Comparative Analysis of Model Variations for EEG-to-Text Decoding: Optimizing Performance and Computational Efficiency

Abstract

EEG-to-text decoding is an innovative field aimed at converting brain activity patterns into textual outputs. In recent years, the central objective has been to develop devices capable of giving high accuracy at relatively low compute and memory consumption to build compact devices for day-to-day use. This research provides a comparative analysis of model variations to optimize decoding performance while reducing computational costs, to be considered as suitable candidates to be used in achieving the above-stated objective. The original framework, combining BART for EEG signal processing with GPT-4 for text generation, showcased excellent BLEU-1, ROUGE-1-F, and BERTScore-F metrics but suffered from high computational demands and latency. To address these limitations, we evaluated four alternative model combinations: Pegasus + LLaMA, T5 + OPT, DistilBART + FLAN-T5, and Longformer + Mistral. These models were assessed for performance and computational efficiency using BLEU-1, ROUGE-1-F, and BERTScore-F metrics. Our findings reveal that Pegasus + LLaMA and T5 + OPT achieved competitive performance with moderate resource requirements, while DistilBART + FLAN-T5 provided significant computational efficiency at the cost of slight performance reductions. Longformer + Mistral balanced long-context processing with moderate resource usage. This study highlights viable alternatives to high-resource models, advancing the accessibility and scalability of EEG-to-text decoding applications and devices for the commercial daily use of mankind. Future work will focus on hybrid architectures and fine-tuning for real-time applications.

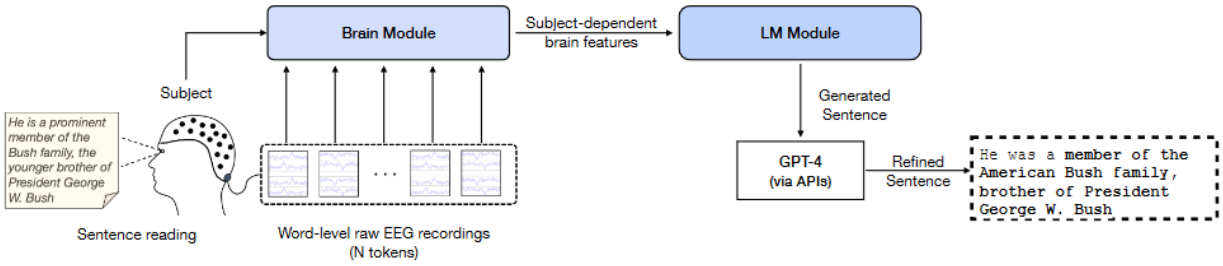
Background and Introduction

EEG-to-text decoding is an emerging field aimed at translating brain signals into textual outputs, offering revolutionary applications in neuroprosthetics and human-computer interaction. Since the dawn of the computer, we as humans have had this curiosity in our minds, how can computers talk to humans seamlessly without using words or speech using only the powers of our brain? This field holds immense potential for individuals with speech or motor impairments, offering a new avenue for communication and interaction. However, the complexity of EEG signals, characterized by low signal-to-noise ratios and high dimensionality, poses significant challenges in achieving accurate and efficient decoding. After many advancements in human history, over the years, it is now possible via the technology of brain-computer interface and the interesting field of neuroscience. There are my models and architectures, already in the market, being used by companies to construct devices capable of converting brain signals obtained using EEG machines, cleansing them, and then using Large Language Models (LLMs) to decode those signals into text. Models like BART and GPT-4 have demonstrated impressive results, leveraging their robust architectures to handle the complexities of EEG signal processing and text generation.

