Thinking of LLM as a Brain: Decoding Text by Aligning Semantic in EEG Signals and Language

Anonymous ACL submission

Abstract

With the rapid development of brain-computer interfaces (BCI) in recent years, the electroencephalography (EEG) to text task has drawn increasing attention. Previous methods have defined it as a sequence-to-sequence translation task. However, their models, trained using teacher-forcing strategies, fail to extract and utilize EEG information. To address this issue, we propose a novel framework in this paper, which innovatively treats the EEG-to-text task as a fine-grained controllable text generation task. Specifically, since large language models (LLMs) have strong text generation capabilities, we treat the LLM as a "brain" and guide it to generate desired sentences by leveraging EEG representations that are aligned with the semantic space of text. Therefore, our approach focuses on training an EEG representation model that can effectively align EEG representation with text semantics, avoiding the limitations introduced by teacher-forcing strategies. Extensive experiments on the ZuCo benchmark demonstrate the effectiveness of our approach, which achieves state-of-the-art performance in multi-subject and single-subject settings. Furthermore, experimental results in cross-subject scenarios further verify that our method has a strong generalization ability and can be applied to unseen subjects.

1 Introduction

Devastating neurological diseases and traumatic brain injuries can cause patients to lose the ability to communicate (Stanger and Cawley, 1996; Pels et al., 2017), even though their language and cognitive abilities may remain intact. Therefore, the high expectations for brain-computer interfaces (BCIs) have been raised. Some recent studies (Anumanchipalli et al., 2019; Sun et al., 2019; Willett et al., 2021) in this field have demonstrated that we can decode brain signals into concrete and understandable representations, such as speech or text.

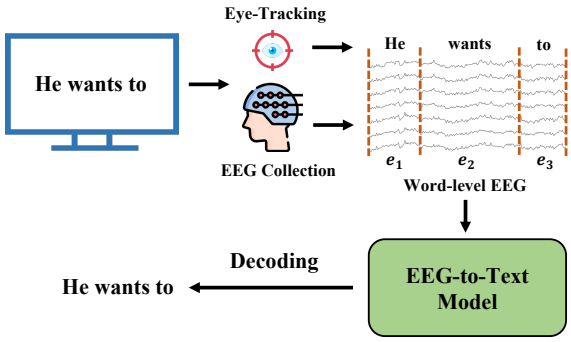

Figure 1: Illustration of the EEG-to-Text task. A subject reads a sentence on a screen while EEG signals are collected and eye-tracking is performed. Eye-tracking fixations are used to divide EEG segments corresponding to different words. These word-level EEG segments are input into the model to decode the sentences.

However, these works usually rely on intracranial signals and require invasive neurosurgery, which is costly and limits their application. In contrast, electroencephalography (EEG) signals are non-invasive and easy to record, so EEG-to-Text tasks have gained much attention.

Early EEG-to-Text research was limited to decoding small closed vocabularies (Makin et al., 2019; Nieto et al., 2022), which were insufficient for natural communication. Therefore, Wang and Ji (2021) proposed an open vocabulary EEG-to-Text task and treated it as a sequence-to-sequence translation task. They built an encoder-decoder structured EEG-to-Text model that translates word-level EEG feature sequences into corresponding sentences (as shown in Figure 1). Subsequently, significant efforts (Zhou et al., 2023; Feng et al., 2023; Duan et al., 2023) have followed this paradigm, aiming to improve the decoding accuracy of EEG-to-text translation models.

Although these methods can enhance performance metrics, as recently discovered by Jo et al. (2024), their experimental settings have a crucial

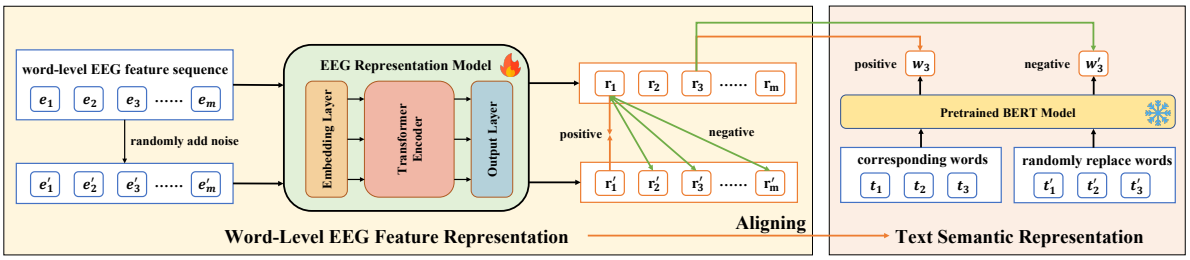

Figure 2: The overall training process of our proposed EEG representation model. We obtain word-level EEG and text semantic representations through the EEG representation model and the pre-trained BERT model, respectively. During training, we freeze the BERT parameters and optimize the EEG representation model through dual contrastive learning to align the EEG representation to the text semantic space.

limitation. Specifically, they still employed the teacher-forcing strategy during evaluation, which led to exaggerated performance. Furthermore, the experimental results in Jo et al.'s paper show that the EEG-to-Text model (Wang and Ji, 2021) performs similarly on pure noise compared to actual EEG data, which indicates that the model does not really utilize EEG signals. We believe this is due to the teacher-forcing strategy employed during training. EEG signals contain much noise, so compared to the context information provided by teacher-forcing, the help supplied by EEG is negligible. Therefore, the model gradually relies only on contextual information in the decoder to predict subsequent words while disregarding the EEG information provided by the encoder.

