



Thought2Text: Text Generation from EEG Signal using Large Language Models

☀ 상태	완료
≡ Journal	arXiv(University of Texas)
≡ Year	2024.10
≡ Summary	EEG 데이터로 instruction-tuned LLMs를 fine-tuning하여 뇌의 활동을 해독하고 표현하는 Thought2Text를 제안
≡ Limitations	객체 인식의 잠재적 모호성, 데이터 부족, BCI 남용 위험 등 한계가 있지만, 다양한 task와 대규모 데이터셋으로 학습시켜 일반화를 향상시키는 모델을 만들 수 있음
🔗 Link	https://arxiv.org/abs/2410.07507v1#
≡ category	EEG Fine-tuning Image LLM Text Generation

<https://github.com/abhijitmishra/Thought2Text>

Introduction

EEG 신호로부터 텍스트 설명을 생성하는 접근법에는 세 가지 단계가 포함됨: (a) 시각적 자극을 통해 language-agnostic EEG 신호 캡처. (b) 이 신호를 deep multichannel neural encoder에 임베딩. (c) 이미지 및 EEG 임베딩을 토큰 임베딩 공간에 투영하여 언어 모델을 fine-tuning하고 응답 생성.

LLM의 응답은 표준 이미지 설명과 비교하여 training loss를 계산함. 추론 중에는 EEG 신호와 일반 텍스트 프롬프트만 LLM에 입력으로 사용되어 응답이 생성됨. 따라서, 본 방법은 training을 위해 이미지, EEG 데이터 및 이미지 설명이 필요하며, inference는 bimodal로서 텍스트 생성을 위해 오직 EEG만 사용함.

실험은 여섯 명의 참가자가 시각적 자극을 보면서 수집된 public 128채널 EEG 데이터셋을 사용하여 진행함. 이미지 설명은 GPT-4-Omini에 의해 생성하고 이를 전문가에 의해 검증하여 text modality를 제공함. MISTRAL-V3, LLaMa-V3 및 QWEN2.5와 같은 pre-trained instruction-based 언어 모델을 활용하여 LLM 파인튜닝 진행함.

Our paper's key contributions include:

- Integration of brain signals with LLMs.
- Fine-tuning models on EEG signals captured for visual stimuli, leveraging its language-agnostic nature to enhance LLM interaction and scalability across languages.
- Validation of model efficacy on a popular public dataset that contains EEG signals captured using affordable devices in less constrained environments compared to traditional methods.

Related Work

- Task
 - EEG classification ; [\[link\]](#), [\[link\]](#), [\[link\]](#)
 - image generation with GANs ; [\[link\]](#)
 - image generation with latent diffusion model ; [\[link\]](#), [\[link\]](#), [\[link\]](#)
 - EEG2Text ; [\[link\]](#)
- Datasets
 - ZuCo 2.0 [\[link\]](#) ; captures EEG and eye-gaze during natural language reading
 - MOABB [\[link\]](#) ; 120,000 EEG samples from 400+ subjects from various BCI tasks
 - CVPR 2017 [\[link\]](#) / MindBigData[\[link\]](#) ; EEG data from responses to handwritten and open vocabulary object-based image stimuli

Dataset and the Need for Visual Stimuli

<eeg, text> 데이터셋은 언어의 복잡성 측면에서 부적절함. 데이터셋은 두 가지 주요 측면에서 생각해야 함: (a) 언어 처리가 최소한으로 방해받는 language-agnostic neural signals 활용, (b) 이러한 신호를 사용하여 목표 언어로 텍스트를 생성하는 것. → 따라서, <eeg, text, image>로 구성된 삼변량 데이터셋이 필요함.

