Is EEG-to-Text Feasible in Real-World Scenarios? An In-Depth Analysis Using a Neuropsychology-Inspired Benchmark

Zihan Zhang[1*], **Yu Bao**[1,2*], **Xiao Ding**[1†], **Tianyi Jiang**[3], **Kai Xiong**[4]
[1]Research Center for Social Computing and Interactive Robotics, Harbin Institute of Technology
[2]Shanghai Innovation Institute, Shanghai, China
[3]State Key Laboratory for Novel Software Technology, Nanjing University
[4]Zhongguancun Laboratory, Beijing, China
{zihanzhang, ybao, xding}@ir.hit.edu.cn, xiongk@zgclab.edu.cn

Abstract

Translating brain signals into text could restore communication for people with severe paralysis, yet practically usable systems to date rely on invasive electrocorticography (ECoG). Electroencephalography (EEG) offers a non-invasive alternative, and EEG-to-text (EEG2Text) has been widely explored. Interestingly, however, EEG2Text models generally rely on teacher-forcing evaluation; without it, they fail to generate meaningful decoding. This reliance prevents EEG2Text from being applied in real-world, non-academic settings. This has fueled numerous debates about whether EEG2Text is a meaningful direction, by extension, and whether EEG truly contains decodable linguistic information. Here, using a neuropsychology-informed paradigm, we find that existing EEG2Text benchmarks have neglected EEG instability, a flaw that has confounded inference and sparked debate. Our experiments furnish key evidence for the feasibility of teacher-forcing-free EEG2Text decoding. Accordingly, we assemble the **Corpus OF Eeg-To-Text (COFETT)** using a 128-channel high-density EEG cap, providing a benchmark dedicated to evaluating EEG2Text models. In comparisons with multiple existing benchmarks, COFETT achieves SOTA ability to distinguish among model performances and enables robust, teacher-forcing-free evaluation, thereby opening a path toward practical EEG2Text applications. COFETT is open sourced in https://github.com/baoyudu/COFETT.

1 Introduction

Brain–computer interfaces (BCIs) that translate neural activity into text have made notable progress in restoring communication for people with severe paralysis (Silva et al., 2024). In particular, invasive systems based on ECoG can directly decode Inner Speech, enabling near-real-time transformation of cortical signals into verbal output with high accuracy (Willett et al., 2023; Card et al., 2024). However, the requirement for craniotomy and implanted electrodes restricts use to patients who cannot undergo surgery and raises ethical and practical constraints.

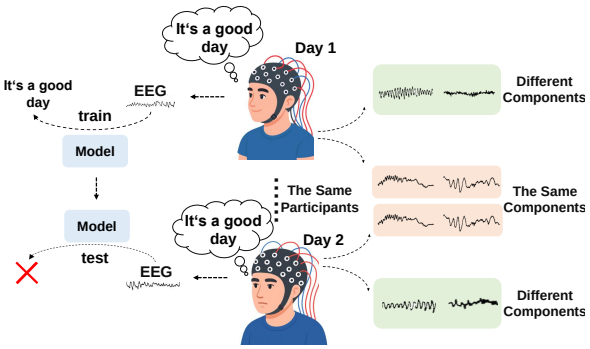

Figure 1: EEG instability in EEG2Text. The EEG elicited by imagining the same sentence on day 1 versus day 2 exhibits different components and features.

As a non-invasive alternative, several studies have attempted EEG2Text decoding, aiming to recover internal semantic intent without overt speech or motor commands (Wang and Ji, 2022; Zhou et al., 2024; Feng et al., 2023; Duan et al., 2023; Xi et al., 2023; Wang et al., 2024b). These studies have enabled EEG2Text to achieve performance comparable to ECoG2Text. So why are today's BCI applications predominantly focused on the skull-penetrating ECoG, rather than the less harmful EEG? Because these evaluations rely on teacher-forcing, a strategy appropriate for training but inappropriate for evaluation: it evaluates sequence models by feeding the ground-truth previous token to predict the next one, thereby masking exposure bias and inflating reported performance. However, it is evident that such ground-truth tokens are unavailable in real-world applications. Given this, what is the real-world performance of EEG2Text?

*Equal Contribution.
†Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1378–1393
July 2-7, 2026

Recent analyses suggest that, under teacher-forced evaluation, some EEG2Text systems perform comparably when fed random noise inputs, while failing to generate meaningful decoding without teacher-forcing, casting doubt on whether models truly learn linguistically meaningful structure from EEG (Murad and Rahimi, 2024; Jo et al., 2024). These concerns have led to broader skepticism about the feasibility of EEG2Text and underscore the need for stricter evaluation protocols.

Motivated by these challenges, we critically re-examined existing EEG2Text studies, observing that prior debates and investigations centred on model architectures while the reliability of the benchmarks themselves remained largely unexamined. We found that existing benchmarks (for example, ZuCo (Hollenstein et al., 2018, 2019)) do not consider EEG instability. EEG is inherently stochastic and markedly non-stationary even within the same participant (Downey et al., 2018). Here, 'instability' denotes trial-to-trial and session-to-session shifts in signal statistics and spatial topography, arising from cognitive-state drift, electrode displacement, impedance changes, and physiological fluctuations. As shown in Figure 1, a model trained on Day 1 data fails to perform effectively on data from the same subject collected on Day 2.

To address these issues, we introduce a neuropsychology-informed paradigm, which collects labelled data from multi-round inner-speech imagery with carefully spaced repetitions to maximise the recoverable linguistic component in EEG. Using 128-channel high-density recordings, we construct COFETT, a benchmark that enables teacher-forcing-free evaluation. In comparative experiments across multiple methods, COFETT shows stronger discriminative power, providing a more reliable benchmark for assessing EEG2Text model.

Our main contributions are summarised below:

- **COFETT dataset.** A novel, high-density, neuropsychology-grounded EEG dataset, specifically designed for EEG2Text decoding, which accounts for EEG instability.
- **Teacher-forcing-free evaluation.** An assessment framework that prohibits teacher-forcing during inference, thereby measuring genuine learning.
- **Feasibility evidence.** Experimental results show that EEG contains linguistically decodable information, providing new evidence for the feasibility of EEG2Text decoding.

2 Related Work

2.1 EEG Instability

EEG instability, defined as temporal drift within the same participant, is a key contributor to EEG noise. It is driven by changes in attention, vigilance and fatigue, by electrode placement and impedance, and by muscle and eye movements (Downey et al., 2018). As a result, recordings collected earlier and later in a dataset can follow extremely different distributions, which greatly hinders decoder training.

Because such instability is intrinsic to EEG, other BCI subfields have developed stability-aware methods and protocols, for example stationary Common Spatial Patterns that bias spatial filters toward invariant subspaces (Samek et al., 2012), and inter-session data-space or domain adaptation to reduce distribution shift across days (Arvaneh et al., 2013); both approaches improve robustness over time.

By contrast, as a nascent, deep-learning–driven field, EEG2Text relies on benchmarks such as ZuCo and ZuCo 2.0 (Hollenstein et al., 2018, 2019) that were built for natural reading and do not provide within-participant, repeated readings of the same sentences across sessions, so stability is rarely measured or exploited. This gap has slowed progress in the field.

2.2 Teacher-Forcing EEG2Text

Teacher-forcing feeds the ground-truth previous token to an autoregressive model when predicting the next token. During training, it stabilizes likelihood optimization and accelerates convergence for RNN/decoder-style architectures (Williams and Zipser, 1989). When applied at evaluation, a use that extends beyond its original purpose, it creates a train–test mismatch. Real generation must condition on the model's own history, but conditioning on the gold history hides error accumulation, commonly referred to as exposure bias, and yields optimistically high token-level metrics.