To address the drawbacks of previous methods, this paper proposes a novel approach by transforming the EEG-to-Text task into a fine-grained controllable text generation task based on large language models (LLMs). LLMs have excellent text generation capabilities and can generate fluent and diverse texts (OpenAI, 2023; Touvron et al., 2023). Therefore, we can think of LLM as a brain and use EEG signals to guide it in generating corresponding text. Specifically, we designed an EEG signal representation model to extract valuable information from word-level EEG signals and generate EEG representations. Then, we obtain the semantic representation for each word using the pre-trained BERT model (Devlin et al., 2018) and propose a dual contrastive learning method to align the EEG representation with the semantic representation of the corresponding word. Finally, during inference, we allow LLMs to generate multiple candidate words at each decoding step and select the word whose semantics are closest to the current EEG representation as the final result.

To the best of our knowledge, this is the first paper that treats the EEG-to-Text task as a controllable text generation task and proposes a comprehensive solution. By aligning semantics in EEG signals and language, our method allows the EEG representation model to thoroughly learn how to effectively extract and utilize the information contained in EEG signals, avoiding the training collapse caused by teacher-forcing. Furthermore, during the decoding stage, our method uses EEG signals to stimulate the "brain" to continuously generate corresponding text, making full use of the powerful capabilities of LLMs.

Extensive experiments on the ZuCo (Hollenstein et al., 2018) benchmark demonstrate the effectiveness of our method. In multi-subject and single-subject scenarios, our method performs significantly better than the previous method, indicating that our method can effectively align EEG signals with text semantics and guide the LLMs to generate correct text. More encouragingly, our method achieves superior performance in the cross-subject setting, demonstrating excellent generalization and applicability to unseen subjects. In summary:

- This paper addresses the shortcomings of previous methods and first regards the EEG-to-Text task as a fine-grained controllable generation task, enabling the model to better capture and utilize the information in EEG.
- We propose an innovative framework to accomplish the EEG-to-Text task. We treat the LLM as a brain and guide it in generating correct and fluent text by leveraging EEG signals aligned with the text semantics.
- We evaluate our approach on the ZuCo benchmark. Our method achieves state-of-the-art performance in various settings and demonstrates excellent generalizability, which verifies the strong potential of our approach.

2 Method

In this section, we provide a comprehensive introduction to our method. First, we highlight the shortcomings of the previous EEG-to-Text task definition and propose a new solution in Section 2.1. Then, in Section 2.2, we describe the training process of our EEG representation model. Finally, in Section 2.3, we present the decoding algorithm that uses word-level EEG signals to guide LLMs in generating desired sentences.

2.1 Task Definition

As shown in Figure 1, the EEG signals are collected during natural reading, and eye-tracking fixations are used to divide word-level EEG segments for the EEG-to-Text task. Similar to previous works (Wang and Ji, 2021; Feng et al., 2023), we use the statistical results of each word-level EEG segment in eight frequency band filters, theta1 (4–6Hz), theta2 (6.5–8 Hz), alpha1 (8.5–10 Hz), alpha2 (10.5–13 Hz), beta1 (13.5–18Hz), beta2 (18.5–30Hz), gamma1 (30.5–40 Hz) and gamma2 (40–49.5 Hz), as its feature vector. Although different word-level segments may have different EEG window sizes, the final feature vector dimensions remain the same. Therefore, the input of the EEG-to-Text model is a word-level EEG feature vector sequence $E = [e_1, e_2, ..., e_m]$, which may contain m feature vectors. Our goal is to decode its corresponding text sequence $T = [t_1, t_2, ..., t_m]$.

Previous studies (Wang and Ji, 2021; Feng et al., 2023; Duan et al., 2023; Zhou et al., 2023; Xi et al., 2023) have defined this task as a sequence-to-sequence translation task, where the EEG sequence E is the source sentence and the text sequence T is the target sentence, and maximize the probability of the decoded sentence:

$$p(T \mid E) = \prod_{i=1}^{m} p\left(t_i \mid E, t_{<i}\right) \quad (1)$$

Similar to traditional machine translation models, these methods employ the teacher-forcing strategy during training, inputting the ground truth sequence into the decoder. As a result, the model predicts the current token based on all previous ground truth tokens $t_{<i}$ and the EEG sequence E. However, due to the low signal-to-noise ratio of EEG signals, the encoder of the EEG-to-text translation model cannot effectively extract valuable information. Consequently, the contribution of the EEG sequence to prediction is much smaller than that of the previous ground truth tokens. Therefore, previous models are prone to degenerate into language models during training, focusing solely on language modeling in the decoder and failing to complete the EEG-to-text task.

Different from previous approaches, in this paper, we define the EEG-to-Text task as a fine-grained controllable text generation task. Since LLMs have powerful text generation capabilities, we can regard them as human brains. Therefore, the key to completing the EEG-to-Text task is controlling LLMs to generate correct text T through EEG sequences E. Since fine-tuning LLMs is expensive, we adopt an efficient post-processing paradigm (Zhang et al., 2022) to guide LLMs in generating specific content. Specifically, we align each word-level EEG representation with the corresponding text semantics. Then, at each decoding step, the LLM generates multiple candidate words, and we select the word closest to the current EEG representation as the final output. Consequently, instead of using the teacher-forcing strategy to train a translation model, we focus on training a robust EEG representation model that can effectively extract and utilize information from EEG signals.

2.2 Model Training

Model Architecture We constructed an effective EEG representation model to extract valuable semantic information from word-level EEG features and output the corresponding word-level EEG representation vector sequence $R = [r_1, r_2, ..., r_m]$. Specifically, our EEG representation model consists of three main modules: embedding layer, Transformer encoder (Vaswani et al., 2017), and output layer. In particular, the embedding layer consists of an MLP layer and a learnable positional encoding, which transforms the dimension of the input word-level EEG features and incorporates position information. The Transformer encoder is used to capture contextual dependencies between word-level EEG features within the same sentence through the attention mechanism. The ablation study (Appendix C.1) shows that the representation, in conjunction with contextual information, is significantly better than independent representations. The output layer consists of an MLP layer, which controls the dimension of the final representation vector to match the text semantic vector. Formally, the overall model is formulated as follows:

