# EEG-GPT: Exploring Capabilities of Large Language Models for EEG Classification and Interpretation

| | |
|---|---|
| 상태 | 완료 |
| Journal | arXiv |
| Year | 2024.02 |
| Summary | Tree of Thoughts 및 few- /zero-shot을 통해 EEG를 체계적으로 활용하여 분류 및 해석을 수행하는 EEG-GPT 제안 |
| Limitations | 임상 EEG feature 적용해 LLM 성능 최적화 필요, 희귀질환 적용 가능성 탐색, LLM의 논리적 추론 평가 및 환각 검출 연구 필수 |
| Link | https://arxiv.org/abs/2401.18006 |
| category | Classification EEG LLM ML Prompt-tuning |

EEG-GPT는 LLM을 활용하여 EEG 분류와 해석을 수행하는 새로운 접근법을 제시함. few-shot learning에서 기존 ML 방법보다 우수한 성능(AUROC 0.86)을 보이며, Tree of Thoughts 추론을 통해 EEG 도구를 체계적으로 활용함. 그러나 임상 적용을 위해서는 LLM의 환각 문제 해결과 신뢰성 검증이 필요함.

## 키워드

- Large Language Model
- EEG
- Machine Learning

## Introduction

LLM은 few- and zero-shot learning 능력이 있고, 이는 EEG 데이터셋과 같이 작은 데이터 환경에 적합한 특징임. 최근의 연구는 LLM이 암 약물 시너지 예측부터 심장 신호 분석까지 다양한 분야에서 few-shot learning을 적용하고 있다고 함. 또한 transformer 아키텍처는 zero-shot 태스크에서 in-context learning이 됨이 증명됨. 따라서, 프롬프트에 제공된 정보를 활용하여 다양한 태스크에서 더 나은 성과를 낼 수 있음.

Chain of Thought와 같은 프롬프트 방식을 통해 다단계 연산을 수행할 수 있음. 더 최근에는 Tree of Thoughts 프레임워크가 개발되었는데, 이는 Chain of Thought의 확장판으로 LLM이 탐색할 수 있는 decision tree를 구성하는 방식임. 즉, LLM은 다양한 방식을 활용해 인간을 대신하는 'experts' 효과를 낼 수 있음.

EEG에 LLM을 적용한 연구는 현재 시작되고 있는 단계임. NeuroGPT의 경우 EEG 신호의 임베딩을 생성하고 예측하는 데 사용되는 LLM 기반 프레임워크임. 다른 연구에서는 BENDR 프레임워크가 있는데, 이는 EEG 기반 수면 단계 분류 작업을 위해 transformer 아키텍처를 사용함.

인간의 EEG 분류 및 해석은 매우 주관적이고 현재 SOTA EEG 해석 시스템인 Persyst에 대한 최근 평가에서는 신뢰할 수 없는 성능이 드러남. 또한, 임상 환경에서 딥러닝 EEG 시스템은 거짓 비율이 높음 [link]. 따라서, 본 논문은 LLM이 EEG 분류 및 해석 작업에서 임상의에게 도움을 줄 수 있는 가능성을 평가하는 것을 목표로 함. LLM 기반 접근 방식이 성능과 투명성 측면에서 현재 딥러닝 기반 EEG 해석 및 분류 방법에 비해 이점을 제공할 수 있는지를 조사함.

## Methods

LLM의 임상 EEG 작업, 특히 EEG를 정상 또는 비정상으로 분류하는 작업에 대한 능력을 탐구하기 위해 두 가지 접근 방식을 사용함. 모두 Temple University Hospital Abnormal Corpus [link] 를 활용함. 2,993명의 피험자에게서 수집된 총 1,140 시간의 EEG 데이터로, 정상 및 비정상 기록이 대략 균형 잡혀 있고, 평가의 일관성을 위해 train과 evaluation sets이 미리 나눠짐.

## Few- and zero-shot Learning

가설 1 : 상대적으로 적은 양의 train 데이터가 주어질 때, fine-tuned LLM(EEG-GPT)이 EEG 정상/비정상 분류 태스크에서 다른 ML 방법에 비해 높은 성능을 발휘할 것이다 !

가설 2 : zero-shot learning을 적용한 base EEG LLM이 정상/비정상 분류 태스크에서 우연히 맞출 확률보다 높은 성능을 발휘할 것이다 !

