the same sentence-level transformer and trained with the same contrastive objective.

D Timing Shortcut Reproduction Across Multiple Subjects

To test whether the overlapping-window shortcut generalises beyond the subject used for training, we evaluate the same checkpoint on all 32 remaining LibriBrain100 subjects. As shown in Figure 5, the shared synthetic decoder remains highly accurate across subjects despite receiving no neural

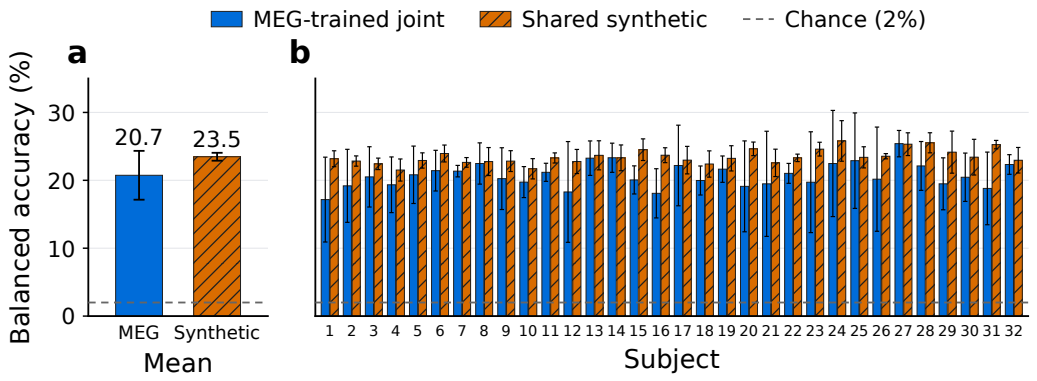


Figure 5: **The overlapping window shortcut reproduces across all subjects.** Subject-0-trained MEG and shared-synthetic decoders evaluated on the 32 other subjects in LibriBrain100. **(a)** Mean across subjects. **(b)** Per-subject results. Error bars show standard deviation across five training seeds.

information, and achieves performance comparable to the MEG-trained decoder. This shows that the shortcut is not specific to subject 0 and the overlap structure alone provides a transferable signal that can be exploited across unseen subjects and unseen speech material.

E Agreement with Duration-Based Word Predictions

To examine whether jointly decoding words captures timing information, we measure how closely its predicted word probabilities align with those inferred from word durations alone. We approximate duration by the interval between consecutive word onsets within a sentence. For each word in the 50-word vocabulary, we fit a Gaussian distribution over log intervals using training occurrences. Combining these likelihoods with training word-frequency priors gives a duration-based posterior $p_{\mathrm{dur}}(w \mid d)$ where w is a word type and d is the duration in samples.

We compare this posterior with predictions from the joint MEG decoder, the joint synthetic decoder, and our isolated MEG decoder. Each decoder's temperature is calibrated by minimising word-level negative log likelihood on validation data. For each training seed, we compute the change in Jensen–Shannon divergence

$$\Delta_{\mathrm{JS}} = \frac{1}{M} \sum_{i=1}^{M} \left[\mathrm{JS}\big(p_i, q_{\pi(i)}\big) - \mathrm{JS}(p_i, q_i) \right], \tag{12}$$

where p_i is the neural word distribution, q_i is the duration-based posterior, and π is a random permutation of test occurrences. Shuffling preserves the collection of duration-based predictions while breaking their correspondence with individual inputs.

The increase in divergence is substantially larger for the joint MEG and synthetic decoders than for the isolated decoder (Figure 6). Thus, both joint decoders exhibit stronger alignment with evidence derived from the duration of words. The similar effect for synthetic inputs supports the interpretation that the overlap structure contributes to performance from jointly decoding words.

F Jointly Decoding Words Without Overlapping Inputs

The synthetic signal experiment in Section 2.3 shows that overlapping word windows are sufficient to reproduce most of the gain from jointly decoding words without neural information. We perform an additional control to test whether this result can instead be explained by a train–test mismatch when overlap is removed.