$$R = MLP_2(Tansformer(MLP_1(E)+P) \quad (2)$$

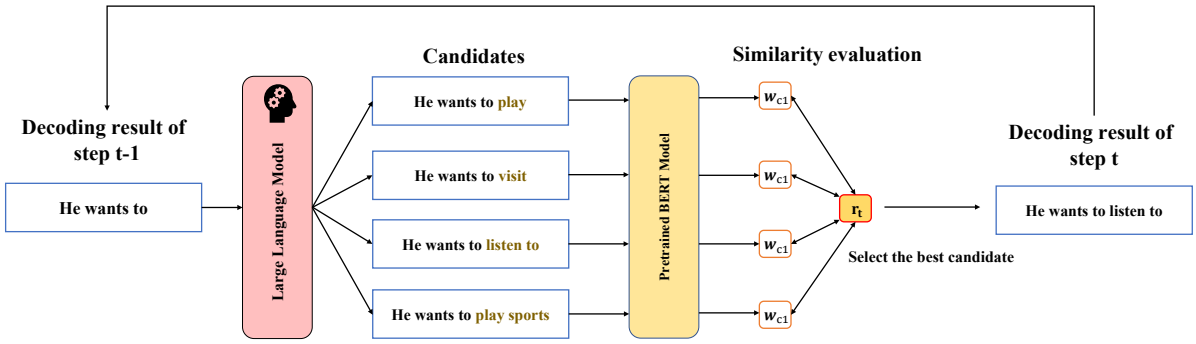

Figure 3: Implementation details of the decoding process. When generating the result of step t, we first let the LLM predict candidate words based on the previous result. Then, we calculate the cosine similarity between the candidate's semantics and the t-th word-level EEG representation r_t to select the best candidate as the final result.

where $P \in \mathbb{R}^{m \times d_{\text{model}}}$ consists of learnable position vectors.

Training Details Figure 2 shows the overall training process. To enable EEG signals to guide LLMs in generating correct text, we need to align each word-level representation with the semantics of its corresponding word. We input the EEG sequence E into our representation model to obtain the word-level EEG representation sequence. We use the pre-trained BERT (Devlin et al., 2018) model to generate the semantic representation for each word. Moreover, since the same word can have different meanings in different contexts, we input the word and its previous context into the BERT model to obtain a more accurate semantic representation. The output from the last input token of the BERT model is used as the representation of the current word. Meanwhile, we obtain representations of unrelated words by randomly replacing the input to the BERT model. Specifically, we replace the current word and randomly replace the previous context with 50% probability. Note that text semantic representations can be reused to improve training efficiency.

After obtaining the EEG representation sequence $R = [r_1, r_2, ..., r_m]$, the corresponding text semantic representation sequence $W = [w_1, w_2, ..., w_m]$, and the unrelated representations $W^{'} = [w_1^{'}, w_2^{'}, ..., w_m^{'}]$, we apply contrastive learning (Hadsell et al., 2006; Schroff et al., 2015) to align the EEG to the text semantic space:

$$loss_1 = \alpha * d(R, W) - \beta * d(R, W^{'}) \quad (3)$$

where α and β are hyperparameters used to balance the impact of positive and negative pairs. We use cosine similarity to measure the distance between representation vectors. We treat W as a positive sample and bring the EEG representation closer to it, while considering $W^{'}$ as a negative sample and increasing the distance between the EEG representation and it. In this way, our EEG representation model can better distinguish subtle semantic differences between different words, thereby generating more accurate representations.

Additionally, we find that focusing solely on aligning EEG with text semantics makes all word-level EEG representations very similar, which leads to poor decoding quality. To address this, we further employ self-supervised InfoNCE loss (van den Oord et al., 2018) to improve the discrimination between each EEG representation vector and enhance the robustness of our model. Specifically, we randomly add noise values sampled from the $\mathbf{N}(\mathbf{0}, \mathbf{1})$ distribution into the EEG feature sequence E to obtain $E^{'}$, and then input it into our EEG representation model to obtain $R^{'} = [r_1^{'}, r_2^{'}, ..., r_m^{'}]$. We regard r_i and r'_i as positive sample pairs while treating the other perturbed vectors in the same batch as negative samples:

$$loss_2 = -\log \frac{\exp\left(\text{sim}\left(r_i, r_i^{'}\right)/\tau\right)}{\sum_{k=1}^{N} \exp\left(\text{sim}\left(r_i, r_k^{'}\right)/\tau\right)} \quad (4)$$

where τ is the temperature, which controls the contrast strength. We sum these two contrastive learning losses to obtain the final training loss:

$$loss = \alpha * d(R, W) - \beta * d(R, W^{'}) + loss_2 \quad (5)$$

Through our proposed dual contrastive learning training method, we can train a superior EEG representation model that effectively aligns word-level EEG features with the corresponding text semantics and distinguishes EEG signals corresponding to different words (as shown in Figure 7).

Model	BLEU-N (%)			ROUGE-1 (%)		
	N=1	N=2	N=3	R	P	F
EEG-to-Text	17.4	9.2	5.3	17.5	17.7	17.4
Random LLM	14.2	8.8	6.1	15.1	18.3	16.3
Ours	**21.1**	**11.7**	**7.6**	**19.8**	**19.6**	**19.7**
Oracle LLM	41.2	31.7	25.6	43.2	45.9	44.3

Table 1: Performance of our method and baseline models in the multi-subject setting. To evaluate model performance accurately, we do not employ teacher-forcing strategies during evaluation.

2.3 Decoding Process

In the decoding stage, we use EEG signals to guide the LLM in generating the corresponding text. First, we input the word-level EEG feature sequence E into the trained EEG representation model to obtain the corresponding representation sequence R. Then, we use R to re-rank the results generated at each step by the LLM, controlling them to produce the desired output. Precisely, as shown in Figure 3, at the t-th decoding step, LLM predicts a set of word candidates based on the previously decoded results. Similar to the training stage, we concatenate the previous context with the candidate words and input them into the pre-trained BERT model to obtain the semantic representation vector for each candidate. Finally, we re-rank the candidates according to the cosine similarity between the candidate representation vector and the t-th word-level EEG representation vector r_t, selecting the candidate with the highest cosine similarity as the result of the t-th decoding step. Therefore, after m steps, our method can decode the entire sentence corresponding to E.