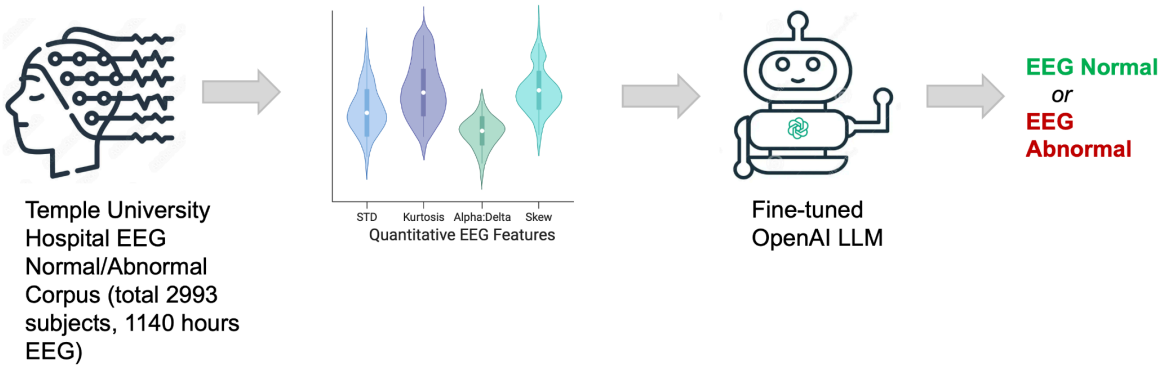


Figure 2: Pipeline for few-shot experiment

1. **feature selection and fine-tuning**
    a. **feature selection** : EEG 파일을 겹치지 않는 20초의 epochs로 나누고, 각 에폭에 대해 채널별 특징을 계산함. 각 채널에 대해 동일한 특징 집합이 계산되어 각 epochs에 대해 30개의 features(6-features x 5-channels)를 갖는 샘플을 생성함. 각 feature 샘플은 해당 부모 파일이 정상 또는 비정상으로 레이블링된 것에 따라 정상 또는 비정상으로 레이블링됨.

| Features calculated (over 20-second epoch) |
|---|
| 90th percentile of voltage amplitudes |
| standard deviation |
| kurtosis |
| alpha:delta power ratio |
| theta:alpha power ratio |
| delta:theta power ratio |

Table 1: Features calculated per channel

| Channels used for feature calculation |
|---|
| Cz |
| T5 |
| T6 |
| O1 |
| O2 |

Table 2: Channels used for feature calculation

    b. **fine-tuning** : 뽑은 feature를 언어적 표현으로 변환하고, 이 언어적 표현을 사용하여 OpenAI의 text-davinci-003 모델을 fine-tuning함. fine-tuning할 때, feature의 언어적 표현을 프롬프트로 하고, 정상/비정상 레이블을 완성하는 prompt-completion 쌍 샘플을 제공함. 즉, 주어진 파일이 정상인지 비정상인지에 대한 예측을 학습함.

```
{"prompt":
    " Quantitative EEG: In a 20 second period,
        at channel Cz:[
            the 90th percentile of voltage amplitudes = 0.35 microvolts,
            standard deviation = 0.27,
            kurtosis = 0.33,
            alpha:delta power ratio = 0.22,
            theta:alpha power ratio = 9.11,
            delta:theta power ratio = 0.49];
        at channel T5:[
            the 90th percentile of voltage amplitudes = 0.10 microvolts,
```

```
                standard deviation = 0.09,
                kurtosis = 1.53,
                alpha:delta power ratio = 0.13,
                theta:alpha power ratio = 2.43,
                delta:theta power ratio = 3.13];
            at channel T6:[
                the 90th percentile of voltage amplitudes = 0.18 microvolts,
                standard deviation = 0.20,
                kurtosis = 9.47,
                alpha:delta power ratio = 0.09,
                theta:alpha power ratio = 3.80,
                delta:theta power ratio = 2.95];
            at channel O1:[
                the 90th percentile of voltage amplitudes = 0.34 microvolts,
                standard deviation = 0.28,
                kurtosis = 0.11,
                alpha:delta power ratio = 0.82,
                theta:alpha power ratio = 1.33,
                delta:theta power ratio = 0.92];
            at channel O2:[
                the 90th percentile of voltage amplitudes = 0.30 microvolts,
                standard deviation = 0.26,
                kurtosis = 2.47,
                alpha:delta power ratio = 0.39,
                theta:alpha power ratio = 1.35,
                delta:theta power ratio = 1.89];.

            Cumulative Effect Category:",

    "completion":" normal"}
```

Figure 3: Example prompt-completion pair (formatted for readability)

2. **Few- and zero-shot learning** : EEG-GPT가 어떤 fine-tuning도 없이 few- and zero-shot에서는 분류 작업이 어떻게 수행되는지 평가함.