In EEG2Text, because of the instability of EEG, several representative systems have to rely on teacher-forced evaluation; without it, EEG2Text models cannot generate useful information that can be meaningfully assessed. This practice essentially redefines generation as next-token classification conditioned on the gold context. For example, DeWave explicitly states that its evaluation uses teacher-forcing to eliminate accumulation error, turning decoding into word-level classification

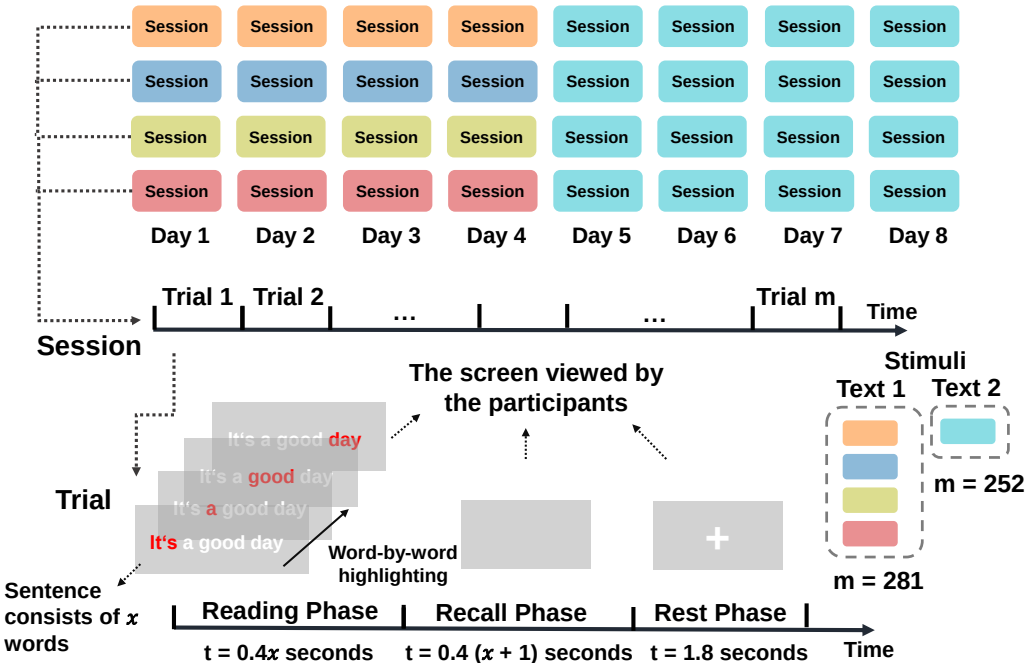

Figure 2: The full experimental procedure for each participant (8 days $\times$ 4 sessions/day $\times$ m trials/session $\times$ 3 phases/trial). Identical colors indicate that the participant read and imagined (recalled) the same text during that period. Specifically, days 1 to 4 involved a set of texts (Text 1), repeated 4 times, while days 5 to 8 involved a different set of texts (Text 2), repeated 16 times.

given the previous gold token (Duan et al., 2023). Subsequent work also acknowledges that reported gains still depend on teacher-forcing (Wang et al., 2024b). Independent analyses further show that, under teacher-forced evaluation, some pipelines perform similarly when the EEG input is replaced by random noise. This finding suggests that many decoded tokens primarily reflect the pretrained language model's prior rather than linguistic content in EEG (Jo et al., 2024). Together, these observations underscore the need for stricter, teacher-forcing-free protocols.

3 Methods

Existing benchmarks could not evaluate model performance without teacher forcing. To enable genuine, teacher-forcing-free assessment, our benchmark makes two core advances. First, the dataset: we collect EEG from participants who repeat the same sentence at carefully spaced intervals to counter EEG instability (Methods 3.1). Second, the evaluation framework: we replace BLEU with a vector-alignment metric that evaluates teacher-forcing-free outputs without gold conditioning (Methods 3.2). Additionally, to furnish evidence for the feasibility of EEG2Text, we analyse large-scale statistical properties of EEG and conduct scaling-law analyses inspired by the LLM literature (Methods 3.3).

3.1 Construction of COFETT

In EEG2Text, instability has a narrower definition because we focus only on the language-related components of EEG. Figure 1 illustrates this: instability refers to the variant signal components when the same individual imagines the same sentence at different time points. Although EEG2Text is formally framed as a machine translation task, it differs fundamentally from conventional machine translation. Traditional translation involves relatively fixed and invariant mappings between sentences in two languages. By contrast, EEG2Text requires mapping highly variable neural signals to sentences in natural language. Unlike static sentence representations, EEG signals are dynamic, time-varying features that evolve with cognitive state and recording conditions. Therefore, the core challenge for EEG2Text lies in collecting EEG datasets that minimise the noise introduced by such instability.

Therefore, guided by insights from neuropsychology (Proix et al., 2022), we designed an experimental paradigm as shown in Figure 2. Participants completed an eight-day experimental protocol, with each day consisting of four sessions. Each session included m trials, where $m = 281$ on days 1–4

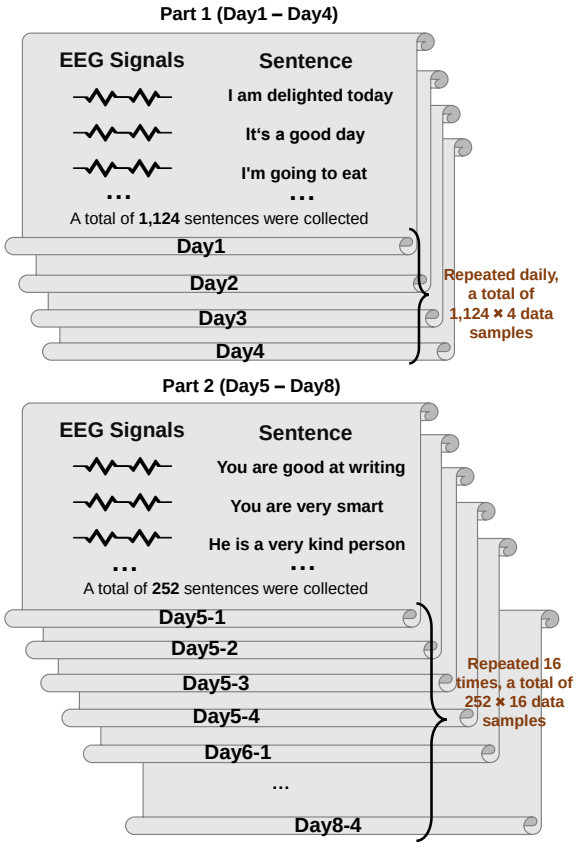

Figure 3: The EEG dataset obtained from the experiment is divided into 2 main parts. In the first part, each sentence was repeated 4 times, capturing EEG signals from the same participant recalling the identical sentence on 4 occasions. In the second part, sentences were repeated 16 times.

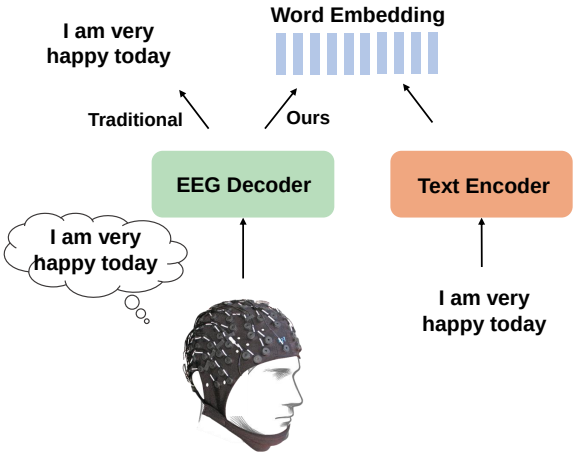

Figure 4: Comparison of supervision methods for EEG2Text models.

and $m = 252$ on days 5–8. Each trial, in turn, was composed of word-by-word reading, word-by-word recall, and a rest period. The textual materials span 39 categories, covering nearly all routine expressions.

The word-by-word recall phase is the focus of our subsequent analysis, as it represents the Inner Speech. At this stage, participants were instructed to recall, verbatim, the content they had just read during the reading phase, with the intention of verbalizing it but refraining from doing so. The Inner Speech produced in this process aligns with the principles of EEG2Text (Zhang et al., 2024).