In addition, the word-level EEG segments, divided by eye-tracking fixations, may correspond to multiple words. To address this issue, we allow the LLM to generate candidate sets based on different segment lengths using beam search. We then merge these candidates of various lengths to form a final candidate set, which has a higher probability of containing candidates that align well with the corresponding EEG segments. In summary, our proposed decoding algorithm leverages LLM's powerful text generation capabilities, treating them as brains and using EEG signals to guide them step by step to generate the desired text, providing a novel and feasible solution to the EEG-to-text task.

3 Experiments

3.1 Experimental Setup

We conduct experiments on ZuCo (Hollenstein et al., 2018), a large-scale public benchmark that records eye-tracking and EEG data during natural reading tasks. ZuCo consists of 12 subjects reading the same material, collected from movie reviews (Socher et al., 2013) and Wikipedia articles, segmented into 1,107 sentences. During the reading process, EEG signals were collected using a 128-channel EEG Geodesic Hydrocel system with a sampling rate of 500 Hz and a frequency band filter ranging from 0.1 Hz to 100 Hz. Nine EOG channels were used for artifact removal, and the channels lying mainly on the neck and face were discarded before data analysis. Finally, 105 EEG channels were used for scalp recordings.

Similar to the previous methods (Wang and Ji, 2021; Feng et al., 2023), we concatenate the statistical results of each word-level EEG segment across eight frequency band filters into a feature vector with a dimension of 840 ($e \in \mathbb{R}^{840}$). For both single-subject and multi-subject settings, we split data into the training (80%), validation (10%), and test (10%) sets by unique sentences, i.e., the sentences in the test set are totally unseen. For the cross-subject setting, we use leave-one-out cross-validation based on different subjects. Specifically, each time, we select the data of one subject as the test set, the data of another subject as the validation set, and the data of the remaining 10 subjects as the training set. Additionally, we filter out samples where the number of word-level EEG segments divided by eye-tracking fixations is less than half the total number of words.[1] We adopt BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) as evaluation metrics to measure the quality of the decoded text. The implementation details and baselines can be found in Appendix A and Appendix B.

3.2 Main Results

Multi-Subject Setting Similar to previous approaches (Wang and Ji, 2021; Duan et al., 2023), we first conduct experiments in the multi-subject setting, where the data from all subjects are mixed. As shown in Table 1, our method achieves excellent performance in multi-subject scenarios compared to the previous method. Specifically, without us-

[1]This indicates that the subjects have skipping issues while reading, which makes it challenging to align the collected EEG signals with the text accurately.

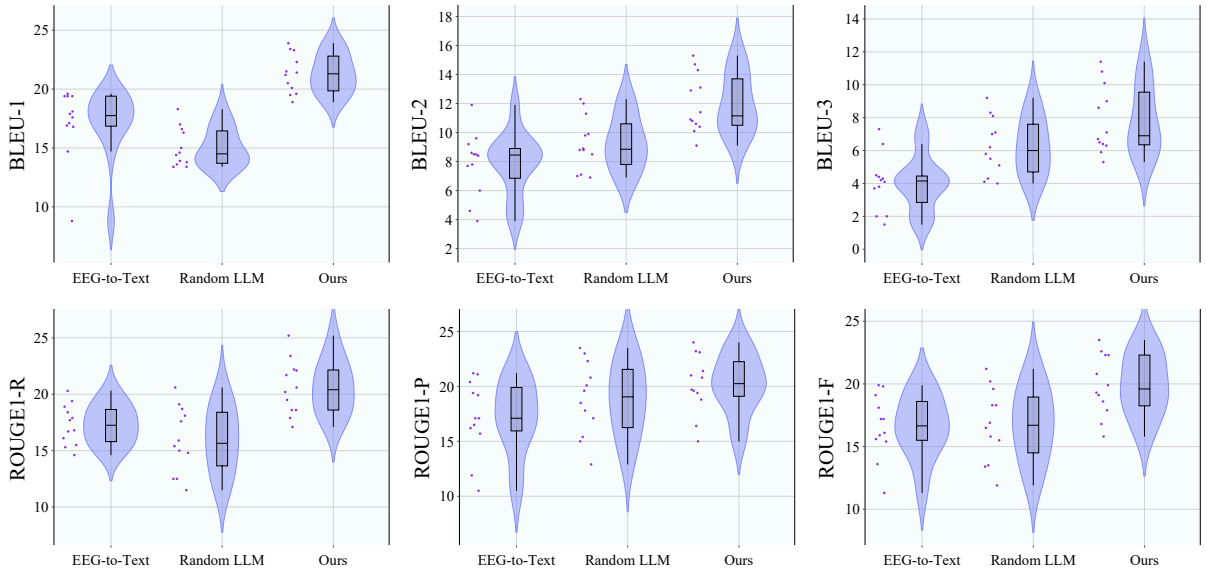

Figure 4: Performance of our method and baselines in the single-subject setting. Each scatter point represents a subject, and we use violin plots to show the distribution of metrics for different methods across the 12 subjects.

ing teacher-forcing, the performance of the EEG-to-Text translation model is similar to that of the Random LLM across most metrics. This suggests that the EEG-to-Text model ignores the EEG information on the encoder side during training and degenerates into a language model, highlighting the shortcomings of previous methods. In contrast, our method achieves a significant performance improvement compared to the Random LLM, increasing the BLEU-{1, 2, 3} scores by **6.9 (+48.6%)**, **2.9 (+33.0%)**, and **1.5 (+24.6%)**, respectively, and the ROUGE1-{R, P, F} scores by **4.7 (+31.1%)**, **1.3 (+7.1%)**, and **3.4 (+20.9%)**.