We preserve the original word sequence of each natural sentence, but replace every position with a separately recorded occurrence of the same word. Any two windows presented in the same sentence must either come from different recordings or begin at least three seconds apart, ensuring

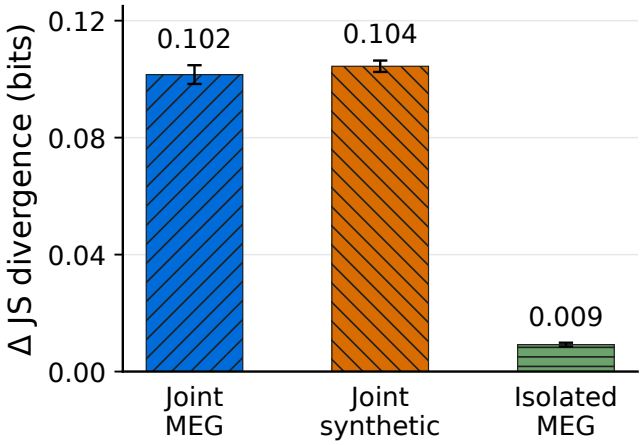


Figure 6: **Joint decoders aligns with duration-based predictions.** We measure how much the Jensen–Shannon divergence between the decoder probabilities and a duration-based word predictor's probabilities increase when their pairing across test occurrences is randomly shuffled. Thus, larger values indicate stronger alignment with the duration-based word predictor. Bars show means across five training seeds and error bars indicate sample standard deviations.

that they share no MEG samples. This preserves the linguistic structure of the natural sentences while removing the overlap between neighbouring neural inputs. We construct new versions of our training, validation, and test sets in this way independently where any neural response to a word may only be used once. Sentences for which we could not avoid overlaps were discarded. We were able to retain 91% of all the original sentences in the test recordings and 96% of the original sentences in the train recordings.

We compare three models: d'Ascoli et al. (2025)'s decoder trained on overlapping natural sentences; the same architecture retrained on non-overlapping sentences; and the isolated neural decoder used by SimpleB2T. All models are evaluated with one observation per word using balanced top-1 accuracy over the same 50-word vocabulary as in Section 2.3.

Figure 7 shows that removing overlap at test time causes the original model to collapse, but this alone could reflect a train–test mismatch. Retraining the same architecture without overlap recovers some performance, yet still does not outperform isolated decoding. Constructing non-overlapping sentences from separate recordings also removes coherent neural information shared across neighbouring words, leaving the joint decoding model primarily able to exploit regularities in the word sequence (effectively acting as an implicit language model).

G Predictability from Subsequent Words

This analysis tests whether SimpleB2T benefits from information about later words contained within each individual three-second window.

For each core test occurrence, we identify all later words whose annotated durations fall entirely within the same three-second window. We then use frozen Qwen3-8B-Base to measure how strongly this continuation predicts the target word. Specifically, for each of the 92 candidate target words, we compute the likelihood of the observed continuation conditioned on that candidate, and normalise these likelihoods assuming a uniform prior over candidates. This gives a probability distribution over possible target words using only the subsequent transcript. For each sentence position, we average these distributions across its five assigned occurrences.

We next test whether SimpleB2T performs better when this subsequent context is more predictive. Predictability varies substantially across positions, from 0.04% to 84.6%, with a median of 20.7% and an interquartile range of 8.1–36.9%. Thus, the benchmark contains both continuations that pro-

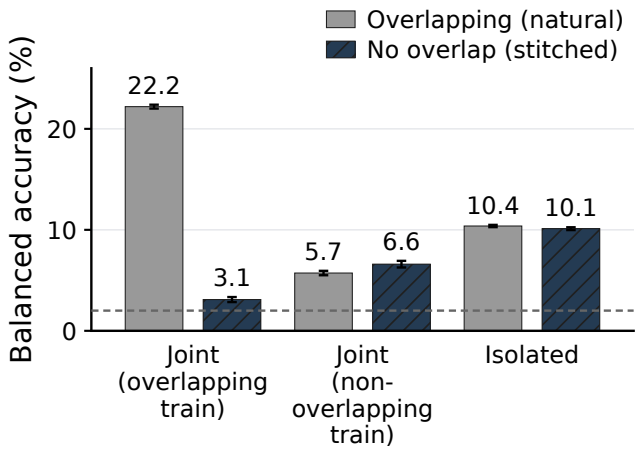


Figure 7: **Jointly decoding words with and without overlapping inputs.** We evaluate d'Ascoli et al. (2025), the same architecture trained without overlapping neighbouring windows, and isolated decoding on both natural overlapping sentences and non-overlapping versions of the same word sequences. Error bars indicate ± 1 sample standard deviation across training seeds.

vide little information about the target word and continuations that strongly constrain its identity. We split examples into low- and high-predictability groups. Because some words may be intrinsically easier to decode than others, we also construct a word-matched split. For each target word separately, we rank its positions by subsequent-context predictability and assign equal numbers to the low- and high-predictability groups.