The experimental design for single trials is inspired by previous psycholinguistic studies on imagined speech, where reading guides the imagination process. The imagined speech is considered to be embedded within the participant's Inner Speech (Zhang et al., 2024; Nieto et al., 2022). Moreover, the reading phase activates brain regions associated with language processing, such as the left temporal lobe and inferior frontal gyrus, which are also involved in Inner Speech, showing partial overlap in neural representations (Tian and Poeppel, 2010; Hickok and Poeppel, 2007). Thus, the experimental paradigm in this study offers two main advantages: first, it constrains the content of participants' imagination, mitigating dataset bias that could arise from free imagination; second, the activation of similar brain regions during both phases aids in a smoother transition to the Inner Speech state. These factors collectively contribute to more effective training of EEG2Text models.

Ultimately, we obtained the EEG data shown in Figure 3, which can also be represented as the following:

$$\mathcal{D} = \{(EEG_i, S_i) \mid i = 1, 2, \ldots, M,$$
$$EEG_i = EEG_{i1}, EEG_{i2}, \ldots, EEG_{ix}\},$$

where EEG_i denotes a set of EEG signals associated with sentence S_i, and $EEG_{i,j}$ corresponds to the EEG recording obtained from the j-th repetition of sentence S_i by a given participant. S_i is a sentence chosen from a predefined sentence library $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$, where each S_i is a natural language sentence.

3.2 Teacher-Forcing-Free Evaluation

Existing work shows that, without teacher-forcing, end-to-end EEG2Text models fail to extract meaningful information from the training data irrespective of architecture (Jo et al., 2024). We therefore take an approach inspired by reinforcement learning. When the ultimate task becomes too challenging, causing the model to engage in passive learning, we can pragmatically decompose the overarching

Model	Input	ZuCo (1.0+2.0) BLEU-1		BLEU-2		BLEU-3		BLEU-4		ROUGE-1 F (%)		WER (%)		Emb.(%)		COFETT Emb.(%)
		w/o tf	w/tf	w/o tf	w/tf	w/o tf	w/tf	w/o tf	w/tf	w/o tf	w/tf	w/o tf	w/tf	w/o tf	w/tf	w/o tf
Brain-BART	**EEG**	13.75	39.12	2.88	21.98	0.81	12.46	0.37	7.22	11.84	28.55	108.24	77.97	5.98	63.89	**14.92**
	Noise	14.18	39.50	2.94	22.23	0.88	12.45	0.36	7.17	11.17	28.17	111.10	78.17	5.96	64.05	**5.96**
BELT	**EEG**	15.55	42.04	4.68	25.06	1.42	13.91	0.46	8.21	13.75	32.53	110.09	74.24	6.52	62.47	**16.72**
	Noise	15.52	42.16	4.35	25.03	1.09	13.95	0.55	8.29	13.83	32.83	109.82	75.38	6.40	62.28	**6.40**
Dewave	**EEG**	14.36	41.25	3.83	24.04	1.13	13.90	0.56	8.22	12.83	30.64	109.88	79.26	6.01	63.62	**15.21**
	Noise	14.42	41.62	3.96	24.24	1.03	14.07	0.38	8.37	13.03	30.87	110.06	79.25	5.92	63.88	**5.92**
Pegasus	**EEG**	8.39	38.09	2.57	21.20	0.89	11.69	0.34	6.07	0.10	28.29	99.87	78.38	4.52	61.10	**11.76**
	Noise	9.14	39.12	2.53	21.58	0.98	11.87	0.14	6.12	0.10	29.12	99.09	78.13	3.74	61.38	**3.74**
T5	**EEG**	16.59	43.33	5.83	25.60	2.12	15.25	0.75	8.76	11.76	24.96	111.17	81.23	7.21	64.85	**16.88**
	Noise	15.50	43.60	5.10	25.60	1.79	15.34	0.77	8.86	11.01	25.29	111.66	81.48	6.31	64.32	**6.31**
EEGnet	**EEG**	7.30	–	2.05	–	0.92	–	0.38	–	0.08	–	120.06	–	5.47	–	**16.29**
	Noise	7.54	–	2.05	–	0.92	–	0.38	–	0.08	–	121.06	–	5.02	–	**5.02**
EEG-Conformer	**EEG**	12.54	–	2.86	–	0.61	–	0.12	–	10.11	–	105.01	–	1.97	–	**9.13**
	Noise	11.30	–	2.83	–	0.26	–	0.27	–	9.65	–	106.13	–	1.44	–	**1.44**
EEGPT	**EEG**	16.34	43.00	6.19	26.77	1.76	15.29	0.80	8.92	14.92	31.96	109.85	80.09	7.04	65.01	**17.13**
	Noise	16.04	43.34	5.98	26.90	1.81	15.18	0.98	9.04	14.73	31.20	109.92	80.11	6.51	64.55	**6.51**

Table 1: Comparison of models on ZuCo and COFETT. 'w/o tf' denotes evaluation without teacher-forcing; 'w/tf' denotes evaluation with teacher-forcing. 'Emb.' refers to the Pearson correlation between EEG embeddings and text embeddings (Section 3.2). Underlined entries denote metrics commonly used in existing work; bold entries indicate our proposed methods; the remaining entries are our extensions of existing baselines.

goal into smaller, more manageable sub-goals. In tasks such as machine translation or EEG2Text, the semantic distance in the embedding space can serve as a proxy for this decomposition. Semantic distance is a key metric for evaluating the performance of multilingual models. It measures the distance in embedding space between different languages under similar semantic conditions, serving as an important indicator of machine translation performance (Bajpai and Chakraborty, 2024).

Accordingly, we adopt the model shown in Fig. 4. Rather than training EEG2Text with text tokens as the supervision signal, we supervise the model using text embeddings from a pretrained language model. At test time, we evaluate performance by the Pearson correlation between EEG embeddings and text embeddings, instead of BLEU between text outputs. To prevent data leakage, EEG data in the training and test sets were drawn from different sessions of the same participant, and a separate model was trained for each participant.

3.3 Feasibility Evidence

We test whether EEG carries linguistic information by comparing the similarity of signals when the same participant imagines the same sentence at different times with the similarity when the same participant imagines different sentences at different times. COFETT uniquely enables this analysis by providing within-participant, repeated renditions of identical sentences—capability absent from existing benchmarks. We quantify EEG similarity using 'MINDFUL' (Pun et al., 2024) and 'CCA' (Dmochowski et al., 2012), with full mathematical details in the Appendix.

From a complementary angle, we also apply an LLM-style scaling analysis (Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022): we vary the proportion of training data and examine whether model performance scales predictably with data size. Subsequent Experiments report the results.

4 Experiments

4.1 Benchmark Comparison

Most existing EEG2Text studies evaluate on the ZuCo corpus and report BLEU, ROUGE-1 or WER under teacher-forcing. Accordingly, we conducted controlled comparisons on ZuCo 1.0/2.0, replacing EEG with pure random noise at both train-