These experimental results demonstrate that our proposed method effectively learns how to extract semantic information from EEG signals, thereby guiding the LLM to generate accurate results. We attribute this success to our approach's ability to eliminate the distraction of the teacher-forcing strategy during training, allowing the model to focus on EEG representation and alignment. However, since this is the first paper to complete the EEG-to-text task in the form of fine-grained controlled text generation, there remains a significant gap between the performance of our method and that of the Oracle LLM. This indicates that our proposed paradigm holds great potential, and further efforts are required to develop more powerful EEG representation models in the future.

Single-Subject Setting We also validated the performance of our method in the single-subject scenario, where we trained and evaluated the model for each subject independently. The experimental results for 12 subjects are shown in Figure 4. Since the EEG-to-text model essentially performs language modeling without exploiting the information provided by the EEG signal, its performance is similar to that of the Random LLM. In contrast, our method achieves statistically better performance (paired t-test with $P < 0.01$) than the Random LLM in the single-subject setting. Since there are differences between subjects, the model's performance fluctuations are observed across subjects. However, our method outperforms both the EEG-to-Text model and the Random LLM on each subject. These experimental results demonstrate the robustness of our approach, which applies to different subjects and can be effectively trained with only a small amount of data.

Cross-Subject Setting In real-world applications, the model is likely to encounter data from unknown subjects, so its ability to generalize to unseen subjects is crucial. To assess the cross-subject generalization ability of our method, we perform leave-one-out cross-validation based on 12 subjects. The experimental results are presented in Figure 5. As we can see, our method still outperforms the EEG-to-Text model on unseen subjects. Specifically, the performance of the EEG-to-Text model is similar to that of the above two settings because it does not utilize EEG information. In contrast, our method achieves an average of **20.9** BLEU-1 scores

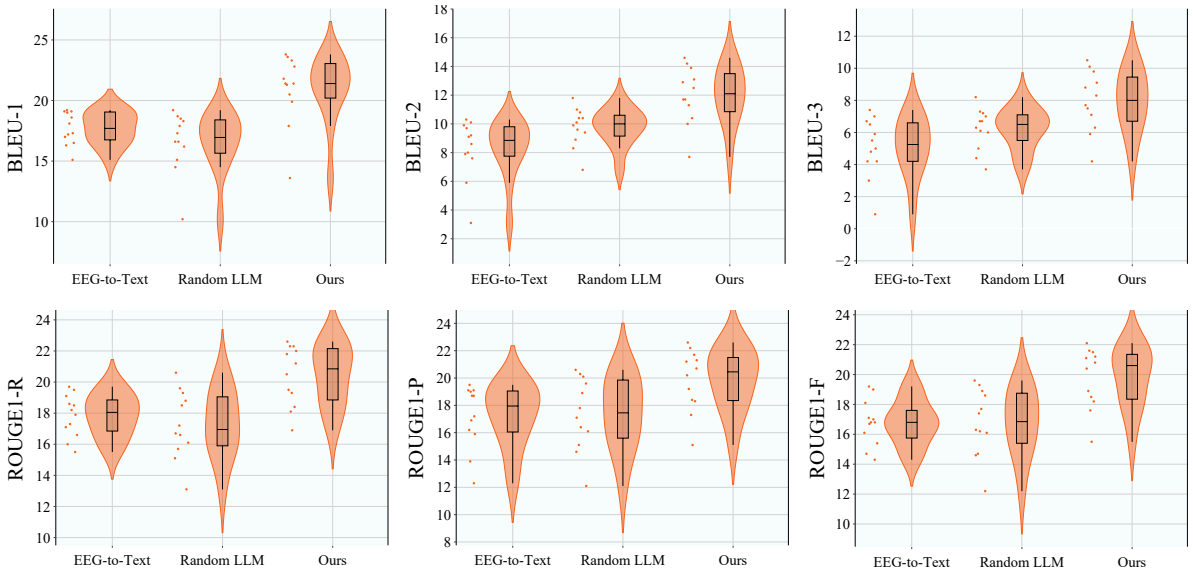

Figure 5: Performance of our method and baselines in the cross-subject setting. To effectively evaluate the generalization ability of our method, we performed leave-one-out cross-validation on all 12 subjects.

Hyperparameters	$\beta = 1$	$\beta = 2$	$\beta = 3$	$\beta = 4$
$\alpha = 1$	18.9 / 17.2	15.8 / 18.4	13.0 / 16.0	8.6 / 16.5
$\alpha = 2$	20.5 / 17.8	19.8 / 18.3	16.0 / 16.9	12.8 / 16.3
$\alpha = 3$	21.0 / 19.0	20.8 / 18.7	19.1 / 18.7	14.9 / 17.0
$\alpha = 4$	21.8 / 18.9	**21.9 / 19.3**	20.5 / 18.4	16.9 / 18.4

Table 2: Evaluation metrics (BLEU-1 / ROUGE1-F) of our method on the validation set with different training hyperparameters. Since α and β are strongly correlated, we use the grid search method to find the best setting.

and **19.8** ROUGE1-F scores, which surpasses the EEG-to-Text model by margins of **3.2 (+18.1%)** and **3.1 (+18.6%)**. These results show that our model can capture the common knowledge between data from different subjects and generalize its representation ability to unseen subjects. We believe this is because we introduced EEG self-supervised contrastive learning (van den Oord et al., 2018) during the training phase, enabling our model to learn a distinguishable and robust representation space, thereby enhancing its generalization ability.

3.3 Analysis

3.3.1 Impact of Hyperparameters

We conduct experiments on the validation set in the multi-subject setting to explore the impact of the hyperparameters on our approach.

