Decoding Thoughts, Refining Words: EEG-to-Text Meets LLMs

Cagan Bakirci Neval Cam Asmaa Hassan Arya Miryala Grace Wu Yaqi Zhang
University of Southern California
{cbakirci, ncam, hassanas, miryala, gkwu, yaqiz763} @usc.edu

Abstract

Electroencephalogram (EEG) to text translation is a promising non-invasive approach for brain-computer interfaces, particularly for individuals with severe speech impairments. However, existing state-of-the-art models face critical challenges: performance metrics are inflated by teacher forcing during evaluation and real-world decoding produces poor quality text with high error rates. In this project, we replicated and extended the R1 Translator model, which combines a bidirectional LSTM EEG encoder with a pretrained BART decoder, to better understand these limitations. Using the ZuCo V1 dataset, consisting of 1,107 sentences from 12 participants across three reading tasks, we performed hyperparameter tuning and systematically evaluated decoding and post-processing strategies. Our best configuration improved BLEU-1 in non–teacher-forced decoding from approximately %12.16 to %18.81, showing that careful decoding choices can give some gains. However, techniques such as external language model rescoring and context-aware decoding did not substantially reduce the gap between teacher-forced and non-teacher-forced decoding. Our goal was to contribute to the ongoing research challenge of developing EEG-to-text systems that can function effectively without teacher forcing.

1 Introduction

Decoding language from brain signals such as Electroencephalogram (EEG) has important real-world applications. EEG to text translation can offer a non-invasive alternative to more invasive brain-computer interfaces (BCIs). Any advancements to these systems can especially be useful for individuals with severe speech impairments such as Amyotrophic Lateral Sclerosis (ALS) and Locked-In Syndrome (LIS).

Although EEG is advantageous because it is non-invasive and low-cost, it comes with many significant challenges. EEG has lower spatial resolution than other signals such as electromyography (EMG). EEG signals are high-dimensional, noisy, and vary across individuals, making them difficult to interpret and map to natural language.

Existing EEG-to-text decoding studies have many issues: some are restricted to closed-vocabulary, limiting real-world usability, and evaluation relies on teacher forcing, which inflates reported results and fails to reflect the same accuracy when the ground truth is not present Jo et al. (2024).

Overcoming these obstacles to achieve robust, open-vocabulary EEG-to-text translation would advance assistive technologies and improve our understanding of how neural activity maps to natural language. To summarize, our goal is to explore modern ML and NLP methods, including various neural architectures and signal processing techniques, to improve existing baselines and develop models that can more accurately decode brain activity into text.

2 Related Work

Early EEG-to-text systems were mostly closed-vocabulary and relied on classical sequence models to decode very limited lexicons. For example, Porbadnigk et al. (2009) applied hidden Markov models to classify speech from EEG over a fixed five-word vocabulary. Such approaches showed feasibility, but were restricted to small, predefined vocabularies, limiting real-world applications.

In later years, researchers explored open-vocabulary decoding and needed standardized datasets. The ZuCo dataset emerged as the standard benchmark (Hollenstein et al., 2018, 2020). It provided a rich corpus of EEG and eye-tracking data collected during both natural

reading and a task-specific annotation. With over 700 sentences from 18 participants, ZuCo enabled systematic evaluation of EEG-to-language models and created the foundation for comparing methods across studies.

Building on ZuCo, recent research has introduced a series of open-vocabulary EEG-to-text models using neural seq2seq architectures. Wang and Ji (2022) proposed such a framework by combining EEG encoders with pretrained BART decoders, while Liu et al. (2024) and Duan et al. (2023) explored discrete EEG wave encodings. Most recently, Murad et al. (2025) presented the R1 Translator, which integrates a bidirectional LSTM encoder with BART, a pretrained transformer decoder, achieving state-of-the-art performance on ZuCo and serving as our baseline.