Model	Input	BLEU-1		BLEU-2		BLEU-3		BLEU-4		ROUGE-1 F (%)		WER (%)		Emb.	
		w/o rep	*w/ rep*	*w/o rep*	*w/ rep*	*w/o rep*	*w/ rep*	*w/o rep*	*w/ rep*	*w/o rep*	*w/ rep*	*w/o rep*	*w/ rep*	*w/o rep*	*w/ rep*
Brain-BART	**EEG**	13.77	14.36	2.90	3.03	0.82	0.95	0.36	0.42	11.83	11.33	108.26	110.30	6.03	**14.92**
	Noise	14.18	14.18	2.94	2.94	0.88	0.88	0.36	0.36	11.17	11.17	111.10	111.10	5.96	**5.96**
BELT	**EEG**	15.57	15.73	4.65	4.47	1.44	1.20	0.47	0.64	13.77	14.00	110.07	109.10	6.49	**16.72**
	Noise	15.52	15.52	4.35	4.35	1.09	1.09	0.55	0.55	13.83	13.83	109.82	109.82	6.40	**6.40**
Dewave	**EEG**	14.35	14.58	3.85	4.06	1.12	1.12	0.55	0.46	12.84	13.21	109.90	109.41	5.98	**15.21**
	Noise	14.42	14.42	3.96	3.96	1.03	1.03	0.38	0.38	13.03	13.03	110.06	110.06	5.92	**5.92**
Pegasus	**EEG**	8.41	9.27	2.56	2.60	0.90	1.04	0.33	0.19	0.11	0.16	99.86	98.54	4.32	**11.76**
	Noise	9.14	9.14	2.53	2.53	0.98	0.98	0.14	0.14	0.10	0.10	99.09	99.09	3.74	**3.74**
T5	**EEG**	16.58	15.70	5.85	5.22	2.13	1.90	0.76	0.85	11.78	11.16	111.19	110.76	7.08	**16.88**
	Noise	15.50	15.50	5.10	5.10	1.79	1.79	0.77	0.77	11.01	11.01	111.66	111.66	6.31	**6.31**
EEGnet	**EEG**	7.32	7.62	2.05	2.12	0.94	0.98	0.36	0.41	0.10	0.11	120.05	119.80	5.67	**16.29**
	Noise	7.54	7.54	2.05	2.05	0.92	0.92	0.38	0.38	0.08	0.08	121.06	121.06	5.02	**5.02**
EEG-Conformer	**EEG**	12.56	11.45	2.88	2.92	0.63	0.32	0.12	0.30	10.13	9.82	105.00	104.60	2.01	**9.13**
	Noise	11.30	11.30	2.83	2.83	0.26	0.26	0.27	0.27	9.65	9.65	106.13	106.13	1.44	**1.44**
EEGPT	**EEG**	16.33	16.23	6.20	6.08	1.77	1.89	0.79	1.05	14.93	14.90	109.86	109.12	6.96	**17.13**
	Noise	16.04	16.04	5.98	5.98	1.81	1.81	0.98	0.98	14.73	14.73	109.92	109.92	6.51	**6.51**

Table 2: COFETT Benchmark ablation results. w/o rep denotes the ablation group without within-participant repetitions, whereas w/ rep denotes the control group with within-participant repetitions.

ing and test time to assess whether scores change. For models, we evaluate three EEG2Text systems, BrainBART (Wang and Ji, 2022), BELT (Zhou et al., 2024) and DeWave (Duan et al., 2023), two general-purpose Transformer baselines, Pegasus (Zhang et al., 2020) and T5 (Raffel et al., 2020), two EEG feature-extraction architectures, EEGConformer (Song et al., 2022) and EEGNet (Lawhern et al., 2018), and a decoder-style large EEG foundation models, EEGPT (Wang et al., 2024a), to cover the full range of model classes. To extract language embeddings, we employ LaBSE (Feng et al., 2022), a multilingual model designed for sentence-level representation.

Additionally, to strengthen the comparison, we extended the conventional setups with vector-alignment variants and non-teacher-forcing evaluation. As shown in Table 1, across all ZuCo 1.0/2.0 conditions, scores obtained with true EEG are not statistically distinguishable from those obtained with noise. This indicates that the canonical ZuCo benchmark primarily reflects models' language priors, rather than their use of EEG, and thus cannot demonstrate learning of meaningful neural information.

By contrast, on our COFETT benchmark with teacher-forcing-free evaluation, performance separates clearly across methods and collapses under the same noise controls, indicating that COFETT discriminates models by their ability to exploit EEG rather than by linguistic priors.

4.2 Benchmark Ablations

We base the benchmark on two design choices that jointly enable teacher-forcing-free evaluation. First, within-participant exact repetitions expose the invariant, language-bearing component of EEG by implicitly averaging over time-varying noise, thereby mitigating EEG instability and improving the statistical reliability of comparisons across methods. Second, a sentence-level embedding-alignment metric evaluates semantic compatibility between EEG-derived representations and text under teacher-forcing-free decoding, which curbs negative learning behaviors that arise with token-level or n-gram scoring and discourages solutions that exploit language priors while ignoring EEG. To verify the necessity of these ingredients, we conduct two ablations under the same teacher-forcing-free protocol.

First, we remove within-participant repetitions and add new, unique sentences to keep the total EEG duration constant; this choice diminishes the benchmark's discriminative power and destabilizes method rankings. Second, we replace the embedding-alignment objective with teacher-

ID	Condition			MINDFUL	CCA			
	Subject	Time	Text		I(%)	II(%)	III(%)	Mean
1	Same	Different Day	Same	1.64	17.42	18.58	18.81	18.27
2	Same	Different Day	Different	1.67	16.59	18.00	18.52	17.70
3	Same	Close Time	Same	1.54	23.23	20.81	21.17	21.74
4	Same	Close Time	Different	1.79	23.17	19.97	21.30	21.48
5	Different	-	Same	1.96	13.55	14.41	14.90	14.29
6	Different	-	Different	2.07	13.48	14.98	15.82	14.76

Table 3: Similarity of EEG signals under various conditions. For instance, ID 1 represents the EEG similarity of the same participant attempting to articulate the same text on different days. Higher CCA values and lower MINDFUL scores indicate greater similarity between EEG signals.

forcing-free BLEU, ROUGE, and WER, which increases dependence on the pretrained language backbone and weakens sensitivity to EEG versus noise controls. The experiments show that a clean separation between models is achieved only when both core components are retained, as summarized in Table 2.

4.3 Statistical Evidence of Linguistic Information in EEG

To quantify whether EEG carries recoverable linguistic signal, we estimate EEG similarity with two complementary, label-free metrics and aggregate their outcomes via Monte Carlo sampling. Specifically, for many randomly drawn pairs of EEG we compute (i) MINDFUL (Pun et al., 2024), a recent instability/similarity measure designed for long-term neural interfaces that detects distributional changes without intention labels, and (ii) CCA (Dmochowski et al., 2012), which extracts components that are maximally correlated across paired recordings and thereby indexes reliable shared structure. We repeat this procedure across six pairing conditions (spanning participant identity and time separation, and controlling the sentence identity) and summarize the results (Table 3).

The pattern is consistent across both metrics: similarity is lowest across different participants, increases within the same participant across days, and is highest within the same participant on the same day. After conditioning on participant and time, sentence identity contributes only a small additional effect, indicating that—within our task—the time-varying physiological and hardware factors dominate the variance, whereas the language-related component is comparatively weak. This is compatible with biological accounts of organismal regulation and homeostasis (Bechtel and Bich, 2024): neural systems actively resist rapid internal fluctuations, so adjacent measurements can appear similar even when the imagined sentence differs.

Taken together, these results indicate that EEG contains linguistically decodable structure, even if weak, and that decoding is feasible.

4.4 Scaling-Law Perspective on Linguistic Information in EEG

To further substantiate the presence of linguistic components in EEG and to quantify how within-participant repetitions and data scale affect vector alignment, we partition the corpus by recording day. For each participant, data from Days 1–4 constitute $\mathcal{D}_1$, and data from Days 5–8 constitute $\mathcal{D}_2$. We then split $\mathcal{D}_2$ into a training set $\mathcal{D}_2^{\text{train}}$ and a test set $\mathcal{D}_2^{\text{test}}$.

$$\mathcal{D}_2^{\text{train}} = \{(EEG_i, S_i) \mid i = 1, 2, \ldots, M', \\ EEG_i = EEG_{i1}, EEG_{i2}, \ldots, EEG_{ij}\}.$$

$$\mathcal{D}_2^{\text{test}} = \{(EEG_i, S_i) \mid i = 1, 2, \ldots, M, \\ EEG_i = EEG_{i15}, EEG_{i16}\}.$$

Here, the test set $\mathcal{D}_2^{\text{test}}$ consists of two repetitions for every sentence in the corpus, while the training set $\mathcal{D}_2^{\text{train}}$ contains j repetitions drawn from a subset of sentences M', where $M' \subset M$. We scale data by jointly varying the subset size $|M'|$ and the repetition count j while keeping $\mathcal{D}_2^{\text{test}}$ fixed, enabling an LLM-style scaling-law analysis on EEG.