Impact of Training Hyperparameters As shown in Equation (5), the primary hyperparameters in the training process include α and β. The experimental results in Table 2 show that different hyperparameter settings can impact model performance, and the trade-off between α and β must be carefully considered. Specifically, when both α and β are small, the self-supervised InfoNCE loss dominates the training process. In this case, although the EEG representation model can effectively distinguish different word-level EEG features, the generated representation vector is not aligned with the text semantic space. As a result, during the decoding stage, EEG representations cannot provide accurate guidance for LLMs, leading to poor model performance. On the other hand, if α and β are too large, the model fails to distinguish the EEG signals corresponding to different words. Furthermore, when β is greater than α, the EEG representation model tends to generate representations that are far from the negative samples, regardless of whether the generated representations are close to the text semantics of the ground truth. This also leads to a decrease in decoding quality. Based on these experimental results, we set α to 4 and β to 2 to ensure that the EEG representation is effectively aligned with the text semantics while maintaining good discrimination.

Impact of Inference Hyperparameters During inference, we need to choose the appropriate candidate length and beam size. First, we fix the beam size to 300 and analyze the effects of various candidate lengths on decoding quality. As shown in Figure 6, different candidate lengths significantly impact final performance. In particular, generating

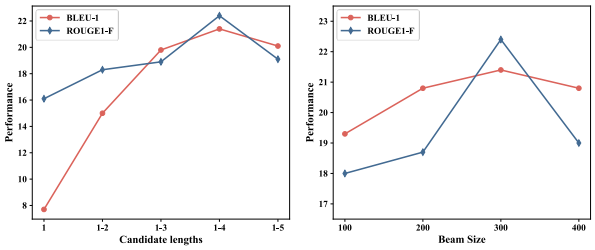

Figure 6: The impact of different candidate lengths (left) and beam sizes (right) on the performance of our method. **1-3** means that we let LLMs generate candidates of length 1, 2, and 3 at each decoding step and merge them to form the final candidate set.

only a candidate set of length 1 at each decoding step results in inferior performance, indicating that most word-level EEG features correspond to more than one token. Therefore, our method for generating candidate sets based on multiple lengths is necessary and can effectively improve alignment accuracy during decoding.

Next, we fix the candidate length to 1-4 and analyze the impact of different beam sizes. As shown in Figure 6, when the beam size is small, the ground truth may be excluded from the candidate set, resulting in poor decoding quality. On the other hand, a large beam size leads to a larger candidate set, which increases the difficulty of guiding the generation of LLMs using EEG representations. Additionally, memory usage and inference latency increase as the beam size grows. Therefore, we finally set the candidate length to 1-4 and the beam size to 300. Due to space constraints, more analysis is presented in Appendix C.

4 Related Works

Decoding brain signals into text is an active research direction at the intersection of artificial intelligence and neuroscience, with high research value and a wide range of application scenarios. There are many types of brain signals, such as electrocorticography (ECoG) and stereotactic electroencephalography (sEEG) for intracranial recordings and functional magnetic resonance imaging (fMRI) and EEG for non-invasive recordings. Intracranial recordings offer high temporal and spatial resolution, as well as a high signal-to-noise ratio, enabling the decoding of high-quality text from them (Herff et al., 2015; Willett et al., 2021). However, their acquisition requires expensive invasive neurosurgery, which cannot be effectively promoted.

Therefore, researchers began to explore how to use non-invasive recordings for Brain-to-Text decoding. For example, Tang et al. (Tang et al., 2022) tried to reconstruct continuous language using fMRI. Similar to our method, they introduced a language model to generate candidates and select the best based on fMRI. However, due to the low temporal resolution of fMRI, each brain image can be affected by over 20 words, so it is difficult to give accurate guidance during decoding. In contrast, we employ high temporal resolution EEG signals, which can provide fine-grained guidance for LLMs at each decoding step. Additionally, EEG acquisition is more portable and cost-effective than fMRI, enhancing its practical value.

Recently, Wang and Ji (2021) proposed an open vocabulary EEG-to-Text task and regarded it as a translation task to solve it. A series of subsequent works have followed this idea and improved upon it (Zhou et al., 2023; Xi et al., 2023; Duan et al., 2023; Feng et al., 2023; Liu et al., 2024). For example, DeWave (Duan et al., 2023) uses a quantized variational encoder to derive discrete codex encoding, which reduces the interference caused by individual differences in EEG signals.

However, Jo et al. (2024) found that this series of works employed teacher-forcing during the evaluation, which caused their performance to be inflated. Moreover, the experimental results show that the EEG-to-Text translation model performs similarly on pure noise compared to actual EEG data, meaning it does not utilize the EEG information. Unlike previous methods, we only focused on aligning the word-level EEG representations with the corresponding text semantics without using the teacher-forcing strategy. Therefore, we could effectively build an EEG representation model to extract semantic information and guide LLMs in generating correct sentences.

5 Conclusion

This paper proposes a novel framework for the EEG-to-Text task. Our approach regards the LLM as a brain and guides it in generating correct text by leveraging aligned EEG signals. Experimental results on the ZuCo benchmark demonstrate the effectiveness of our method, which not only achieves significantly better performance than previous methods in both multi-subject and single-subject settings but also shows strong generalization capabilities in cross-subject scenarios.

6 Limitations

Although the new paradigm for the EEG-to-Text task proposed in this paper addresses the shortcomings of previous methods and achieves significant performance improvements, some issues still require further exploration.

First, our method treats the LLM as a brain and uses EEG signals to guide it in generating the desired text. It is curious whether this decoding process resembles brain activity during human thinking and whether the LLM shares similarities with the human brain. We believe that with the continuous development of neuroscience, these questions can be answered more clearly, which will further guide us in designing better methods.

Second, there is a significant performance gap between our method and the Oracle LLM, indicating that our EEG representation model can be further improved. We believe better handling the low signal-to-noise ratio in EEG signals is key to improving model performance.