Despite these advancements, the validity of reported results has been questioned. Jo et al. (2024), in Are EEG-to-Text Models Working?, demonstrated that much of the observed performance is inflated by teacher forcing and may not truly depend on EEG features. This raises a need for rigorous baselines and other evaluation methods. Therefore, in our work, we aim to evaluate both with and without teacher forcing, ensuring fair comparison with the latest baselines.

Finally, while not yet applied extensively to EEG, language model rescoring has proven effective in other signal-to-text tasks. For example, in speech recognition and EMG decoding, rescoring has improved fluency and robustness of the generated text (Shi et al., 2024). Therefore, incorporating similar strategies into EEG decoding pipelines is a promising direction we can take.

3 Dataset

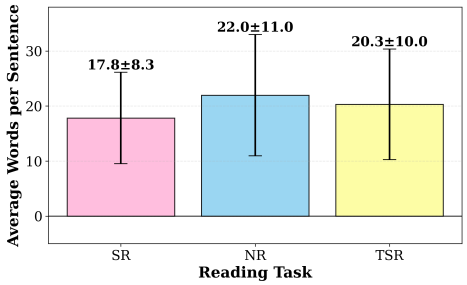
We are utilizing the ZuCo (Zurich Cognitive Language Processing Corpus) V1 dataset, a high-density EEG and eye-tracking resource collected from 12 adult native English speakers reading natural English text over 4-6 hours per participant. The recordings include 21,629 words across 1,107 sentences with 154,173 eye-tracking fixations. Each participant performed 3 reading tasks: 1 sentiment reading (SR) task, 1 normal reading (NR) task, and 1 task-specific reading (TSR) task. This allows exploration of differences in cognitive processing when participants read naturally versus focusing on specific semantic information.

- **Task 1 – Sentiment Reading (SR):** Subjects read movie review sentences labeled by sentiment (positive/negative).
- **Task 2 – Normal Reading (NR):** Subjects read neutral sentences with labeled entity and relation information for information-extraction tasks.
- **Task 3 – Task-Specific Reading (TSR):** Similar to NR but with an explicit relation-annotation objective, causing subjects to focus on semantic relations.

Table 1: ZuCo V1 Dataset Overview

Task	Sentences	Words	Subjects
SR	400	7.1K	12
NR	300	6.6K	12
TSR	407	8.0K	12
Total	**1,107**	**21.6K**	**12**

Figure 1: Sentence length across three ZuCo V1 tasks



We also intended to use ZuCo V2, which records NR and TSR in the same session, addressing session-related EEG variability that existed in V1. However, all download attempts of V2 data set on various devices failed due to 403 errors during OSF client downloads. All preliminary analyses and experiments were ran on ZuCo V1.

4 Methods

4.1 Replication of Baseline

As our baseline model, we replicated the R1 model from Murad et al. (2025)[1].

4.1.1 Model Architecture

EEG Encoder: A bidirectional LSTM to capture temporal and contextual dependencies within the EEG sequential data.

[1]We utilized and modified the code and model architecture from the official repository for R1 baseline: https://github.com/saydulakbarmurad/EEG-To-text

Linear Projection: The output hidden states from the LSTM are mapped via a linear + ReLU layer into the embedding space required by BART.

Pretrained BART Decoder: A transformer-based seq2seq model that generates text from the EEG-derived embeddings.

4.1.2 Training

We used Google Colab NVIDIA A100 for both training and evaluation.

Stage 1: To align the EEG encoder outputs with BART input space, we train the Bi-LSTM from scratch while freezing most BART parameters except for a few input layers to avoid disruptive updates to the pretrained BART representations.

Stage 2: Once a stable alignment was achieved, all model parameters were unfrozen for comprehensive end-to-end full fine-tuning (all parameters of both BART and the Bi-LSTM).

Optimization: Minimize cross-entropy loss using Stochastic Gradient Descent with momentum.

4.2 Model Expansions Over Baseline

Decoding Variants and Beam Size Analysis Evaluated different decoding strategies, including greedy decoding, moderate beam search, and wider beam settings.