In this section, we select EEGPT(Wang et al., 2024a), the best-performing model in the previous experiments, as the EEG encoder, together with LaBSE(Feng et al., 2022) for text representation.

As shown in Figure 5(a), alignment performance increases with both the repetition count j and the number of sentences M', with the effect of M' exceeding that of j. To examine the regime $M' > M$, we proportionally augment the training set with samples from $\mathcal{D}_1$. The results in Figure 5(b) show no further gains once M' exceeds M. We attribute this to an in-/out-of-domain effect: $\mathcal{D}_2^{\text{test}}$ and $\mathcal{D}_2^{\text{train}}$

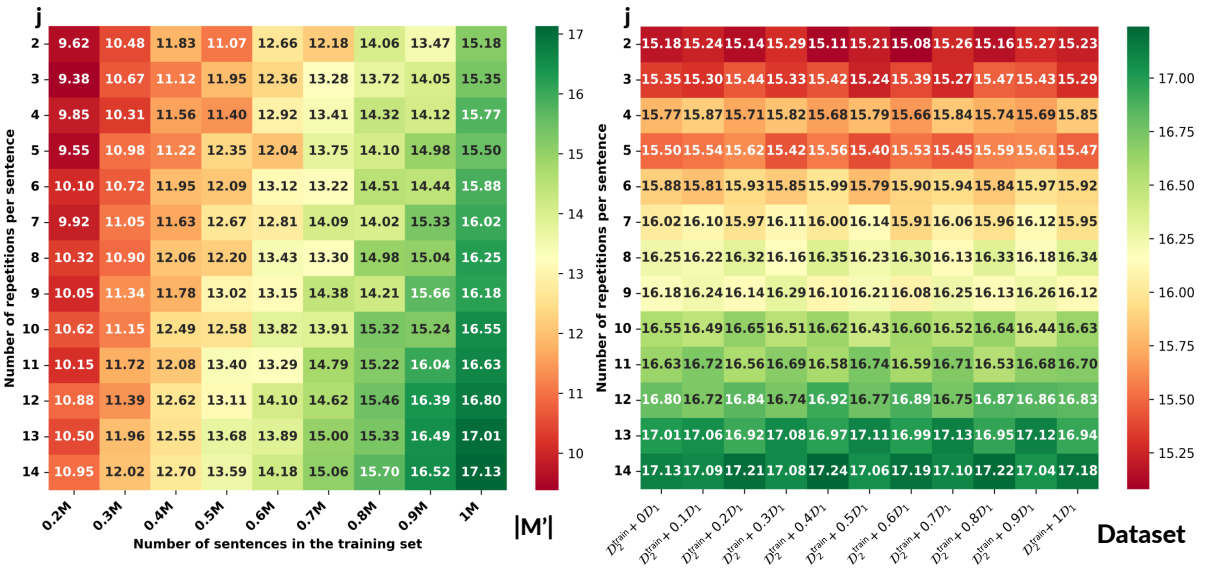

Figure 5: Alignment performance at different data scales.

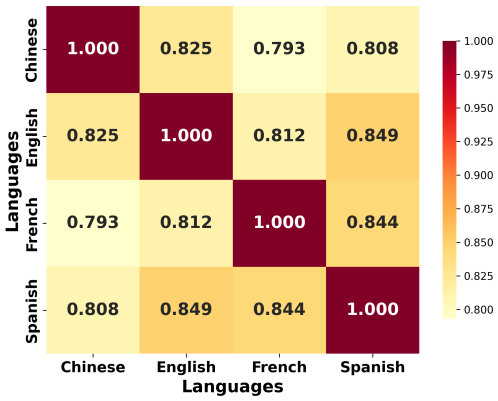

Figure 6: Embedding distances of the same sentence expressed in different languages.

are in-domain when both are drawn from the same sentence set M; sentences added beyond M are out of domain and do not improve EEG2Text performance.

Taken together, these findings not only support the presence of linguistically decodable information in EEG, but also indicate that training and test sets must be drawn from the same semantic domain. This helps explain prior failures on ZuCo: ZuCo is a natural-reading corpus in which stories progress semantically rather than repeat semantically similar content, rendering many samples effectively out of domain with respect to one another.

4.5 Predictive Analysis

In this section, we explore an intriguing question: to what extent are EEG2Text and current machine translation systems comparable in performance? To examine this, we compare the embedding distances of the same sentence expressed in different languages. As shown in comparison to Figure 6, it is clear that in existing multilingual pretraining models the semantic distance between EEG and natural language remains substantially larger than the distances observed between different natural languages.

We now provide a rough projection under the linear trend inferred above. If the in-domain upper envelope is ≈ 0.17, then gains must primarily come from increasing within-participant repetitions. Extrapolating to ≈ 0.85 at the lower right of Figure 6 suggests ~ 700 repetitions per sentence, i.e., a $\sim 40\times$ increase over our current setting. This means that EEG2Text requires $\sim 40\times$ more paired samples to approach machine-translation-like performance, without accounting for diminishing returns at larger scales and the instability of EEG over longer time scales. This estimate highlights the need for more EEG–text resources.

5 Conclusion

We provide a systematic synthesis of the emerging EEG2Text literature and identify a core limitation of current benchmarks: their design obscures feasibility by ignoring EEG instability, thereby fuelling the field's central debate. In response, we introduce a new benchmark COFETT, comprising two components: a within-subject repetition paradigm and a vector-alignment evaluation protocol. These enables teacher-forcing-free assessment. Using this framework, we obtain convergent

evidence that EEG2Text is a feasible and scientifically meaningful direction and the EEG2Text model holds promise for practical applications.

Limitations

This work centres on benchmarking rather than proposing new architectures; as such, we did not explore the model-design space exhaustively. It remains possible that alternative architectures or training regimes could further reduce the data requirements observed in our predictive analyses. Nevertheless, any ultimate ceiling on performance is constrained by the signal-to-noise ratio of EEG, which imposes a theoretical upper bound on decodable linguistic information.

Ethics Statement

This study involved human participants in the collection of EEG data. All participants provided written informed consent for participation and data reuse prior to the experiments. The study was reviewed and approved by the Ethics Commission of Harbin Institute of Technology (Approval Number: HIT2024035) and was conducted in accordance with the Declaration of Helsinki (2013). The study involved only non-invasive EEG recordings and standard cognitive tasks, and therefore posed minimal risk to participants.

Acknowledgments

The research in this article is supported by the New Generation Artificial Intelligence of China (2024YFE0203700), National Natural Science Foundation of China under Grants U22B2059 and 62576124.

References

Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. 2013. Eeg data space adaptation to reduce intersession nonstationarity in brain-computer interface. *Neural computation*, 25(8):2146–2171.

Ashutosh Bajpai and Tanmoy Chakraborty. 2024. Multilingual llms inherently reward in-language time-sensitive semantic alignment for low-resource languages. *arXiv preprint arXiv:2412.08090*.

William Bechtel and Leonardo Bich. 2024. Situating homeostasis in organisms: maintaining organization through time. *The Journal of Physiology*, 602(22):6003–6020.

Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A Robbins. 2015. The prep pipeline: standardized preprocessing for large-scale eeg analysis. *Frontiers in neuroinformatics*, 9:16.

Nicholas S Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R Willett, Erin M Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R Deo, and 1 others. 2024. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618.

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

Jacek P Dmochowski, Paul Sajda, Joao Dias, and Lucas C Parra. 2012. Correlated components of ongoing eeg point to emotionally laden attention–a possible marker of engagement? *Frontiers in human neuroscience*, 6:112.

John E Downey, Nathaniel Schwed, Steven M Chase, Andrew B Schwartz, and Jennifer L Collinger. 2018. Intracortical recording stability in human brain–computer interface users. *Journal of neural engineering*, 15(4):046016.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. 2023. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36:9907–9918.