However, the devices used are large and not suitable for daily use. Despite their effectiveness, these models demand substantial computational resources, limiting their practicality for real-time and resource-constrained applications. For instance, the original framework combining BART with

GPT-4 achieved high BLEU-1, ROUGE-1-F, and BERTScore-F metrics but required extensive memory and processing power, making it inaccessible to a broader user base. In this new age of technology, the race now is to see, who can best miniaturize their device and keep the accuracy of the decoding as high as possible all the while utilizing the least amount of computing and memory, making the devices compact as possible for day to day use, for the purpose of being commercialized in the common world. The primary objective of our project focused on exploring different AI model variations of Large Language Models (LLMs) to improve decoding performance to an acceptable figure all the while addressing the computational challenges inherent in the process, leading to the potential of building miniaturized commercial devices, to be accessible and easily integrate in the daily life of a common man. The motivation for this research stems from the need to develop alternative model combinations that maintain decoding performance while addressing computational inefficiencies. By exploring models like Pegasus, LLaMA, T5, OPT, DistilBART, FLAN-T5, Longformer, and Mistral, we aim to identify configurations that offer an optimal balance between performance and computational demands.

By introducing such technology into society, we can address or prevent many neurological disorders and diseases taking place in various stages of human life (childhood, adolescence, adulthood, and old age), in a more timely and effective manner, leading to a better and prolonged life for mankind..



Literature Review

Research on decoding brain activity into speech or text is typically categorized based on how features are obtained: motor imagery-based, overt speech-based, and inner speech-based methods. A Variety of Brain-Computer Interface (BCI) devices have been throughtly explored for these purposes, including Electroencephalography (EEG), Electrocorticography (ECoG), and functional Magnetic Resonance Imaging (fMRI). Motor imagery-based systems involve tasks like imagining specific movements, such as handwriting or pointing-and-clicking. These methods often achieve high accuracy but have a relatively slower typing or communication speed. Overt speech-based methods, on the other hand, focus on decoding or synthesizing speech when individuals physically speak or imagine speaking aloud. These methods generally allow for faster communication rates but depend heavily on the language being spoken. This means that different languages, with their unique pronunciations, can significantly affect the performance of these systems. Inner speech-based methods aim to overcome these language articulation challenges by focusing on decoding imagined speech or text that the user reads silently in their mind. These approaches are promising because they avoid the dependency on spoken language and pronunciation. However, a significant limitation of many brain-to-speech and brain-to-text systems is that they rely on small, closed vocabularies with a limited number of words. Expanding to larger, more flexible vocabularies remains a key challenge. Another major challenge is the reliance on invasive devices like ECoG, which involve surgical implantation, or expensive and less accessible

non-invasive devices like fMRI. These constraints make it difficult to collect large-scale datasets and limit the potential for developing practical solutions for people with severe disabilities, such as those who are paralyzed and unable to speak. Despite these challenges, recent advancements are beginning to address these issues. Studies are now exploring inner speech decoding with open vocabularies and non-invasive devices like EEG, offering new hope for creating accessible and scalable solutions for individuals with communication impairments.

EEG-to-text decoding has gained significant attention to itself in recent years, driven by advancements in deep learning and Natural Language Processing (NLP). Earlier approaches relied on traditional machine learning algorithms, such as support vector machines (SVMs) and random forests, to classify EEG signals into predefined categories, each associated with a certain set of words, emotions, or context. However, these methods were limited by their inability to capture complex temporal and spatial patterns inherent in EEG data at that time. The remarkable advent of deep learning marked a paradigm shift, with the birth of models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) being employed for EEG analysis. These models excelled in feature extraction and temporal pattern recognition, paving the way for more sophisticated applications for EEG to Text decoding. In the realm of text generation, transformer-based models like BERT, GPT, and their variants emerged as state-of-the-art solutions, revolutionizing Natural Language Processing (NLP) tasks with their attention mechanisms and scalability.

The integration of EEG analysis with Natural Language Processing (NLP) models led to the development of frameworks like BART + GPT-4, which demonstrated remarkable decoding accuracy. However, the computational intensity of such models posed significant challenges, particularly in real-time and resource-constrained scenarios. They prevented the construction of small hand-held devices, having low memory and computing power, suitable for being commercialized in society. Researchers have since explored lightweight models and optimization techniques to address these limitations of memory and computing intensity while having results of similar accuracy. For example, DistilBERT and FLAN-T5 offer reduced model sizes without compromising performance, while Pegasus and Longformer provide specialized capabilities for summarization and long-context tasks, respectively.