Table 4 reports positional word accuracy using the unchanged $k = 5$, beam-50 predictions. In the word-matched analysis, Brain accuracy rises modestly from 24.0% on low-predictability positions to 27.1% on high-predictability positions. In contrast, Brain + LM accuracy is almost unchanged, at 65.3% and 64.9%, respectively. The overall split shows no consistent advantage for high-predictability positions. Therefore, SimpleB2T does not appear to significantly benefit from other words in each three-second window.

	Overall split		Word-matched split	
Method	Low	High	Low	High
LM	26.3	25.8	28.3	28.3
Brain	$27.5_{\pm 2.0}$	$23.8_{\pm 2.7}$	$24.0_{\pm 2.1}$	$27.1_{\pm 2.6}$
Brain + LM	$62.6_{\pm 1.3}$	$64.0_{\pm 2.2}$	$65.3_{\pm 2.0}$	$64.9_{\pm 1.9}$

Table 4: **Word accuracy grouped by the predictability of subsequent context.** Subscripts denote sample standard deviations across five training seeds. The LM is deterministic.

H Training and Decoding Details

Table 5 summarises the preprocessing, architecture, training, and sentence decoding settings.

Language-model scoring. We use the fixed prompt `A patient communicates a short request to hospital staff.`, followed by a newline and `Patient:`. We do not apply a chat template or insert any additional special tokens. Candidate words are formatted with a leading space, the first word of the sentence is capitalised, and the pronoun "I" is always uppercase.

Hyperparameter	Setting
Input	
MEG channels / sampling rate	306 / 50 Hz
Word window / baseline correction	3 s from onset / subtract channel-wise mean of first 0.5 s
Bandpass filter / clipping	0.1–40 Hz / $[-5, 5]$ after scaling
Channel scaling	Recording-wise, channel-wise RobustScaler (Pedregosa et al., 2011)
Model	
CNN	Temporal CNN with spatial channel merging and temporal attention pooling (Défossez et al., 2023; d'Ascoli et al., 2025)
CNN depth / hidden channels	5 / 160
Kernel size / dilation period	3 / 5
Input dropout / batch normalization	0.1 / enabled
Initial linear width	512
Spatial merger positional dimension	2048
Residual MLP blocks	4; LayerNorm–Linear–GELU–Linear
MLP dimensions	$1024 \to 2048 \to 1024$
Embedding normalisation	Unit ℓ_2 norm before and after MLP
Word targets	Mean embeddings from layer 12 of T5-large (Raffel et al., 2020)
Training and model selection	
Training observations	Individual word occurrences ($k = 1$)
Objective	D-SigLIP (Zhai et al., 2023; d'Ascoli et al., 2025)
Optimizer / learning rate	AdamW (Loshchilov & Hutter, 2019) / 10^{-4}
Weight decay / batch size	0 / 128
Schedule / maximum epochs	Cosine annealing (Loshchilov & Hutter, 2017) / 50
Early-stopping patience	10 epochs without improvement
Checkpoint selection	Validation balanced word accuracy ($k = 5$)
Training seeds	0, 100, 200, 300, 400
Sentence decoding	
Temperature T	0.049
Individual-evidence α	2
Language model	Qwen3-8B-Base (Qwen Team, 2025)
LM weight λ	Default: 0.5; observation-dependent settings below
Beam width	50

Table 5: **SimpleB2T hyperparameters.** The ablations use $\lambda = 0.5$, while other results use $\lambda = (1.5, 1, 0.5, 0.5, 0.5)$ for $k = 1, \ldots, 5$.