Penalty Strategies Based on beam size 3, which our beam analysis identified as optimal, we applied penalty mechanisms during beam search to control output quality and reduce repetitive generations. We tested repetition penalty to discourage word-level repetition, no-repeat n-gram size to block repeated multi-word phrases, and length penalty to adjust output length preference.

External Language Model Rescoring Reranked the candidate outputs from beam search using external pre-trained language models of different sizes (ranging from lightweight models to large-scale LLMs). Evaluated whether additional language modeling knowledge improves fluency, grammar, and semantic coherence in EEG-to-text decoding. Explored rescoring methods such as log-probability averaging, length-normalized scoring, and ranking with added penalties.

Context-Aware Rescoring Incorporated prior decoded sentences as context when rescoring beams to promote overall coherence in multi-sentence decoding tasks.

5 Experiments & Results

5.1 Baseline Replication

Our baseline replication achieved comparable and slightly better results.

Hyperparameter Sweep We perform a sweep over the hyperparameters in Table 3. Each run follows a two-stage training schedule: train for ep-1 epochs with learning rate lr1, then fine-tune for ep-2 epochs with lr2. We vary ep-1, ep-2, lr1, lr2, and the batch size, while keeping all other factors fixed to observe the effects of hyperparameter tuning.

Model	lr1	lr2	ep-1	ep-2	batch
M_1	0.00002	0.00002	20	30	32
M_2	0.00001	0.00002	30	40	32
M_3	0.00005	0.00005	20	30	32
M_4	0.00001	0.00002	20	30	64
M_5	0.00003	0.00005	20	30	128
M_6	0.00003	0.00001	20	30	32
M_7	0.00005	0.00003	20	30	128
M_8	0.00005	0.00003	20	30	32
M_9	0.00002	0.00001	30	40	32

Table 3: Baseline experimentation with different hyperparameter settings

Results & Evaluation We assess the performance using BLEU, SacreBLEU, ROUGE-1/2/L, Word Error Rate (WER), and Character Error Rate (CER) and compare

Table 2: Evaluation of EEG-to-Text Decoding Performance on the ZuCo Dataset (SR, NR, TSR): results with(w/tf) and without teacher forcing.

Model	BLEU-N (%)				ROUGE-1 (%)			ROUGE-2 (%)			ROUGE-L (%)		
	N=1	N=2	N=3	N=4	P	R	F	P	R	F	P	R	F
R1 w/tf	38.62	21.41	11.65	6.15	30.91	25.52	27.79	6.90	6.13	6.45	28.79	23.78	25.90
M_8 **w/tf**	**40.10**	**22.43**	**12.70**	**7.13**	**31.30**	**26.20**	**28.40**	**7.45**	**6.80**	**7.09**	**29.30**	**24.60**	**26.70**
R1	12.16	3.22	1.12	0.34	12.01	9.51	9.96	1.35	0.84	0.96	10.58	8.41	8.75
M_8	**14.20**	3.16	1.02	**0.40**	**12.50**	**11.80**	**11.40**	0.87	**0.85**	0.80	**10.80**	**10.50**	**10.00**

the results with other baselines. Each model M_i is evaluated both with and without teacher forcing. Our best model M_8's results are reported in Table 2. We see that our baseline model with teacher forcing performs slightly better across all BLEU and ROUGE-1/2/L metrics compared to the R1 model by Murad et al. (2025). Without teacher forcing, our baseline remains competitive with R1, outperforming for some metrics.

Additionally, when M_8 is evaluated without teacher forcing, it gets a SacreBLEU of 0.46, WER of 1.10 and CER of 0.87. Under teacher forcing, the model achieves SacreBLEU of 6.07, with WER reduced to 0.78 and CER to 0.60. Our best model M_8 will serve as the default for our subsequent experiments.

Target string:
`These events soon led to the American Civil War.`
Predicted string with TF:
`are are became to the formation Civil War.`
Predicted string without TF:
`He was elected to the United States House of Representatives in 1894.`

Figure 2: Ex. comparison of ground truth vs. M_8 model predictions with and without teacher forcing.