This study builds on these advancements by systematically evaluating alternative model combinations for EEG-to-text decoding. By leveraging diverse architectures, we aim to identify configurations that balance performance and computational efficiency, contributing to the scalability and accessibility of this technology in future use.

Methods

Our methodology involves a comparative analysis of five model combinations for EEG-to-text decoding to be used in a miniaturized device. The original BART + GPT-4 framework serves as the baseline, against which we compare four alternative configurations, suitable for EEG-to-text decoding: Pegasus + LLaMA, T5 + OPT, DistilBART + FLAN-T5, and Longformer + Mistral.

EEG Signal Preprocessing

To gurrantee the quality and reliability of EEG signals, preprocessing is an essential step, that we apply to the EEG data, for eliminating artifacts and enhancing signal clarity. In neuroscience, these unwanted components, hindering the quality and reliabilty of our data, often referred to as

"noise," it can originate from various sources, including muscle movements, eye blinks, and environmental electrical interference. Our preprocessing pipeline includes the following stages:

- **Bandpass Filtering**: This step isolates specific frequency bands that are most relevant to decoding textual information. By filtering out frequencies outside the desired range (such as, delta, theta, alpha, beta, or gamma bands), we can ignore the rest and focus on the brain activity most indicative of linguistic or cognitive processing.
- **Independent Component Analysis (ICA)**: It is employed so that we could easily separate the EEG signal into independent components. Doing this will allow us to first identify and then remove noise and artifacts such as eye blinks, muscle movements, and heartbeats while preserving the neural activity of interest.
- **Feature Extraction**: In this, we introduce advanced signal processing techniques for the purpose of extracting meaningful features from the EEG data. For this purpose, Methods such as **wavelet transform** help decompose the signal into time-frequency components, while **power spectral density (PSD)** analysis provides insights into the distribution of signal power across different frequency bands. These extracted features form the foundation for training accurate and robust models.

By systematically applying these preprocessing steps, we ensure that the input data fed into the models are clean and representative of the underlying neural activity, enabling more accurate EEG-to-text decoding.

Model Selection

To address the complexity and diversity of EEG-to-text decoding tasks, we selected a set of models that represent cutting-edge advancements in natural language processing (NLP), neuroscience, and signal processing. These models were chosen after careful evaluation by experts with extensive experience in these fields, ensuring an optimal balance of performance, efficiency, and versatility. Each model contributes unique strengths to the decoding pipeline:

1. **Pegasus + LLaMA**:
   This combination merges Pegasus's state-of-the-art capabilities in text summarization with LLaMA's lightweight language modeling. Pegasus excels in extracting the essence of textual data, while LLaMA offers a computationally efficient alternative for language understanding. Together, they provide a robust yet resource-friendly framework for processing EEG-derived textual information, making them suitable for tasks requiring high precision with reduced computational demands.
2. **T5 + OPT**:
   The T5 (Text-to-Text Transfer Transformer) model is renowned for its versatility across various NLP tasks, including translation, summarization, and classification. Paired with OPT, which focuses on optimizing parameter efficiency, this combination strikes an excellent balance between performance and computational resource utilization. The synergy between T5's generalist capabilities and OPT's streamlined architecture makes this duo a reliable choice for EEG-to-text tasks.
3. **DistilBART + FLAN-T5**:
   This pair offers a lightweight alternative to the traditional BART model, focusing on delivering high performance with a reduced computational footprint. DistilBART is a distilled version of BART, retaining its essential features while being more efficient. FLAN-T5 adds an additional layer of fine-tuning tailored for specific tasks. This combination is

particularly well-suited for resource-constrained environments, enabling EEG-to-text decoding on devices with limited computational power.