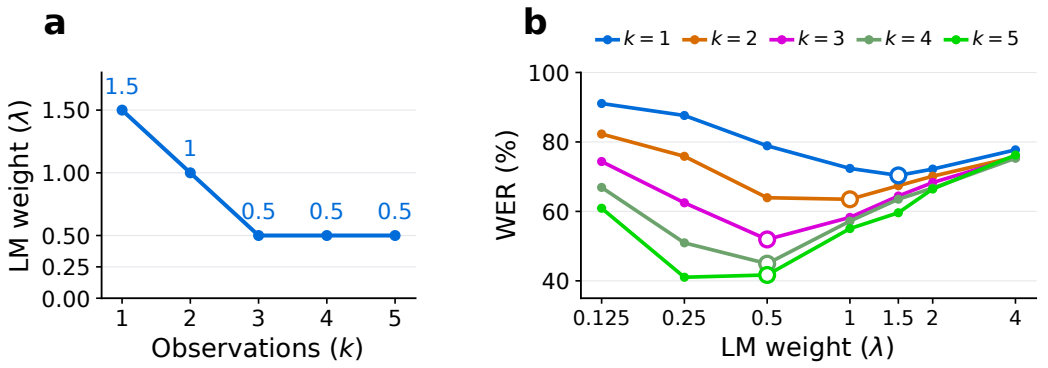


Figure 8: **Language-model weight decreases as neural evidence improves.** **(a)** Development-selected λ falls as more observations are aggregated. **(b)** WER is minimised at lower LM weights for larger k.

If a candidate word w is tokenised as $(u_1, \ldots, u_m)$, we score it by summing the autoregressive log probabilities of its constituent tokens:

$$\ell_{\mathrm{LM}}(w \mid h) = \sum_{j=1}^{m} \log p_{\mathrm{LM}}(u_j \mid h, u_{<j}), \tag{13}$$

where h contains the fixed prompt together with the current candidate sentence prefix. Each token probability is computed using the model's full tokenizer vocabulary.

Beam search. At each sentence position, beam search expands every selected prefix with all 92 candidate words. For each expansion, we add the neural decoder score and the language-model score weighted by λ. The number of word positions is known in advance, and the decoder therefore generates exactly one word at each position.

After the final word, we also score the end of the sentence. For each remaining candidate sentence, we add

$$\lambda \log[p_{\mathrm{LM}}(\,.\mid h) + p_{\mathrm{LM}}(?\mid h)]\,, \tag{14}$$

where both punctuation marks are single tokens and h now contains the complete candidate sentence. The highest-scoring completed candidate is returned.

Language-model weight across observations. We select the language-model weight λ separately for each number of observations k using the 50-sentence development set, and fix the selected value before test evaluation. The selected weights are

k	1	2	3	4	5
λ	1.5	1.0	0.5	0.5	0.5

As more observations are aggregated, the selected language-model weight decreases and then stabilises (Figure 8).

I Clinical Communication Benchmark

Tables 6 (core) and 7 (expanded) list the 200 clinically motivated sentences used in our communication benchmark. The sentences were generated with assistance from GPT-6 Astra (OpenAI, 2026) and subsequently reviewed to ensure they represented plausible patient communication needs. The MEG samples for the benchmark were constructed by assigning recorded word occurrences to positions in predefined sentences. For each position, we select five distinct occurrences of the required word from the held-out test sessions, irrespective of their original sentence context, and extract the corresponding word-aligned MEG windows. Each occurrence is assigned to exactly one position across the 200-sentence benchmark, so no recording event is reused.