5.2 Expansions

5.2.1 Beam Search Size Analysis

We systematically evaluated beam sizes of 1 (greedy), 3, 5 (baseline), 10, 15, and 20 using autoregressive generation without teacher forcing on the ZuCo V1 test set.

Beam	B-1	B-2	B-3	B-4	Sacre	R-1	R-2	R-L
1	15.85	4.43	1.70	0.68	0.79	12.93	**1.57**	**11.59**
3	**18.11**	**5.63**	**2.06**	**0.77**	**0.80**	**13.47**	1.40	11.49
5 (M8)	14.20	3.16	1.02	0.40	0.46	11.40	0.80	10.00
10	13.01	3.24	1.31	0.68	0.75	10.24	0.94	9.50
15	10.81	2.38	0.63	0.00	0.17	10.89	0.89	9.86
20	10.81	2.38	0.63	0.00	0.17	10.89	0.89	9.86

Table 4: Effect of Beam Size on performance metrics. B denotes BLEU and R denotes ROUGE. Beam 5 represents the Baseline (M8).

Table 4 shows that beam size 3 achieved optimal performance across most metrics, improving BLEU-1 from 14.20% to 18.11% and SacreBLEU from 0.46 to 0.80 compared to the standard beam 5 baseline. Surprisingly, greedy decoding (beam 1) remained competitive with BLEU-1 of 15.85% and SacreBLEU of 0.79, outperforming beam 5 while offering significant computational advantages.

In contrast, larger beam sizes showed consistent performance degradation. Beams 15 and 20 converged to identical poor performance (BLEU-1: 10.81%, SacreBLEU: 0.17, BLEU-4: 0.00%), indicating search space saturation at wider beam sizes. These findings suggest that simpler decoding strategies may be more suitable for EEG-to-text tasks, possibly due to the noisy nature of EEG signals.

5.2.2 Penalty Strategies

We systematically evaluated penalty strategies to optimize decoding performance using beam size 3.

Configuration	B-1	B-2	B-3	B-4	Sacre	R-1	R-2	R-L
Baseline (M8)	14.20	3.16	1.02	0.40	0.46	11.40	0.80	10.00
b=3, rep=1.2	13.51	3.82	1.51	0.61	0.74	13.09	**1.60**	**11.55**
b=3, rep=1.5	13.51	3.82	1.51	0.61	0.74	13.09	**1.60**	**11.55**
b=3, ng=2	**18.81**	**5.82**	**2.18**	**0.79**	**0.82**	12.90	1.40	11.01
b=3, ng=3	18.11	5.63	2.06	0.77	0.80	**13.47**	1.40	11.49
b=3, ng=2, lp=0.8	13.51	3.82	1.51	0.61	0.74	13.09	**1.60**	**11.55**

Table 5: Performance comparison of various hyperparameter configurations. 'b' denotes beam size, 'ng' denotes no-repeat n-gram size, 'rep' denotes repetition penalty, and 'lp' denotes length penalty.

Table 5 shows that the no-repeat n-gram constraint proved most effective and improved on the original unoptimized Baseline M8. ng=2 achieved the highest BLEU scores (B-1: 18.81%, B-2: 5.82%, B-3: 2.18%, B-4: 0.79%) and SacreBLEU (0.82), while ng=3 yielded the best ROUGE-1 (13.47%) with slightly lower BLEU performance.

On the other hand, repetition penalty configurations (rep=1.2 and rep=1.5) produced identical results, suggesting that the model generates limited token-level repetition and that increasing the penalty beyond 1.2 provides no additional benefits. These configurations improved ROUGE-2 (1.60) and ROUGE-L (11.55) but underperformed on BLEU metrics compared to the n-gram constraints.

The length penalty configuration (b=3, ng=2, lp=0.8) yielded the same results as the repetition penalty settings, indicating that the length penalty negated the benefits of the n-gram constraint. This suggests that encouraging shorter outputs may not be desirable for this task.

Overall, these findings indicate that n-gram-level repetition control is more effective than token-level penalties for EEG-to-text decoding, with ng=2 offering the best balance across metrics.