본 논문에서는 CVPR 2017 데이터셋으로 진행함. 6명의 참가자가 50개의 이미지를 보는 동안 기록된 EEG 데이터를 포함하고 있는 데이터. 각 EEG 기록은 하나의 참가자-하나의 시각 자극에 해당하며, 128채널이 1kHz의 샘플링 속도로 0.5초 동안 기록됨. 128×N 행렬로 표현되며, 여기서 N은 각 세그먼트에서 채널당 샘플 수를 나타내며 대략 500 정도임. → 전처리 방법 ; [\[link\]](#) → 본 데이터셋은 text가 없기 때문에 GPT-4로 한 줄 text 생성 후 정확성 검증하기 위해 Amazon Mechanical Turk 사용함.

Method

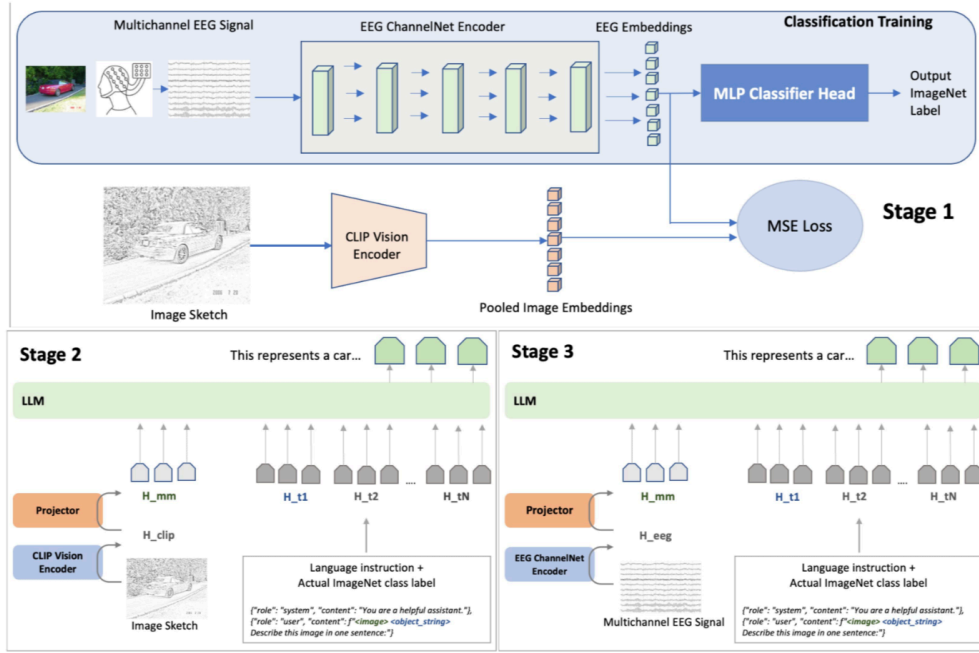


Figure 1: Multi stage training process for *Thought2Text* solution. **Stage1:** EEG ChannelNet generates embeddings from raw EEG signals. The encoder is trained using a combined loss function, which includes: (1) MSE loss from aligning the EEG embeddings with the image sketch pooled embeddings generated by the CLIP model, and (2) CE loss for accurate prediction of EEG classifications based on the ground truth ImageNet labels. **Stage2:** LLMs are fine-tuned using a generic image description prompt along with ground truth object labels. The token embedding layer integrates token embeddings from the prompt with multimodal image embeddings generated by a projector trained on these ground truth labels and available image descriptions. Only the projector is trained in this stage, while the LLM and CLIP encoder remain frozen. **Stage3:** Similar to Stage 2, but in this stage the projector undergoes further training using EEG embeddings derived from EEG signal data instead of CLIP embeddings from image data. The LLM and EEG encoder remains frozen. In both stages 2 and 3, *< image >* token embeddings in the input prompt are replaced by H_{mm} embeddings, while H_{t1} embeddings are extracted from the *< object_string >*.