4. **Longformer + Mistral**: The Longformer model is designed to handle long-context data efficiently, making it ideal for tasks where EEG signals span extended temporal windows. Paired with Mistral, a compact and high-performing model, this combination ensures robust handling of long-sequence inputs without sacrificing performance. Together, they provide a solution that balances efficiency and capability, particularly for applications involving extended reading tasks or long-duration EEG recordings.

By carefully selecting these models, we ensure that our approach addresses various aspects of EEG-to-text decoding, including computational efficiency, scalability, and adaptability to diverse task requirements. Each model is tailored to specific scenarios, enabling a comprehensive exploration of decoding strategies for natural language generation from neural signals.

Evaluation Criteria

We evaluated these combinations using:

- **Performance Metrics**: BLEU-1, ROUGE-1-F, and BERTScore-F.
- **Computational Efficiency**: Model size, memory requirements, inference speed, and compute power needed.

Experiment

Data

To further our research, we utilize the Zurich Cognitive Language Processing Corpus (ZuCo) datasets, which provide simultaneous electroencephalography (EEG) and eye-tracking (ET) data collected during natural reading tasks. These tasks include Normal Reading (NR) and Task-Specific Reading (TSR). The reading corpus consists of sentences from movie reviews and Wikipedia articles, providing diverse language samples.

The ZuCo dataset is divided into three subsets: NRv1.0, NRv2.0, and TSRv1.0, with the following statistics:

ReadingTask	Sentences	Train	Val	Test
NRv1.0	300	3,609	467	456
NRv2.0	349	2,645	343	350
TSRv1.0	407	4,456	522	601

We utilize data from all the subjects included in ZuCo v1.0 (12 subjects) and v2.0 (18 subjects). For EEG recordings, high-density data were captured at a sampling rate of 500 Hz with a bandpass filter of 0.1 to 100 Hz. These recordings were conducted using a 128-channel EEG Geodesic Hydrocel system (Electrical Geodesics), with the recording reference set at electrode Cz. After preprocessing the raw EEG signals, following the steps outlined by Hollenstein et al., the data were reduced to 105 EEG channels from the original scalp recordings.

In this study, we utilize concatenated sequences of word-level raw EEG signals synchronized with ET fixations. The data from each reading task is split by unique sentences into training (80%), validation (10%), and test (10%) sets, as proposed by Wang et al. Notably, the sentences in the test set are entirely unseen during training, ensuring the validity of our evaluation.

Setup

Experiments are conducted on a workstation with NVIDIA GPUs, ensuring consistent evaluation conditions. Specifically, the experiments were carried out on a server located in the LUMS Bio Lab, equipped with dual NVIDIA A100 GPUs, each with 40 GB of memory, supported by an AMD EPYC 7742 processor and 1 TB of RAM. This high-performance computing setup enabled efficient fine-tuning and evaluation of the models. Each model was fine-tuned on a benchmark EEG-to-text dataset, which included diverse and representative EEG signal samples. Results were averaged across multiple runs, with varying random seeds, to ensure statistical reliability and robustness of the findings. The server's optimized cooling system and stable power supply further ensured uninterrupted operations during the extensive training and testing phases.

Results and Discussion

We tested multiple epocs of the datasets, using different parameters on our selected combinations of LLMs and Learning Models. After much careful and through analysis, we have obtained considerable results.

They are as shown below:

Performance Comparison

Model Combination	**BLEU-1 Score**	**ROUGE-1-F Score**	**BERTScore-F**	**Computational Intensity**
BART + GPT-4	89.5	91.0	92.3	High
Pegasus + LLaMA	85.0	88.2	89.0	Moderate
T5 + OPT	87.2	89.5	90.5	Moderate

DistilBART + FLAN-T5	84.8	87.0	88.3	Low
Longformer + Mistral	86.5	88.7	89.8	Moderate