5.2.3 External LM Rescoring

We implemented an external LM rescoring system on top of our best configuration (b=3, ng=2). The goal was to utilize the linguistic knowledge of a pre-trained language model to select the most fluent sentences from the candidates generated during beam search. We used Qwen2-7B for our external scoring, taking the N-best full sentence candidates generated by our optimized beam search and reranking them based on the likelihood scores assigned by the LM. Our hypothesis was that the external LM would prioritize grammatically correct and semantically meaningful sentences that the EEG-decoder might have scored lower due to signal noise.

Model	B-1	B-2	B-3	B-4	Sacre	R-1	R-2	R-L
M8	14.20	3.16	1.02	0.40	0.46	11.40	0.80	10.00
b=3, ng=2	**18.81**	**5.82**	**2.18**	**0.79**	**0.82**	**12.90**	1.40	11.01
+Qwen	16.17	4.33	1.67	0.67	**0.82**	12.65	**1.50**	**11.17**

Table 6: Comparison of the model with Qwen against the baseline (M8) and b=3 ng=2 model from 5.2.2. 'b' denotes beam size and 'ng' denotes no-repeat n-gram size.

While the integration of external LM rescoring improved over our original unoptimized Baseline M8, it did not provide an improvement over the best optimized results of Beam Size 3 with an N-gram penalty of 2.

We believe this is due to two limiting factors:

- Metric Limitations: Evaluation metrics like BLEU and ROUGE rely heavily on exact word overlap. While the resulting sentence may be semantically valid, they might have low n-gram overlap with the ground truth, resulting in lower scores despite potentially higher readability.

- Model Choice: We limited our experiments to a single external model, Qwen2-7B, which is relatively small compared to state-of-the-art LLMs (70B+ parameters). Larger models with greater linguistic knowledge may yield better rescoring results. Additionally, since every LLM is trained differently, Qwen might simply not have been the best fit for our specific task with the ZuCo dataset.

5.2.4 Context-Aware Rescoring

Building on external LM rescoring, we explored whether incorporating contextual information from previously decoded sentences could improve multi-sentence coherence. Our hypothesis was that context-aware rescoring would help maintain narrative flow and topic consistency when decoding sequences of sentences from EEG signals, addressing the problem of independent sentence-by-sentence generation.

We computed P(candidate | previous sentences), while maintaining a sliding window of the N previously decoded sentences to provide relevant context for rescoring (N=5 by default).

Model	B-1	B-2	B-3	B-4	Sacre	R-1	R-2	R-L
M8	14.20	3.16	1.02	0.40	0.46	11.40	0.80	10.00
b=3, ng=2	**18.81**	**5.82**	**2.18**	**0.79**	**0.82**	**12.90**	1.40	11.01
+Qwen	16.17	4.33	1.67	0.67	**0.82**	12.65	**1.50**	11.17
+Context-aware	16.25	3.21	0.95	0.00	0.23	12.67	1.11	**11.19**

Table 7: Comparison of the model with Qwen context-aware rescoring against the baseline (M8), b=3 ng=2 model from 5.2.2, and model with Qwen from 5.2.3. 'b' denotes beam size and 'ng' denotes no-repeat n-gram size.

Context-aware rescoring failed to provide improvements over any previous configuration. In fact, compared to our best results (b=3, ng=2) it significantly degraded performance across most metrics, with B-4 dropping to 0.00 and SacreBLEU falling from 0.82 to 0.23. Only ROUGE-L showed very slight improvement (11.19). This suggests that incorporating previous sentences as context was counterproductive for this task.

We attribute this failure to several factors:

- Error Propagation: Since previously decoded sentences contain errors, using them as context compounds these mistakes. The model conditions on imperfect outputs, leading to increasingly incoherent predictions.

- Dataset Characteristics: The ZuCo dataset consists of sentences from movie reviews that may lack strong dependencies beteween sentences. Unlike narrative text with clear continuity, review sentences can shift topics abruptly, making contextual priors less useful.