Stage1: Training EEG Encoder for Embedding Extraction

두 가지 목적을 가진 인코더 설계: (a) EEG 임베딩을 pretrained visual encoder를 사용하여 이미지 자극으로부터 유도된 임베딩과 정렬하고, (b) EEG 임베딩에서 가장 두드러진 객체(ex. 피아노)를 예측함. Multichannel EEG encoder(inspired by ChannelNet[link])를 사용하여 EEG 신호를 다차원 임베딩(H_{eeg})으로 변환함. 이 임베딩은 MLP classifier를 통해 처리되어 이미지 자극에 있는 객체에 해당하는 객체 레이블(y_{obj})을 예측함. 여기서 레이블 집합은 데이터셋의 일부로 제공되는 ImageNet의 레이블과 일치하며, 각 이미지에서 가장 두드러진 객체를 표현함.

학습은 두 가지 손실을 최소화하는 방식으로 훈련됨: (a) 예측된 객체 레이블과 실제 객체 레이블 간 categorical cross-entropy loss(CE), (b) EEG 임베딩(H_{eeg})과 pretrained CLIP 모델에서 추출된 풀링된 이미지 임베딩(H_{clip}) 간 mean squared error(MSE). 아래는 전체 손실함수의 식이며, 하이퍼파라미터 값은 0.5로 설정함.

$$L = (1 - \alpha) \cdot MSE(H_{eeg}, H_{clip}) + \alpha \cdot CE(y_{obj}, \hat{y}_{obj})$$

EEG 임베딩을 시각 자극과 더 잘 정렬하기 위해, 이미지를 단순화하는 작업을 거침. 색상과 같은 비중심 세부사항 제거 후, Gaussian Blur 및 Canny filter와 같은 기법을 사용하여 이미지를 스케치 형태로 변환함.

세 가지 이유로 H_{eeg} 와 y_{obj} 를 모두 예측함: (a) EEG 임베딩을 이미지 임베딩과 정렬함으로써 multimodal vision-language 모델을 활용하며 EEG-based text generation을 위한 pretrained model을 제공할 수 있음, (b) 공동 최적화는 임베딩이 이미지의 두드러진 객체를 강조하도록 보장함, (c) EEG 임베딩과 결합된 객체 레이블은 나중에 더욱 정확한 생성을 유도하기 위해 multimodal language model에 공급될 수 있음.

Stage2: Priming LLMs with Image Embeddings

LLM이 EEG 및 시각 임베딩과 같은 multimodal input을 처리할 수 있도록 projector(feed-forward layer)를 설계함. 이는 비전 및 EEG 모델의 임베딩(H_{clip} , H_{eeg})을 LLM의 토큰 임베딩 공간으로 변환함. 즉, projected된 임베딩이 LLM의 토큰 임베딩과 동일한 차원을 가지도록 해야 함. → 변환된 임베딩은 입력 프롬프트의 토큰 임베딩과 concat됨.

```
{ "role": "system", "content": "You are a helpful assistant." },
{ "role": "user", "content": "<image> <object_string> Describe this in one sentence:" },
```

Input prompt

객체 레이블을 포함한 프롬프트의 토큰을 LLM 토큰 임베딩 레이어로 임베딩함. 토큰 임베딩 $H_{t1}, H_{t2}, \dots, H_{tN}$ 은 multimodal 임베딩으로 보강됨. multimodal 임베딩 H_{mm} 은 외부 임베딩(H_{clip})을 LLM의 토큰 임베딩 공간으로 투영하여 다음 변환을 통해 계산됨:

$$H_{mm} = W_{mm} \cdot H_{clip} + b_{mm}$$

여기서 W_{mm} 과 b_{mm} 은 프로젝터의 파라미터임.