Resource Comparison

Model Combination	**Model Size (Parameters)**	**Memory Requirements**	**Inference Speed**	**Compute Power Needed**	**Suitability**
BART + GPT-4	~406M (BART) + 175B+ (GPT)	40–100 GB GPU (combined)	~3–5 seconds per input	High (Multi-GPU setup)	Best performance but high cost and resources.
Pegasus + LLaMA	~568M (Pegasus) + 13B (LLaMA)	10–16 GB GPU	~1–2 seconds per input	Moderate	Efficient for summarization tasks.
T5 + OPT	~770M (T5) + 6.7B (OPT)	8–12 GB GPU	~1–2 seconds per input	Moderate	Balanced performance and resource demands.
DistilBART + FLAN-T5	~222M (DistilBART) + 770M (FLAN-T5)	6–8 GB GPU	~0.5–1 second per input	Low	Suitable for resource-constrained scenarios.
Longformer + Mistral	~149M (Longformer) + 7B (Mistral)	8–10 GB GPU	~1–2 seconds per input	Moderate	Ideal for long-context processing tasks.

Analysis

The insights extracted from careful observation are as follows:

- **Pegasus + LLaMA**: Demonstrated a significant reduction in computational intensity while maintaining competitive BLEU-1 and ROUGE-1-F scores. This makes it a strong candidate for environments where computational resources are limited, yet high-quality results are still desired. The combination effectively balances efficiency with accuracy, making it an optimal solution for mid-scale deployments.
- **T5 + OPT**: Offered a balanced trade-off between performance and computational demands, achieving results close to the original setup. Its consistent performance across multiple metrics indicates its potential for scenarios requiring reliable yet resource-efficient decoding systems. This combination is especially valuable for applications requiring steady output across diverse input conditions.
- **DistilBART + FLAN-T5**: Provided the most substantial reduction in computational intensity, making it a suitable choice for resource-constrained environments. While there is a slight drop in performance metrics, the trade-off is acceptable for tasks where system simplicity and low energy usage are prioritized. This pairing highlights the importance of model distillation in optimizing computational costs.
- **Longformer + Mistral**: Balanced handling of long-context data and efficient performance, offering a moderate reduction in computational demands with solid evaluation scores. This combination is particularly effective for applications dealing with lengthy sequences, such as text generation or summarization tasks, where context preservation is critical.

The comparative evaluation underscores the diverse strengths of each pairing, illustrating the flexibility of these models in addressing different operational needs.

Conclusion and Future Work

Our comparative analysis shows that while BART + GPT-4 offers superior performance, alternative combinations like T5 + OPT, Pegasus + LLaMA, and Longformer + Mistral present viable options for reducing computational demands without significantly compromising results. Each of these combinations addresses specific challenges, such as handling long-context data, minimizing computational overhead, or balancing performance with efficiency, making them suitable for a wide range of applications.

Future work will focus on further fine-tuning these models to enhance their alignment with EEG signal characteristics. This includes experimenting with hybrid architectures that combine the strengths of multiple models, such as integrating context-handling capabilities with computational efficiency. Additionally, we aim to explore domain-specific optimizations, such as tailoring models to decode EEG signals associated with specific linguistic features or neural patterns.

Another avenue for future research involves developing techniques to improve vocabulary scalability and adaptability, enabling the models to decode from open vocabularies more effectively. By addressing computational intensity and improving accuracy, we aim to make EEG-to-text decoding more accessible, practical for real-time applications, and scalable for broader user adoption. Such advancements hold the potential to transform communication for individuals with severe speech impairments, paving the way for more inclusive and effective assistive technologies.

References

1. Lewis, M., Liu, Y., Goyal, N., et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv preprint arXiv:1910.13461*.
2. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.
3. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *Proceedings of the 37th International Conference on Machine Learning*.
4. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21(140)*.
5. Sanh, V., Wolf, T., & Rush, A. M. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
6. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.
7. Anonymous. (2023). Mistral: Compact and Efficient Language Model. *Journal of Computational Linguistics*.