Jin Fan, Bruce D McCandliss, Tobias Sommer, Amir Raz, and Michael I Posner. 2002. Testing the efficiency and independence of attentional networks. *Journal of cognitive neuroscience*, 14(3):340–347.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. 2014. Mne software for processing meg and eeg data. *neuroimage*, 86:446–460.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, and 1 others. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2019. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.

Aapo Hyvärinen. 1997. Independent component analysis by minimization of mutual information.

Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. 2017. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, 159:417–429.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. 2024. Are eeg-to-text models working? *arXiv preprint arXiv:2405.06459*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013.

Seo-Hyun Lee, Minji Lee, and Seong-Whan Lee. 2020. Neural decoding of imagined speech and visual imagery as intuitive paradigms for bci communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2647–2659.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xinyu Mou, Cuilin He, Liwei Tan, Junjie Yu, Huadong Liang, Jianyu Zhang, Yan Tian, Yu-Fang Yang, Ting Xu, Qing Wang, and 1 others. 2024. Chineseeeg: A chinese linguistic corpora eeg dataset for semantic alignment and neural decoding. *Scientific Data*, 11(1):550.

Saydul Akbar Murad and Nick Rahimi. 2024. Unveiling thoughts: A review of advancements in eeg brain signal decoding into text. *IEEE Transactions on Cognitive and Developmental Systems*.

Nicolás Nieto, Victoria Peterson, Hugo Leonardo Rufiner, Juan Esteban Kamienkowski, and Ruben Spies. 2022. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. *Scientific Data*, 9(1):52.

Harry Nyquist. 1928. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644.

Jonathan W Peirce, Jeremy R Gray, Samuel Simpson, Max R MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas K Lindeløv. 2019. Psychopy2: experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

Timothée Proix, Jaime Delgado Saa, Andy Christen, Stephanie Martin, Brian N Pasley, Robert T Knight, Xing Tian, David Poeppel, Werner K Doyle, Orrin Devinsky, and 1 others. 2022. Imagined speech can be decoded from low-and cross-frequency intracranial eeg features. *Nature communications*, 13(1):48.

Tsam Kiu Pun, Mona Khoshnevis, Tommy Hosman, Guy H Wilson, Anastasia Kapitonava, Foram Kamdar, Jaimie M Henderson, John D Simeral, Carlos E Vargas-Irwin, Matthew T Harrison, and 1 others. 2024. Measuring instability in chronic human intracortical neural recordings towards stable, long-term brain-computer interfaces. *Communications Biology*, 7(1):1363.

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

Wojciech Samek, Carmen Vidaurre, Klaus-Robert Müller, and Motoaki Kawanabe. 2012. Stationary common spatial patterns for brain–computer interfacing. *Journal of neural engineering*, 9(2):026013.

Alexander B Silva, Kaylo T Littlejohn, Jessie R Liu, David A Moses, and Edward F Chang. 2024. The speech neuroprosthesis. *Nature Reviews Neuroscience*, 25(7):473–492.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. 2022. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719.

Yi-Yuan Tang, Yinghua Ma, Junhong Wang, Yaxin Fan, Shigang Feng, Qilin Lu, Qingbao Yu, Danni Sui, Mary K Rothbart, Ming Fan, and 1 others. 2007. Short-term meditation training improves attention and self-regulation. *Proceedings of the national Academy of Sciences*, 104(43):17152–17156.

Xing Tian and David Poeppel. 2010. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in psychology*, 1:7029.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. 2024a. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. 2024b. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7278–7292.

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

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, and 1 others. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. 2023. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13277–13291.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. 2024. Chisco: An eeg-based bci dataset for decoding of imagined speech. *Scientific Data*, 11(1):1265.

Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. 2024. Belt: bootstrapped eeg-to-language training by natural language supervision. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

A Participants and pre-experiment

EEG data were collected from two subjects: S1 (male, age 26) and S2 (female, age 25). Neither participant reported a history of neurological or psychiatric conditions. All participants were right-handed, had normal hearing, and either had normal or corrected-to-normal vision. None of them had previous experience with brain-computer interfaces (BCIs).

To improve the quality of Inner Speech EEG data, we first conducted a screening process. Seven volunteers (aged 22–47, Chinese native speakers, 4 males) were initially recruited for a five-day session of Integrative Body-Mind Training (IBMT) (Tang et al., 2007), a meditation technique designed to enhance concentration. Previous research has shown that IBMT significantly improves attention capacity and reduces fatigue. Additionally, IBMT training has been associated with increased vigor, as measured by the Profile of Mood States scale, and a marked reduction in stress-related cortisol levels (Tang et al., 2007). This training was aimed at preparing participants for the cognitively demanding experimental paradigm.

After completing the IBMT session, all participants took the Attention Network Test (ANT) (Fan et al., 2002), a psychological assessment that evaluates attentional functions. The ANT measures three key indices: alerting (readiness for an impending event), orienting (ability to focus attention on a specific spatial location), and executive control (capacity to handle conflicting information). Lower ANT scores indicate greater attentional focus. We calculated the mean scores of all participants across the three ANT subtests and selected the two participants with the lowest scores for further experimentation. Subsequently, the two selected participants underwent a Chinese language proficiency test (HSK Level 6 Reading Section) to ensure they could accurately comprehend the text presented during the experiments. Details of each participant are provided in Table 4. No participants withdrew during the experiment.

Upon completion of the experiment, all seven participants received a base compensation of $50. Additionally, two participants selected to complete the full experiment received an extra $400, which, when calculated on an hourly basis, is slightly above the average hourly wage in the region where the experiment was conducted.

Variable	S1	S2	S3	S4	S5	S6	S7
Gender	male	female	male	female	male	male	female
Age	26	25	22	47	40	42	28
ANT(Alerting)	25.1	20.3	30.5	55.2	68.9	50.4	30.5
ANT(Orienting)	40.8	45.7	30.6	110.3	25.0	85.2	60.4
ANT(Executive Control)	70.4	87.3	95.7	115.1	140.9	100.8	120.6
ANT(Mean)	45.4	51.1	52.3	93.5	78.3	78.8	70.5
HSK(Reading Section)	97.0	95.0	-	-	-	-	-

Table 4: Details of all subjects in the study.

B Textual stimuli

The textual stimuli presented to the participants were sourced from the Chisco (Zhang et al., 2024), which comprises daily language texts in Chinese across 39 semantic categories. As shown in Figure 2, our experimental paradigm consists of two sets of text data. Text 1 contains the complete set of 8 semantic categories. Text 2 includes the complete set of 2 semantic categories. For convenience, a description of the Chisco text data (Zhang et al., 2024) has been provided as follows.

To ensure the dataset is suitable for training BCIs for everyday use, the textual materials were designed to encompass a broad range of daily language. To achieve this, Chisco manually selected expressions from the Chinese social media platform Weibo, as well as public datasets ROCstory (Mostafazadeh et al., 2016) and Dailydialog (Li et al., 2017). These expressions were initially categorized into 39 categories using a combination of machine learning clustering algorithms and manual annotation by human experts. The expressions were then rephrased into sentences of 6 to 15 Chinese characters through a crowdsourcing approach, resulting in a dataset representing daily Chinese expressions.

The crowdsourcing process adhered to the following criteria to ensure the quality and usability of the text data:

- Each sentence must be independently understandable, without requiring multiple segments to form a coherent discourse.
- Sentences must avoid clauses and indirect anaphora.
- The number of nouns in a sentence must not exceed three.
- Ambiguous sentences were excluded.
- Texts containing any form of racial or gender discrimination were prohibited.