Stage3: Tuning LLMs with EEG Embeddings

Stage 1에서 훈련된 EEG 인코더에서 추출된 H_{ee} 를 사용하여 multimodal 임베딩 H_{mm} 을 계산함. 이 단계에서 프로젝터의 파라미터 W_{mm} 과 b_{mm} 은 추가 훈련을 거침. Stage 2와 동일하게 프로젝터만 훈련됨. (LLM 및 EEG 인코더는 고정)

Inference

EEG 인코더는 raw EEG 신호로부터 객체 레이블(y_{obj}) 및 EEG 임베딩(H_{ee})을 예측함. 객체 레이블(y_{obj})은 일반 프롬프트와 연결되어 토큰 임베딩을 계산하는 데 사용됨. 이전 단계에서 훈련된 multimodal projector는 H_{ee} 로부터 H_{mm} 을 계산함. 해당 임베딩은 LLM에 공급되어 설명을 생성함. 따라서, 추론 중에는 이미지는 사용 안하는 bimodal임.

Experimental Details

Dataset

- CVPR2017 dataset ; 6명이 50개의 이미지를 보는 동안 기록된 EEG 신호. 40개의 ImageNet 클래스 → 총 2000개의 이미지로 구성됨. 데이터는 training(7954), evaluation(1994), test(1987)로 분리됨.

Model Details

- EEG encoder : ChannelNet 사용 → final linear layer를 수정하여 512차원 임베딩으로 설정 (CLIP vision encoder의 output과 맞추기 위함(openai/clip-vit-base-patch32))
- Training methods : batch size 16, epochs 100, AdamW optimizer, learning rate 1e-4
- Finetuning methods : batch size 16, 5 epochs per stage, learning rate 2e-5
- Platform : PyTorch 및 Huggingface의 transformers 라이브러리
- LLM 평가 : LLaMa-v3(meta-llama/Meta-Llama-3-8B-Instruct), Mistral-v-03(mistralai/Mistral-7B-Instruct-v0.3), Qwen2.5-7B(Qwen/Qwen2.5-7B-Instruct) → 소비자 GPU(예: NVIDIA RTX 3090)에서 효율적으로 작동되는 모델 위주로 선택
- 임베딩 투영 : multimodal 임베딩 H_{mm} 은 각 LLM에서 요구하는 토큰 임베딩 차원과 맞추어 투영
- Training time : LLM 훈련 주기당 약 7 GPU 시간 소요
- Inference settings : batch size 1, top_k, top_p같은 파라미터와 temperature는 LLM 기본값으로 설정
- baselines
 - ONLY_OBJ : LLM에 추가 입력 없이 예측된 object를 기반으로 설명을 생성하는 경우 (ex. object가 "car"일 경우 LLM이 "car"라는 단어에 대한 설명을 생성)
 - ONLY_OBJ + RAND_EMB : 예측된 object label과 함께 임의의 임베딩을 LLM에 전달하는 경우
 - NO_STAGE2 : 섹션 4에서 설명된 priming step을 건너뛰는 경우
 - ONLY_EEG : object label을 무시하고 stage 1의 EEG 임베딩만 입력으로 사용하는 경우

Evaluation

BLEU, METEOR, ROUGE, BERTScore와 같은 표준 NLG 메트릭을 사용함. 추가로, GPT-4를 사용하여 생성 품질 평가함. GPT-4는 두 가지 측면을 측정함: 1) 문법을 평가하기 위한 유창성, 2) 의미 전달의 정확성을 평가하기 위한 적합성. 두 가지 모두 1-5 점 척도로 평가되며, 5는 최고 품질을 나타냄.