- Distribution Mismatch: The LM predicts continuations based on typical language patterns, which may not match the actual ground truth sequences, pushing candidates toward fluent but semantically incorrect outputs.

- Furthermore, as noted in Section 5.2.3, BLEU and ROUGE rely on exact word overlap and

may penalize semantically valid outputs that lack n-gram matches with the ground truth.

These results suggest that for this task, treating each sentence independently may be more effective than attempting to model cross-sentence coherence, at least with current model capabilities. More broadly, our rescoring experiments (Sections 5.2.3 and 5.2.4) indicate that post-processing techniques cannot compensate for fundamental limitations in how the model processes EEG signals. Meaningful improvements will likely require architectural innovations rather than downstream refinements.

6 Limitations and Future Work

Semantic Evaluation Metrics Our analysis highlighted the limitations of metrics like BLEU and ROUGE, which penalize semantically correct paraphrases. Future work should adopt semantic evaluation metrics. These metrics utilize contextual embeddings to measure similarity in meaning rather than just exact word overlap, providing a fairer assessment of the external LLM's ability to improve fluency and coherence.

Addressing the EEG-Text Alignment Problem There is a fundamental limitation in current EEG-to-text architectures: the big performance gap between teacher forcing and realistic decoding due to poor alignment between EEG signal representations and natural language semantics. Both external LM rescoring and context-aware rescoring failed to bridge this gap, demonstrating that post-processing techniques did not compensate for architectural weaknesses. We believe that the bottleneck lies in how the model encodes brain signals, not in linguistic refinement.

Future work should prioritize architectural work that explicitly enforce semantic alignment between EEG embeddings and text representations. Contrastive learning approaches, such as adapting CLIP-style frameworks to learn a shared embedding space for EEG signals and language, could force the encoder to capture meaningful semantic information rather than relying solely on autoregressive patterns from ground-truth tokens. These methods could align EEG and text representations before generation, potentially eliminating the model's tendency to ignore brain signals during inference and achieving better performance.

7 Conclusion

In this project, we extended the R1 EEG-to-text translation model using the ZuCo V1 dataset to better understand the challenges of decoding natural language from EEG signals. We performed hyperparameter tuning and systematically evaluated different decoding and post-processing strategies. These experiments led to small but consistent improvements over the baseline. We also found that simpler decoding methods, such as smaller beam sizes with n-gram constraints, performed better than more complex alternatives. This suggests that standard assumptions from neural machine translation do not directly transfer to EEG-to-text tasks, likely due to the noisy and high-variance nature of EEG signals.

Our experiments showed that post-processing methods, including external language model rescoring and context-aware decoding, did not close the large gap between teacher-forced and real-world decoding. This suggests that current models struggle to align EEG signals with linguistic meaning. Overall, this project helped identify the limits of decoding-based improvements and pointed to the need for more fundamental model changes.

References

Ran Duan, Pengfei Liu, and Heng Ji. 2023. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*.

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*, 5:180291.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 138–146, Marseille, France. European Language Resources Association.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. 2024. Are EEG-to-Text models working? *Computing Research Repository*, arXiv:2405.06459.

Pengfei Liu, Ran Duan, Yizhou Wang, and Heng Ji. 2024. Eeg2text: Open vocabulary eeg-to-text translation with multi-view transformer. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1824–1833. IEEE.

Saydul Akbar Murad, Ashim Dahal, and Nick Rahimi. 2025. EEG-to-Text translation: A model for deciphering human brain activity. *Computing Research Repository*, arXiv:2505.13936.

Anne Porbadnigk, Marek Wester, Jan Calliess, and Tanja Schultz. 2009. Eeg-based speech recognition - impact of temporal effects. In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*, pages 376–381. INSTICC Press.

Linjun Shi, Jinchuan Ma, Qian Dong, Sheng Zhou, and Yicheng Gong. 2024. Progres: Prompted generative rescoring on asr n-best. *arXiv preprint arXiv:2409.00217*.

Yizhou Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3611–3625. Association for Computational Linguistics.