Table 6: **Core communication benchmark sentences.**

ID	Sentence	ID	Sentence
1	Can you help me?	51	Can you open the door?
2	I would like some help.	52	Can you open the window?
3	Can you come here?	53	Can you put the light on?
4	Can you come back?	54	Can you take this away?
5	Can you be here with me?	55	Can you put that over me?
6	Can you give me more time?	56	Can you put this by my hand?
7	I would like to be on my own.	57	Can you put this on the other side?
8	Do not go away.	58	Can you put that down?
9	Can you come back in a little while?	59	Can you put that here?
10	Can you ask for help?	60	Can you put that back?
11	I would like some water.	61	Can you tell me your name?
12	Can you give me some water?	62	Can you tell me what that is?
13	I would like more water.	63	Can you tell me what this is for?
14	That is enough water.	64	Can you say that again?
15	Can you put the water here?	65	Can you tell me more?
16	Can you help me with the water?	66	Can you give me some paper?
17	Can you take the water away?	67	Can you look at this paper?
18	I would like a little more.	68	I don't see that.
19	That is too much.	69	I don't see your face.
20	No more for now.	70	Can you ask me one thing at a time?
21	Can you help me up?	71	Can you give me a little time?
22	Can you help me down?	72	I have something to say.
23	Can you put my head up?	73	That is not what I said.
24	Can you put my head down?	74	Yes, that is right.
25	Can you put my hand here?	75	No, that is not right.
26	Can you put my hand there?	76	I do not know.
27	Can you take my hand?	77	I would like to know why.
28	Can you put this under my head?	78	Can you tell me when?
29	Can you put this under my back?	79	Can you tell me how?
30	Can you help me get on my side?	80	I would like to ask something.
31	I would like to be on my left side.	81	I would like to see my friend.
32	I would like to be on my right side.	82	Can you tell my friend to come here?
33	I would like to be on my back.	83	Can you ask my friend to come back?
34	Can you help me with my head?	84	Can you tell him I am here?
35	Can you help me with my back?	85	Can you tell her I am here?
36	Can you help me with my left hand?	86	I would like to see him.
37	Can you help me with my right hand?	87	I would like to see her.
38	Do not put that on my head.	88	Can you ask them to come in?
39	Do not put that on my back.	89	Can you ask them to come back?
40	That is not the right place.	90	I would like more time with them.
41	Can you wash my face?	91	Can you help me now?
42	Can you wash my hands?	92	That is too much light.
43	Can you wash my head?	93	There is something in my eyes.
44	Can you wash my back?	94	There is something on my face.
45	Can you wash my left hand?	95	Can you look at my hand?
46	Can you wash my right hand?	96	Can you look at my head?
47	Can you wash here?	97	Can you look at my back?
48	Can you wash there?	98	Can you look at my eyes?
49	Can you help me wash?	99	Can you tell me the time?
50	Do not get water in my eyes.	100	Can you come back before night?

Table 7: **Expanded communication benchmark sentences.**

ID	Sentence	ID	Sentence
1	I would like some water now.	51	Tell her to come back in a little while.
2	Give me a little water now.	52	Tell him to come back before night.
3	I would like water before you go.	53	I would like to see them now.
4	Take the water away now.	54	I would like to have more time with her.
5	I have enough water for now.	55	I would like to have more time with him.
6	I would like the water by my hand.	56	Ask them to be here with me.
7	Do not give me more water.	57	I do not know your name.
8	Give me some more time with the water.	58	Tell me your name again.
9	There is water on my face.	59	Tell me why you have come back.
10	There is water in my eyes.	60	I would like to know what this is.
11	I would like to wash my face.	61	I would like to know what that is for.
12	I would like to wash my hands.	62	I do not know why you said that.
13	Wash my face with a little water.	63	Say that one more time.
14	Do not wash my eyes.	64	Say one thing at a time.
15	Wash my hands again.	65	Give me time to say what I would like.
16	I would like to wash before night.	66	I have more to say.
17	Wash the other hand too.	67	That is what I would like.
18	There is something under my back.	68	That is not what I would like.
19	There is something under my head.	69	I said no.
20	I would like my head a little more up.	70	I said yes.
21	I would like my head a little more down.	71	I would like to ask you something.
22	My hand is not in the right place.	72	Ask me again in a little while.
23	Take this away now.	73	Do not ask me now.
24	I would like to be on the other side.	74	I do not know how to say this.
25	Do not take my hand away.	75	I would like some paper now.
26	Put this under my hand.	76	Give me the paper again.
27	Put that by my side.	77	Look at what is on the paper.
28	Put the water on the other side.	78	I don't see the paper.
29	Put this paper by my hand.	79	I would like the paper here.
30	I would like my back down a little.	80	There is something I would like you to see.
31	The light is in my eyes.	81	Look at my face when I say this.
32	I would like a little more light.	82	I don't see what is in your hand.
33	I would like the light on now.	83	I would like to see your face.
34	Do not have the light on at night.	84	I don't know what to do.
35	The window is open.	85	Tell me what you would like me to do.
36	I would like the window open for a little while.	86	Tell me before you wash my face.
37	Do not open the window now.	87	Tell me before you take this away.
38	I would like the door open.	88	I would like to know when you would be back.
39	Do not open the door now.	89	How much time do I have?
40	Do not have this by my eyes.	90	What is this water for?
41	Come here for a little while.	91	Is that my name on the paper?
42	Be here while I have some water.	92	Is my friend here now?
43	I would like you here with me.	93	Do you know when my friend would be here?
44	Do not go while I have something to say.	94	Do you know when I would see him again?
45	Come back at night.	95	I would like you to look at my eyes.