Results

LLM	ROUGE-N		ROUGE-L	BLEU-N		METEOR	BERT Score	GPT-4 Fluency	GPT-4 Adequacy
	N=1	N=2		N=1	N=4				
LLaMA3-8B _{ONLY_OBJ}	9.8	1.5	8.5	7.3	1.3	12.6	0.84	3.44	1.30
LLaMA3-8B _{OBJ+RAND_EMB}	3.8	0.4	3.3	2.8	0.4	5.9	0.84	4.72	1.08
LLaMA3-8B _{ONLY_EEG}	28.9	7.3	26.2	24.1	5.2	23.7	0.89	4.80	1.49
LLaMA3-8B _{NO_STAGE2}	26.9	6.1	23.9	22.6	4.3	23.7	0.88	4.83	1.41
LLaMA3-8B _{ALL}	30.0	8.1	26.6	25.5	5.5	26.3	0.89	4.82	1.58
Mistral-7B-v0.3 _{ONLY_OBJ}	17.6	3.4	14.8	14.5	2.5	23.2	0.86	4.46	1.52
Mistral-7B-v0.3 _{OBJ+RAND_EMB}	17.9	3.6	15.1	15.7	2.9	22.8	0.87	4.89	1.55
Mistral-7B-v0.3 _{ONLY_EEG}	26.7	5.3	23.5	23.3	4.2	22.0	0.88	4.82	1.25
Mistral-7B-v0.3 _{NO_STAGE2}	29.2	7.3	26.5	24.1	5.0	24.0	0.89	4.77	1.60
Mistral-7B-v0.3 _{ALL}	30.6	8.8	28.0	26.0	6.1	26.2	0.89	4.79	1.65
Qwen2.5-7B _{ONLY_OBJ}	17.6	2.8	14.5	14.8	2.4	21.0	0.85	3.91	1.47
Qwen2.5-7B _{OBJ+RAND_EMB}	1.7	0.1	1.6	1.3	0.3	6.4	0.84	4.73	1.01
Qwen2.5-7B _{ONLY_EEG}	25.2	3.6	21.5	21.9	3.2	20.2	0.88	4.77	1.10
Qwen2.5-7B _{NO_STAGE2}	24.4	4.1	20.9	20.7	3.3	20.2	0.88	4.66	1.24
Qwen2.5-7B _{ALL}	26.4	4.6	22.8	22.7	3.7	21.1	0.88	4.75	1.28

Table 2: Averaged Evaluation Metrics (%) and GPT-4 assessment of text generated from EEG signals using different LLMs. A comparison is made between chance-level performance (with only object label given as input (*ONLY_OBJ*), and the object label and a random embedding given as input (*OBJ + RAND_EMB*) and only EEG embeddings given as input (*ONLY_EEG*) and our solutions without Stage 2 (*NO_STAGE2*), and the complete solution with all stages (*ALL*).

Comparison with Stage 2 Omission(NO_STAGE2)

전반적으로 표는 stage 2(EEG 임베딩과 CLIP-based supervision strategy를 사용한 이미지 임베딩을 일치시키는 것)의 통합이 더 높은 품질의 텍스트 생성에 기여한다는 것을 보여줌. BERT Score가 일관되게 높은 것은 stage 2가 의미론적 일관성을 유지하는 데 도움을 준다는 것을 강조하며, 이는 fluency, adequacy 개선과 일치함.

Generation performance without object labels in the input

원래 초기 가정은 EEG 임베딩은 noisy하고 multichannel이기 때문에 복잡한 생각은 완벽히 포착하지 못할 것이라고 생각해서 object label을 추가적으로 넣어줬음. 하지만, ONLY_EEG와 ALL이 비슷한 성과를 냄. 이는 stage 1에서 EEG 임베딩과 vision 임베딩을 정렬하고 stage 2에서 LLM, 특히 projector를 vision 임베딩으로 pretraining한 효과로 보여짐.