This approach was designed to minimize syntactic complexity, thereby reducing the likelihood of errors during the imagined speech process. Examples of sentences with their categories include:

- "Today's dinner tasted great" - *Food and Dining*
- "She is going to listen to the concert" - *Performing Arts*
- "I am deeply sorry for my behavior" - *Apologies*

C EEG data acquisition

The study was conducted in an electrically shielded room. Participants were seated comfortably in front of a computer screen positioned 80 cm away, directly facing its center. Stimuli were presented on a 26-inch screen with a refresh rate of 60 Hz and a resolution of 1920×1080 pixels. The visual stimuli consisted of sentences from everyday language, ranging from 6 to 15 Chinese characters in length. These sentences were displayed in a 28-point Song typeface, on a single line, centrally placed on a grey background with white text to minimize glare and eye strain. During breaks, participants were provided with snacks and water, and encouraged to rest.

To ensure consistent data quality, a personalized head case was designed for each participant with three main objectives: (1) to control head movements and reduce noise from electrode displacement; (2) to provide head support, alleviating physical fatigue during prolonged sessions; and (3) to maintain consistent electrode positioning across multiple sessions, as participants underwent 8 days of experiments. The head case facilitated the reapplication of the EEG cap, minimizing internal variance in the data.

EEG data were collected using the SynAmps-2 128-channel amplifier and the 128-channel Quik-

Cap, both manufactured by Compumedics Neuroscan. The Curry 9 software, also from Compumedics Neuroscan, was used for hardware control and impedance checks before each experiment, ensuring impedances remained below 25 kΩ. The conductive GREENTEK® GT5 Gel was used to fill the gap between the scalp and the electrodes.

After each experimental session, PC2 generated and stored an EDF file containing continuous recordings from 125 EEG channels, 6 external channels (for vertical and horizontal electrooculography signals, as well as voltages at the mastoids on both sides), and marker signals. The experiments were conducted with a sampling rate of 1 kHz in alternating current (AC) mode, with an accuracy of 3 nV/LSB. Stimulus presentation was managed using the open-source software package PsychoPy (Peirce et al., 2019) (v2023.2.3).

D Quality control of EEG data

To help participants better adapt to the experimental paradigm, each participant underwent a pre-experiment training session prior to the main experiment. This session allowed participants to become familiar with the procedure, which involved word-by-word reading, followed by imagination, and then a rest period.

In addition, to ensure the quality of the data collected, participants' attention was monitored throughout the experiment. Random attention checks were incorporated into the paradigm: at the end of 10 randomly selected trials in each session, participants were required to verbally recall the content they had memorized. The accuracy of these recitals was evaluated by the experimenters, and a recall error was defined as a discrepancy of more than four characters from the original text. If a participant made two or more errors in these 10 checks, they were considered to have experienced lapses in concentration during that period. A rest period was then provided to help restore attention before restarting the session.

E EEG data preprocessing

Preprocessing was performed in Python, primarily utilizing the MNE library (Gramfort et al., 2014). Minimal preprocessing was applied to retain the maximum amount of valid information, allowing for flexibility in further processing tailored to specific research needs. The detailed steps of the preprocessing pipeline are described as follows, with an equivalent flowchart provided in the supplementary information.

Resampling: The raw data stored in the .edf file contained both EEG signals and event markers, sampled at 1,000 Hz. We resampled the data to 500 Hz. Given ongoing debates regarding the frequency bands associated with Inner Speech (Lee et al., 2020; Proix et al., 2022; Crone et al., 2006), we employed a conservative downsampling approach. According to the Nyquist sampling theorem (Nyquist, 1928), the 500 Hz sampling rate captures EEG information up to 250 Hz, which includes the ultra-high gamma band (Crone et al., 2006).

The PREP pipeline (Bigdely-Shamlo et al., 2015): The PREP pipeline is a standardized early-stage EEG processing procedure, applied after data resampling. The key steps of the PREP pipeline are as follows:

1. Removing line noise without committing to a specific filtering strategy.
2. Robust referencing of the signal relative to an estimate of the "true" average reference.
3. Detection and interpolation of bad channels relative to this reference.
4. Retaining sufficient information to allow users to re-reference or undo interpolation of specific channels.

Filtering: In this step, we applied a 50 Hz notch filter to remove power-line noise and a zero-phase high-pass finite impulse response (FIR) filter with a 1 Hz cutoff. No low-pass filtering was performed, as explained in the 'Resampling' section.

Autoreject (Jas et al., 2017): The Autoreject algorithm was used to automatically identify and reject bad data segments and artifacts. This method optimizes rejection thresholds on a per-channel basis, ensuring an adaptive cleaning process that minimizes the loss of valid data.

Remove Ocular Artifacts: To effectively remove ocular artifacts from the EEG signals, we employed a multivariate linear regression model. In this approach, the horizontal electrooculography (H-EOG) and vertical electrooculography (V-EOG) signals were used as regressors. Specifically, by inputting the H-EOG and V-EOG signals into the regression model, we quantified the contribution of these two ocular movements to the EEG signal. The multivariate linear regression model was used to fit the linear relationship between the EOG signals and the EEG data, allowing for the estimation of

their respective regression coefficients. These coefficients were then used to predict the ocular artifact components, which were subsequently subtracted from the original EEG data, resulting in purified signals. This method not only improved the signal-to-noise ratio of the EEG data but also enhanced the accuracy and reliability of subsequent neuroelectric analyses.

Independent Component Analysis (ICA) (Hyvärinen, 1997): ICA is a widely used blind source separation method to remove artifacts from EEG signals. For our dataset, ICA was applied only to the EEG channels using the MNE implementation of extended Infomax ICA. Due to the wide frequency range of the data, we set the number of independent components to 30, which is higher than in other neural language decoding studies (Mou et al., 2024), ensuring that the components capture the majority of relevant information. Noise component identification was conducted using MNE-ICALabel automatic annotation, followed by manual evaluation to ensure accuracy without introducing excessive manual processing.

Data Segmentation: The continuous EEG data were segmented into three types of segments based on the experimental paradigm: reading, Inner Speech, and meditation segments.

F Mindful

The Mindful algorithm follows a pipeline process, beginning with feature extraction from two EEG signals, followed by dimensionality reduction using Principal Component Analysis (PCA) and Z-score normalization. The Kullback-Leibler (KL) divergence between the two segments is then computed, which can be mathematically represented as:

$$F_i = \text{Normal}(\text{PCA}(\text{Extract\_Features}(x_i))), i = 1, 2$$

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} F_i^{(j)}$$

$$\Sigma_i = \frac{1}{N-1} \sum_{j=1}^{N} \left(F_i^{(j)} - \mu_i\right) \left(F_i^{(j)} - \mu_i\right)^\top$$

$$\text{Mindful}(x_1, x_2) = D_{\text{KL}} \left(\mathcal{N}(\mu_1, \Sigma_1) \,\|\, \mathcal{N}(\mu_2, \Sigma_2)\right)$$

G Correlated Components Analysis (CCA)

The computation of CCA involves the following six steps. The two input EEG signal sets are denoted as X_1 and X_2. The outputs from Step 6, ρ_1, ρ_2, and ρ_3, correspond to CCA-I, CCA-II, and CCA-III as mentioned in Figure 1(c).

1. **Covariance Matrix Computation**

   Description: Compute the covariance matrices of the input data matrices $X1$ and $X2$.

   Mathematical Representation: Let T be the number of time steps or samples. The covariance matrices are calculated as:

   $$R_{11} = \frac{1}{T} X1 \cdot X1^T$$

   $$R_{12} = \frac{1}{T} X1 \cdot X2^T$$

   $$R_{22} = \frac{1}{T} X2 \cdot X2^T$$

2. **Eigenvalue Problem Solving**

   Description: Construct matrix M and find its eigenvalues and eigenvectors.

   Mathematical Representation:

   $$M = (R_{11} + R_{22})^{-1}(R_{12} + R_{12}^T)$$

   Solve the eigenvalue equation:

   $$M\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

   where λ_i are the eigenvalues and $\mathbf{v}_i$ are the corresponding eigenvectors.

3. **Extract and Sort Eigenvalues and Eigenvectors**

   Description: Sort the eigenvalues in descending order and reorder the eigenvectors accordingly to prioritize the most significant components.

   Mathematical Representation: Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$ be the sorted eigenvalues, and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K$ be the corresponding eigenvectors.