Third, our method relies on eye-tracking fixations to divide EEG signals into word-level segments, which limits its practical application. Furthermore, the word-level EEG segments divided by eye-tracking fixations are not always strictly aligned with the text, presenting challenges for training EEG representation models and generating candidate words during decoding. Therefore, research on text decoding methods based on raw EEG signals is crucial.

References

Gopala Krishna Anumanchipalli, Josh Chartier, and Edward F. Chang. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568:493 – 498.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu kai Wang, and Ching-Teng Lin. 2023. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *ArXiv*, abs/2309.14030.

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*, 31:3874–3883.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742.

Christian Herff, Dominic Heger, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9.

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.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. 2024. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *ArXiv*, abs/2405.02165.

Joseph G. Makin, David A. Moses, and Edward F. Chang. 2019. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23:575 – 582.

Nicolás Nieto, Victoria Peterson, Hugo Leonardo Rufiner, Juan E. Kamienkowski, and Rubén D. Spies. 2022. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. *Scientific Data*, 9.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Elmar G. M. Pels, Erik J. Aarnoutse, Nick F. Ramsey, and Mariska J. Vansteensel. 2017. Estimated prevalence of the target population for brain-computer interface neurotechnology in the netherlands. *Neurorehabilitation and Neural Repair*, 31:677 – 685.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*.

Carol A. Stanger and Michael F. Cawley. 1996. Demographics of rehabilitation robotics users. *Technology and Disability*, 5:125–137.

Pengfei Sun, Gopala Krishna Anumanchipalli, and Edward F. Chang. 2019. Brain2char: a deep architecture for decoding text from brain recordings. *Journal of Neural Engineering*, 17.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. 2022. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Zhenhailong Wang and Heng Ji. 2021. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *AAAI Conference on Artificial Intelligence*.

Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. 2021. High-performance brain-to-text communication via handwriting. *Nature*, 593:249 – 254.

Nuwa Xi, Sendong Zhao, Hao Wang, Chi Liu, Bing Qin, and Ting Liu. 2023. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Annual Meeting of the Association for Computational Linguistics*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56:1 – 37.

Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu kai Wang, and Chin-Teng Lin. 2023. Belt: Bootstrapping electroencephalography-to-language decoding and zero-shot sentiment classification by natural language supervision. *ArXiv*, abs/2309.12056.

A Implementation Details

We conducted experiments on NVIDIA RTX A6000 48GB GPUs. The hidden dimension of the two MLP layers in our EEG representation model is 1536, and the output dimension is 768. Our Transformer encoder consists of 3 layers, each with 8 attention heads and a hidden dimension of 2048. The pre-trained BERT-base model (Devlin et al., 2018) is used to generate text semantic representations. During training, we set the hyperparameters $\alpha = 4$, $\beta = 2$, and $\tau = 0.3$. We use the Adam optimizer with a learning rate of 5×10^{-4}, a batch size of 128, and a dropout rate of 0.3. We train for 30 epochs and select the checkpoints with the best performance on the validation set as the final model. During the decoding process, we adopt Llama2-7b (Touvron et al., 2023) for inference, using the ground truth corresponding to the first word-level EEG feature as the prompt and the subsequent EEG feature sequence to guide it in decoding text. We set the candidate lengths to [1, 2, 3, 4] and the beam size to 300 for Llama2-7B, generating a total of 1200 candidate words and selecting the best result from them.

B Baselines

To verify the effectiveness of our method, we compare our method primarily with three baselines:

1. **EEG-to-Text Model** (Wang and Ji, 2021): This model treats the EEG-to-Text task as a translation task and extends the pre-trained BART (Lewis et al., 2019) model to construct an EEG translation model. For a fair comparison, the same prompts used in our method are input to its decoder during decoding.

2. **Random LLM**: During the decoding process, this baseline randomly selects results from the candidate set, with the rest of the process identical to our method. By comparing with this baseline, we can assess whether our method

effectively extracts and utilizes information from EEG signals to guide LLMs correctly.

3. **Oracle LLM**: This model selects the best candidate using the semantic representation of the ground truth, representing the ideal case where EEG representations are perfectly aligned with text semantics.

C More Analysis

C.1 Ablation Study

To verify the effectiveness of each proposed module, we conduct detailed ablation experiments in the multi-subject setting. Specifically, we compare our approach with several variants:

- w/o InfoNCE loss: Directly aligning EEG representations and text semantics using a contrastive learning method similar to CLIP (Radford et al., 2021).
- w/o text negative samples: Focusing only on narrowing the distance between EEG representations and the ground truth semantics without constructing negative samples by randomly replacing words.
- w/o adding noise: Using only the original word-level EEG features to compute the self-supervised InfoNCE loss.
- w/o previous context: Excluding previous context when obtaining the text semantics.
- w/o Transformer Encoder: Removing the Transformer Encoder module from our EEG representation model so that each word-level EEG feature is encoded independently. To prevent a significant change in the number of model parameters, we replaced the Transformer Encoder with a three-layer MLP of similar size.

As shown in Table 3, removing any module leads to performance degradation of our method. First, the w/o InfoNCE loss variant performs poorly, indicating that directly aligning EEG representations with text semantic representations is not feasible. Therefore, we propose a dual contrastive learning method, which effectively trains the EEG representation model through modal alignment of EEG signals and text semantics, as well as self-supervised learning with noise, achieving excellent performance. Second, when obtaining text semantics, concatenating the previous context to the word can effectively reduce ambiguity, allowing EEG representations to be better aligned with the corresponding semantics. In addition, it is necessary to construct negative text examples by randomly replacing words. Finally, independently encoding each word-level EEG results in a significant decrease in decoding quality. This shows that the Transformer Encoder in our EEG representation model can effectively capture and utilize contextual dependencies in word-level EEG sequences, thereby improving representation capabilities.