Subject-wise Analysis

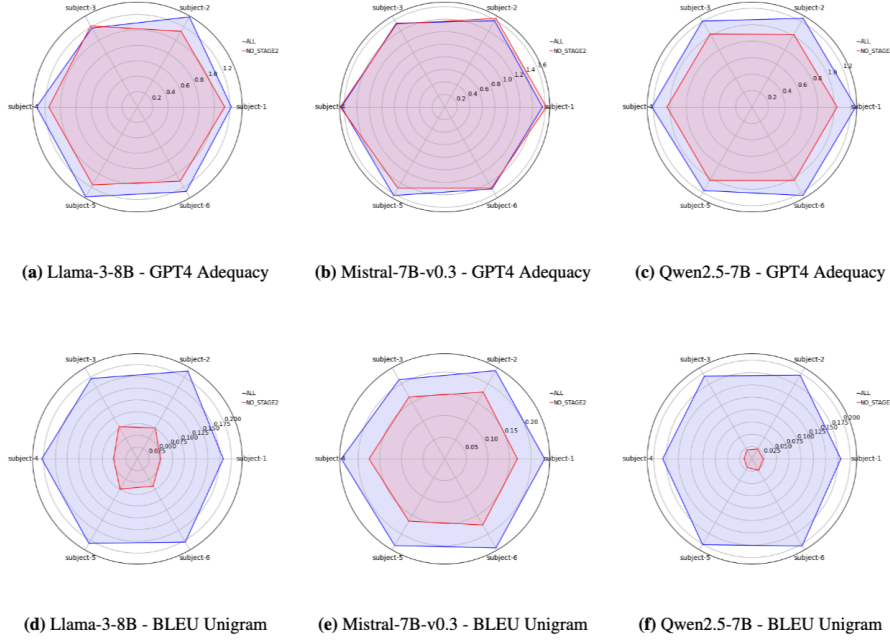


Figure 2: Spider graphs comparing the *ALL* and *NO_STAGE2* variants across subjects for GPT-4 Adequacy and BLEU Unigram metrics, evaluated using different models. For BLEU scores, the *ALL* variants show a noticeable improvement across all six subjects compared to the *NO_STAGE2* variants. Although numerically smaller, a consistent improvement in Adequacy is also observed with the *ALL* variants, which is significant in context of noisy EEG data.

데이터셋이 6가지 주제로 구성되어 있어, 각 개인의 분석을 통해 참가자들 간 접근 방식의 안정성을 평가할 수 있음. 주제 내 실험에서 ALL은 일관되게 우위를 보이며, 이는 stage 2가 포함될 때 적합성 점수에서 상당한 향상을 나타낸다고 볼 수 있음. NO_STAGE2는 EEG 임베딩과 이미지 임베딩 간 필수 정렬 단계를 생략하기 때문에 주제 간 일관되게 낮은 성능을 보여줌. → stage 1과 stage 2 없이 직접 fine-tuning 하는 것은 EEG 데이터에 불충분함.

thought-to-text 변환 모델의 경우, 개인정보보호 환경 내에서 개발되고 배포되어야 함. 즉, 본 논문에서 제안하는 multi-stage 접근법은 이미지와 같은 non-EEG 데이터에서 pretraining을 하고, 소량의 특정 주제 EEG 데이터로 fine-tuning하기 때문에 개인정보보호가 유지되는 개인화된 EEG-LLM 모델 개발의 가능성을 열 수 있음.

Qualitative Inspection and Basic Error Analysis









ID	Original Image Stimuli	Input	Ground Truth Object	Predicted Object (Stage 1)	Ground Truth Description (GPT-4)	Predicted Description (Thought2Text)
1		EEG	mushroom	flower	A large yellow mushroom with a brown stem and a brown cap, surrounded by green foliage.	A group of mushrooms growing on a log.
2		EEG + OBJ	piano	piano	A black grand piano in a living room.	A grand piano with a stool in front of it.
3		EEG	piano	piano	A man in a red coat and black pants is playing a piano in a room with a chandelier.	A man is playing the piano in a dimly lit room.
4		EEG + OBJ	pumpkin	pumpkin	A carved pumpkin with a face and eyes, sitting on a table.	A carved pumpkin with a spooky face on it.
5		EEG + OBJ	flower	mushroom	A black and white photograph of a single daisy with a white center and a dark brown center.	A group of mushrooms growing on a log.
6		EEG + OBJ	coffee mug	coffee mug	A hand holding a mug with a blue background and a handprint design.	A person holding a coffee mug with the words "World's Best Dad" written on it.
7		EEG	guitar	watch or watches	A young boy sitting on a chair playing a guitar.	A man is holding a guitar in front of a microphone.
8		EEG + OBJ	camp or camping	camp or camping	A tent in a mountainous area with trees and fog.	A tent set up in a forest with a campfire nearby.