4. **Extract Principal Components**

   Description: Select the top three eigenvectors corresponding to the largest eigenvalues to form the projection matrix W.

   Mathematical Representation:

   $$W = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$$

5. **Projection into New Feature Space**

   Description: Project the original data matrices $X1$ and $X2$ into the new feature space defined by the projection matrix W.

   Mathematical Representation:

$$C1_{\text{full}} = W^T X1$$

$$C2_{\text{full}} = W^T X2$$

   Here, $C1_{\text{full}}$ and $C2_{\text{full}}$ represent the projected data in the new feature space.

6. **Compute Pearson Correlation Coefficients**

   Description: For each of the first three projected components, calculate the Pearson correlation coefficient between $C1_{\text{full}}$ and $C2_{\text{full}}$. If the standard deviation of any component is zero (indicating a constant), set the correlation coefficient to zero.

   Mathematical Representation: For each component $j = 1, 2, 3$:

$$\rho_j = \text{corr}\left(C1_{\text{full}}[j, :], C2_{\text{full}}[j, :]\right)$$

   If $\sigma\left(C1_{\text{full}}[j, :]\right) = 0$ or $\sigma\left(C2_{\text{full}}[j, :]\right) = 0$, then set $\rho_j = 0$.

H EEG Experimental Setup

To familiarize the participant with the experimental procedure and the room environment, a detailed explanation of all experimental steps was provided during the placement of the EEG cap and external electrodes. This setup process took approximately 30 minutes. Figure 7 shows the main experiment setup. During the experiment, stimuli were presented to the subjects via a screen connected to PC1. The EEG cap recorded the signals, which were filtered and amplified by the headbox. These signals were further amplified by an external amplifier and tagged with markers generated by PC1 to enhance signal integrity. The marker signals enabled accurate segmentation of the EEG data. The processed EEG signals were then transmitted to the experimenter's computer (PC2) for storage, with the experimental operator monitoring the session through PC2. To minimize interference with the subject, only the PC1 screen was placed inside the data collection room, while all other equipment was located in the control room.

I EEG Data Preprocessing Pipeline

Raw EEG data were initially recorded at a sampling rate of 1,000 Hz and stored in the original .edf format. Subsequently, the data were downsampled to 500 Hz to facilitate processing. The PREP algorithm was then applied to detect and remove bad channels and to perform re-referencing. Following this, power line noise was eliminated, and a high-pass filter was applied to further enhance signal quality. The data were then segmented according to the experimental paradigm. Bad data spans and artifacts were identified and rejected using the Autoreject algorithm, as previously described by Jas et al. (2017). Finally, independent component analysis (ICA) was employed to remove physiological noise sources such as electrooculogram (EOG) and electromyogram (EMG) artifacts, resulting in a clean dataset ready for further analysis. The complete process is shown in the Figure 8.

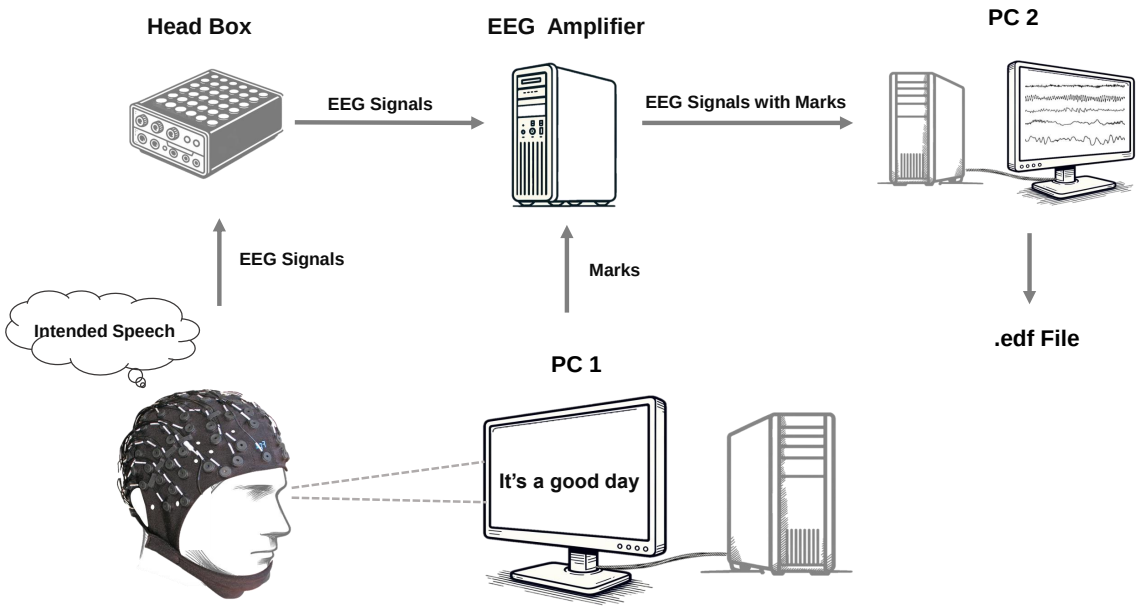

Figure 7: Main experimental setup.

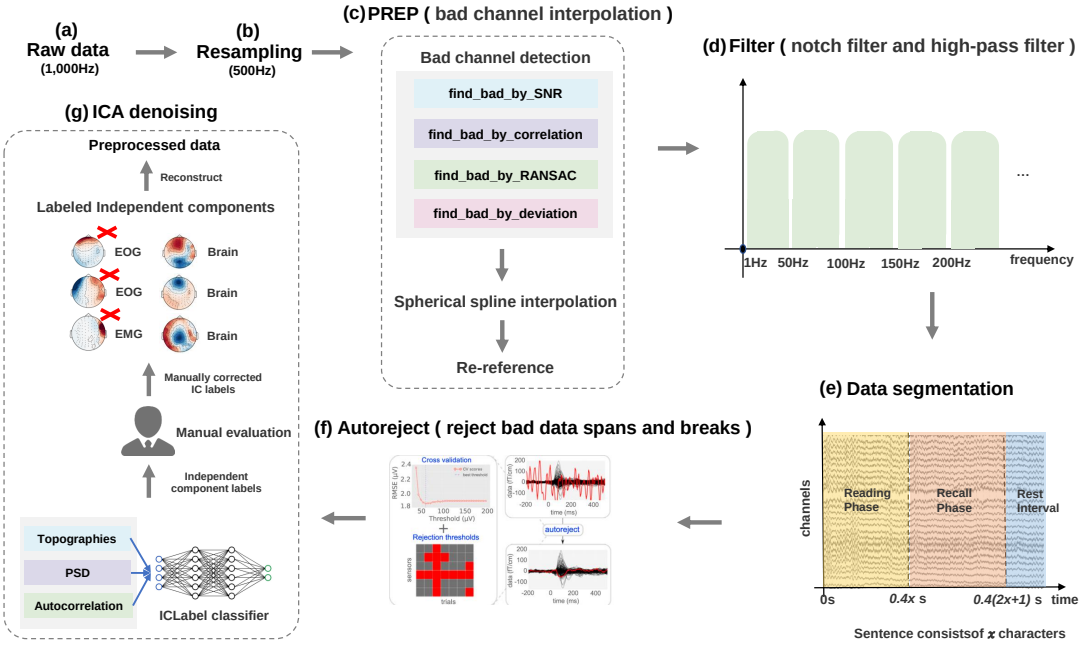

Figure 8: EEG preprocessing pipeline (part (f) cites previous work(Jas et al., 2017)). (a) Raw data were recorded at a sampling rate of 1,000 Hz, stored in the original .edf files. (b) The raw data were then downsampled to 500 Hz. (c) The PREP algorithm was applied for bad channel detection and re-referencing. (d) Power line noise was removed, and high-pass filtering was performed. (e) The data were segmented according to the experimental paradigm. (f) Bad data spans and breaks were rejected using the Autoreject algorithm. (g) ICA algorithm was used to remove noise such as EOG and EMG.