Method	BLEU-1	ROUGE1-F
Ours	**21.1**	**19.7**
w/o InfoNCE loss	15.8	16.4
w/o adding noise to EEG	19.1	17.9
w/o previous context	18.4	17.1
w/o text negative samples	18.5	18.5
w/o Transformer Encoder	18.5	18.1

Table 3: Ablation experiments of our method in the multi-subject setting.

C.2 Representation Visualization

To verify whether our EEG representation model can effectively distinguish different word-level EEG features, we use t-SNE (van der Maaten and Hinton, 2008) to visualize the EEG representations of both the EEG-to-Text translation model and our model. As shown in Figure 7, the representations of the EEG-to-Text model are disorganized and cannot distinguish between different word-level EEG signals. Consequently, the EEG-to-Text model struggles to utilize EEG signals for decoding effectively. In contrast, our model enables word-level EEG representations corresponding to the same word to form clear clusters, even when sourced from different subjects. These experimental results demonstrate that our method allows the EEG representation model to learn to identify semantic information across various EEG signals during training and to acquire robust representation capabilities. Based on these distinct EEG representations, our approach can provide more accurate guidance to LLMs during the decoding stage.

C.3 Case Study

We present a case study in Table 4 to specifically analyze the decoding results of different methods. As we can see, even though the open vocabulary EEG-to-text task is very challenging, our model can still

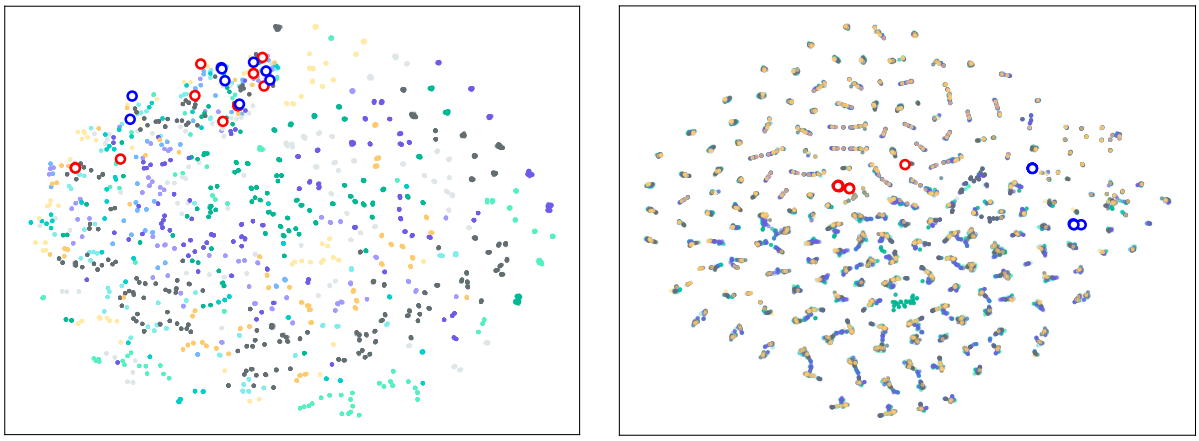

Figure 7: T-SNE visualization of word-level EEG representations of the EEG-to-Text model (left) and our model (right). Different colors mean different subjects. Each dot represents a word-level EEG representation. The red or blue box dots represent EEG representations corresponding to the same word.

(1)	Ground Truth: I did go **back** and check out the **last** 10 minutes, **but** these were more repulsive than the first 30 or 40 minutes.
	EEG-to-Text: I am born in New York City, and graduated from Yale University.
	Random LLM: I made another one but this time j couldnt sit still so added photos of him along the way thanks for looking and any swizzers out there pls tell us how you managed to
	Ours: I also went **back** and watched the **last** part of the movie, **but** it was very bad for me that I don't want to watch.
	Oracle LLM: I did go back and check out the last 10 minutes, but these were rushed in comparison to the the first 30 or more minutes.
(2)	Ground Truth: A **strong** script, powerful direction and splendid production design allows us to be transported into the life of wladyslaw szpilman, **who is** not only **a pianist**, but a good human being.
	EEG-to-Text: A thinly veiled layer of fog settled over the quiet town, casting an eerie glow in the early morning light.
	Random LLM: A network of partners and suppliers that have been tested in the real world. What brought you...Building a successful brand involves a lot of effort other than...Building Smarter CitiesHardware is just the beginningWhat excites you most about P
	Ours: A **strong** influence on his career. He once told us that bruce was the nephew of William Bruce, Jr., **who is a pianist**, but also a person with a good temper.
	Oracle LLM: A strong foundation, powerful direction, excellent production presentation makes it easy to be transported away into the life of Jacob Jensen, who is not only a pianist, but is a good human being, too.

Table 4: Examples of EEG-to-text decoding results using different methods in the multi-subject setting. **Bold words** indicate exact matches. Underlined words denote fuzzy matches with similar semantics.

decode some meaningful results corresponding to the EEG signals, achieving better performance than previous methods.

First, although the EEG-to-Text model can generate fluent sentences, the content is unrelated to the ground truth. This outcome indicates that the EEG-to-Text model collapsed into a language model during training and failed to learn how to extract and utilize information from the EEG. In contrast, our method successfully decodes some words that exactly match the ground truth or phrases with similar semantics. This result demonstrates that our EEG representation model can align word-level EEG features with the corresponding text semantics, effectively guiding LLMs to generate accurate text. Second, our approach is better at matching simple nouns or adjectives but has difficulty generating the correct named entities, such as 'wladyslaw szpilman' in example (2). We believe this is because the model cannot distinguish semantic differences between various entities of the same type. In addition, the correct entity may be rare and not necessarily included in the candidate set predicted by LLMs. Third, there is a decoding quality gap between our method and Oracle LLM. Oracle LLM can generate results that are mostly the same as the ground truth. These results show that it is reasonable to consider the EEG-to-Text task as a fine-grained controllable text generation task. There remains significant room for improvement in our EEG representation model, and exploring ways to enhance its performance warrants further investigation.