Continued on next page

Table 7 continued from previous page

ID	Sentence	ID	Sentence
46	Come back in a little while.	96	I would like you to look at my back.
47	I would like a little time on my own.	97	Do not give me that now.
48	I would like my friend here with me.	98	That is enough for now.
49	Ask my friend to come in.	99	I would like more time before you go.
50	Ask my friend to come back at night.	100	I would like to know why you said no.

J TRAIN–TEST MISMATCH CONTROL IN THE CLINICAL BENCHMARK

One possible explanation for the poor transfer of jointly decoding words to our clinical benchmark is a mismatch between training and evaluation. The original model is trained on sentences with overlapping neighbouring windows, whereas benchmark positions are assembled from separately recorded occurrences and therefore do not overlap. To test this, we retrain the same model on non-overlapping sentence constructions matched to the evaluation setting.

We construct non-overlapping versions of the natural training sentences while preserving their original word sequences. For each word position, we replace the original MEG window with a three-second window from a different recorded occurrence of the same word. We require that no two windows assigned to the same constructed sentence overlap in the underlying recording. They must either come from different recordings or begin at least three seconds apart. The model therefore sees exactly the same linguistic sequence as before, but cannot recover word intervals from shared MEG samples.

To construct these sentences without reusing data, we permute recorded occurrences among positions of the same word type, so that each retained occurrence is assigned to exactly one position per epoch. We then resolve any remaining overlaps within sentences by swapping assignments between positions with the same target word. Sentences for which no valid assignment can be found are discarded. This retains *96.3% of eligible training occurrences*. Assignments are re-randomised between epochs while preserving the same non-overlap constraint.

Retraining does not materially improve sentence reconstruction (Table 8). With LLM rescoring at $k = 5$, the matched model obtains 79.2%, 85.5%, and 82.6% WER on Core, Expanded, and Full, respectively, compared with 79.4%, 85.5%, and 82.7% for the original overlap-trained model. Our clinical benchmark is constructed from separately recorded occurrences at each sentence position, so there is no coherent cross-word neural activity for the d'Ascoli et al. decoder to exploit. The control shows that matching the decoder to the non-overlapping, word-by-word setting does not recover the performance of independent decoding.

		Core		Expanded		Full	
Method	k	WER (%) ↓	SMR (%) ↑	WER (%) ↓	SMR (%) ↑	WER (%) ↓	SMR (%) ↑
d'Ascoli	1	$97.7_{\pm 0.4}$	$0.0_{\pm 0.0}$	$97.0_{\pm 0.7}$	$0.0_{\pm 0.0}$	$97.3_{\pm 0.6}$	$0.0_{\pm 0.0}$
d'Ascoli + LM	1	$79.9_{\pm 1.5}$	$2.3_{\pm 0.2}$	$86.2_{\pm 0.9}$	$0.6_{\pm 0.0}$	$83.2_{\pm 0.2}$	$1.4_{\pm 0.1}$
d'Ascoli	5	$98.0_{\pm 0.6}$	$0.0_{\pm 0.0}$	$97.2_{\pm 1.1}$	$0.0_{\pm 0.0}$	$97.5_{\pm 0.8}$	$0.0_{\pm 0.0}$
d'Ascoli + LM	5	$79.2_{\pm 1.6}$	$2.3_{\pm 0.6}$	$85.5_{\pm 0.7}$	$1.0_{\pm 0.0}$	$82.6_{\pm 0.5}$	$1.7_{\pm 0.3}$

Table 8: **Controlling for train–test mismatch in the joint word decoder.** We retrain the decoder of d'Ascoli et al. (2025) on non-overlapping sentence constructions and evaluate at $k = 5$. Jointly decoding words remains much less accurate than SimpleB2T. Values are means across training seeds and subscripts show sample standard deviations.