Table 3: Sample positive (in green) and negative (in red) anecdotal examples using the MISTRAL-7B-v0.3 *ALL* and *EEG_ONLY* variants that take different inputs: EEG signals + object information and EEG signals alone, respectively.

GPT-4와 Mistral-7B-v0.3 모델(ALL과 ONLY_EEG)이 생성한 이미지 설명을 비교함. 거의 모든 경우 EEG+predicted OBJ를 입력으로 한 방식은 매우 정확한 설명을 생성한 반면, 아예 객체 분류가 잘못된 경우도 있음. 이러한 잘못된 식별은 객체 분류의 개선 가능성을 나타냄.

EEG만 입력으로 넣었을 때 주요 장점은 예측된 object label이 잘못된 경우에도 비교적 일관된 설명을 생성할 수 있다는 점임. 이는 EEG 임베딩이 언어 모델의 생성을 안내하는 데 있어 우수한 견고성을 가지고 있다는 뜻임.

Conclusion and Future Work

이 논문에서 EEG 신호를 텍스트로 변환하는 새로운 접근 방식을 소개하였으며, EEG 데이터로 fine-tuning된 instruction-tuned LLM을 활용함. 소개된 방식은 크게 세 단계로 진행됨: (1) 특징 추출을 위한 EEG 인코더 training, (2) multimodal 데이터에 대한 LLM fine-tuning, (3) neural signal에서 직접 텍스트 생성을 위한 EEG 임베딩으로의 추가 정제.

본 논문의 방법론은 교차 주제 및 주제 내 분석 모두에서 양적인 평가를 통해 검증됨. 질적 평가는 EEG 신호와 object label 또는 EEG 신호만을 텍스트 생성 입력으로 사용하는 다양한 시나리오에 대한 통찰력을 추가로 제공함. 이러한 평가는 효율적인 텍스트 생성을 위한 본 방법론의 효과를 강화할 뿐만 아니라 원하는 결과를 달성하기 위해 EEG 데이터만을 활용할 수 있는 잠재력을 강조함. 향후 연구는 모델 아키텍처 최적화, 다양한 데이터셋에서의 foundational pretrained EEG 모델 활용, 대규모 이미지 데이터셋과 다양한 작업에 대한 stage 2 training을 통해 EEG-text 정렬 개선, 의료 및 보조 기술에서 실용적인 응용 프로그램 탐색에 초점을 맞출 것이며, thoughts-to-text 시스템의 접근성을 높일 것임.

Limitations

EEG 신호에서 세밀한 정보를 추출하는 것은 높은 noise 비율과 낮은 공간 해상도 때문에 어려움이 있음. 이러한 어려움에도 불구하고 EEG 신호는 객체 범주를 식별할 수 있으며, 이를 일반 프롬프트와 결합하여 텍스트 생성에 도움을 줄 수 있음.

그러나 유사한 형태 객체의 잘못된 분류는 꽃을 버섯으로 착각하는 경우와 같이 객체 인식에서의 잠재적 모호성을 강조함. 이를 해결하는 한 가지 방법은 각 주제에 대한 개인화된 모델을 훈련하고 그들의 성능을 평가하는 것임. 또한, 주제 간 예측의 일반화를 향상시키는 방법을 탐색할 수 있음.

또 다른 도전 과제는 데이터 부족임. 통제된 실험 조건에서 더 많은 고품질 다채널 EEG 데이터를 수집하는 것이 중요하며, 이는 noise를 줄이는 데 필요함.

생각을 읽는 것은 의사소통 능력이 제한된 개인에게 유익할 수 있으나, 동의 없이 생각을 침해할 수 있는 BCI의 잠재적 남용 위험이 있음. 그러나 적절한 조치와 규제가 있다면 이점이 한계보다 큰 기술임.