K WORD-LEVEL ANALYSIS

We further examine which words are easiest to decode. Figure 9 relates mean Brain + LM word accuracy across five training seeds to properties of the 92 vocabulary words. Phonetic distinguishability is the minimum phoneme edit distance to another vocabulary word, normalised by the longer

pronunciation's length, using CMUdict pronunciations with stress markers removed and taking the minimum if there are pronunciation variants. Semantic distinguishability is one minus the maximum cosine similarity to another word's frozen T5 target embedding. Neither measure is strongly associated with accuracy ($r = -0.02$ and $r = -0.11$, respectively). Training frequency ($r = 0.09$) and mean word duration ($r = -0.10$) also show weak associations. Duration variability, computed as the sample standard deviation of annotated word durations across the used training occurrences, has a stronger negative association ($r = -0.26$). Therefore, words with more consistent durations tend to be decoded more accurately. The strongest association among these measures is with LM predictability ($r = 0.42$), defined as the target word's probability under the frozen LM, conditioned on the fixed prompt and preceding reference words. All correlations are Pearson correlations. The accuracy-binned word groups illustrate the range of performance, with both function and content words appearing across accuracy bins.

L Effect of the Linguistic Prior

We examine how the language-model prior changes with the prompt (Table 9). These experiments were conducted *after* all main SimpleB2T experiments were complete and were not used to select the prompt reported in the main results. The prompt used throughout the paper was fixed before this ablation. We compare our prompt with several alternatives ranging from a minimal prefix to broader descriptions of English, everyday speech, and clinical communication. The results show that specifying the intended patient-communication setting improves decoding relative to more generic prompts. This supports the view that part of SimpleB2T's success comes from using a linguistic prior matched to the constrained communication task. Future work may benefit from optimising the prompt on a development set to further improve the linguistic prior.

		Core		Expanded		Full	
Prompt	λ	LM only	Brain + LM	LM only	Brain + LM	LM only	Brain + LM
A	0.25	97.3	$55.2_{\pm 1.2}$	98.1	$53.9_{\pm 0.8}$	97.7	$54.5_{\pm 1.0}$
B	0.5	96.0	$51.3_{\pm 3.6}$	91.0	$52.5_{\pm 2.6}$	93.3	$51.9_{\pm 3.1}$
C	0.25	91.8	$51.8_{\pm 3.3}$	90.0	$49.8_{\pm 3.1}$	90.8	$50.7_{\pm 3.2}$
D	0.25	84.4	$49.3_{\pm 2.0}$	95.0	$48.1_{\pm 3.2}$	90.0	$48.7_{\pm 2.6}$
E (ours)	0.5	**73.8**	$\mathbf{36.6_{\pm 1.7}}$	91.8	$44.8_{\pm 3.7}$	83.4	$41.0_{\pm 2.2}$
F	0.5	78.8	$38.7_{\pm 4.3}$	**87.2**	$\mathbf{42.6_{\pm 5.1}}$	**83.3**	$\mathbf{40.8_{\pm 4.6}}$

ID	Exact prompt
A	Sentence:
B	The following is a sentence from an English passage:
C	The following is a sentence from everyday speech:
D	The following is a sentence spoken in a clinical setting:
E	A patient communicates a short request to hospital staff. Patient:
F	A patient who cannot speak communicates a short message to hospital staff about comfort, positioning, personal care, their surroundings, or interaction with other people. Patient:

Table 9: **Prompt specificity improves sentence reconstruction.** Brain + LM uses $k = 5$, beam width 50, and frozen Qwen3-8B-Base. Subscripts report sample standard deviations. LM only decoding is deterministic. Bold, shaded cells mark the lowest WER within each condition and split. A line break before `Patient:` is part of prompts E and F.

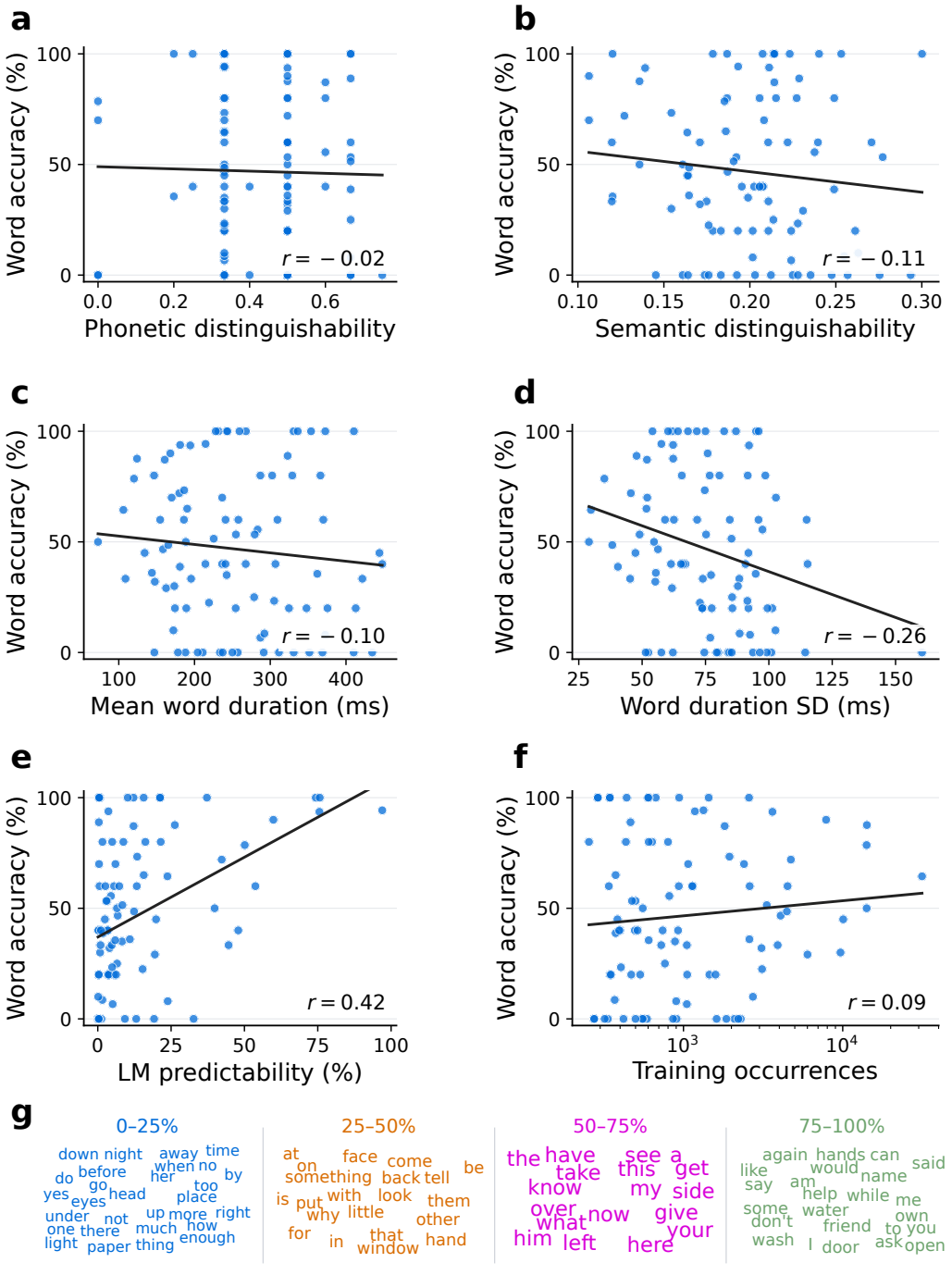


Figure 9: **Word-level decoding analysis.** Word accuracy is weakly related to **(a)** phonetic distinguishability, **(b)** semantic distinguishability, and **(c)** mean word duration. It is negatively related to **(d)** word duration variability and positively related to **(e)** LM predictability, with little association with **(f)** training frequency. **(g)** Words grouped by accuracy bins.