## Evaluation of reasoning capability in EEG tool usage

가설 3 : LLM 기반 프레임워크가 EEG 정상/비정상 분류 태스크에서 전문 소프트웨어 툴을 효과적으로 활용할 수 있을 것이다 !

이 프레임워크를 개발하기 위해 임상 뇌전증 전문의가 주어진 EEG 파일의 비정상 여부를 분석하는 방식을 고려함. 이 과정을 decision tree로 개념화함.

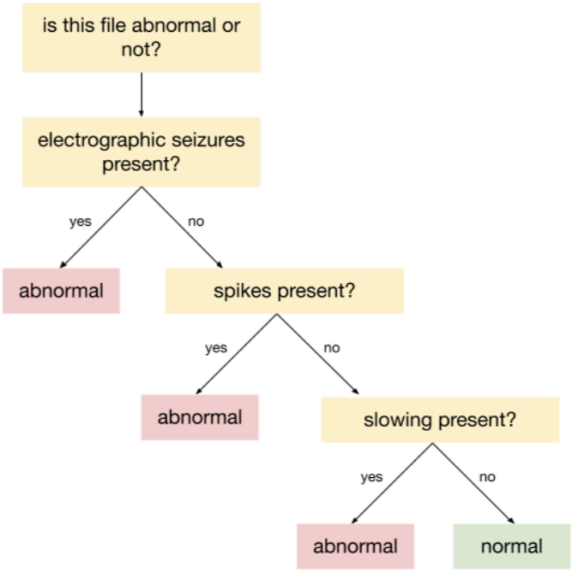


Figure 4: Simplified diagram of clinical epileptologist workflow

1. **Tree of Thought** [link, link] : LLM이 복잡한 문제를 해결할 때, decision-tree style 탐색 공간을 서칭하고 되돌아갈 수 있도록 함.

2. **Specialized software tools** : 특정 EEG 파일을 평가하기 위해 프레임워크는 decision tree 공간을 탐색하고, 아래 언급되는 3개의 모델이 각 분기점에서 모델을 탐색하여 계속 진행할지 여부를 결정함.
    a. Automated seizure detection model
        i. Temple University Hospital seizure corpus를 사용하여 CNN을 활용한 EEG 기반 automated seizure detection model을 훈련함. 10초 간격의 데이터를 입력받아 주어진 epoch에 seizure 활동이 포함되어 있는지 여부를 반환함.
    b. Automated spike detection model
        i. Esteller et al. [link]의 연구를 참고하여 EEG spike에 대한 계산의 biomarker로 선의 길이를 탐지함.
    c. qEEG feature comparison tool
        i. qEEG feature의 하위 집합에 대한 연령 기반 규범 범위를 계산함. 특정 EEG 파일이 주어지면, qEEG feature 집합을 계산하고 연령에 맞는 규범에 대한 코사인 유사도 점수를 계산함. 점수가 특정 입계값보다 높으면 “similar to a reference of normal files” 결과를 반환하고, 그 반대의 경우에도 마찬가지 결과를 반환함.

# Results

## EEG-GPT demonstrates few- and zero-shot learning proficiency

EEG-GPT를 few- 및 zero-shot 맥락에서 전통적인 ML과 최근 DL 기반 접근법과 비교 평가함. ML 비교를 위해, 각 접근법 별 동일한 training features의 세트를 활용함. DL은 해당 연구의 논문에서 보고된 성능 지표를 활용함. 왜냐하면, 각 논문의 아키텍처가 EEG-GPT가 사용하는 제한된 feature 대신 raw EEG 입력을 사용하기 때문임.


Figure 5: Resulting AUROCs of normal/abnormal classification task, plotted against proportion of training data used to fit model

사용 가능한 데이터의 2%로 훈련했을 때, EEG-GPT는 0.86의 AUROC 달성함. 전통적인 ML은 훈련 데이터 양이 증가할수록 성능이 향상되지만, 대략 0.8%에서 정체됨. EEG-GPT는 모든 데이터를 학습한 DL과 성능이 일치하고, 더 최근 접근법에는 뒤쳐짐. zero-shot context에서 EEG-GPT 성능은 우연보다는 나은 상태이며, in-context learning이 제공될 때 0.63으로 향상됨.

## EEG-GPT navigates EEG tool usage with tree-of-thought reasoning

tree-of-thought의 두 가지 예시를 제시함. 첫 번째 데모에서, seizure가 포함된 것으로 알려진 EEG 파일을 프레임워크에 전달함. 문제 해결 절차를 진행하면서 automated seizure detector가 즉시 발작을 감지하여 해당 파일이 비정상임을 나타내고 문제 해결 절차의 조기 중지를 가능하게 함을 알 수 있음.

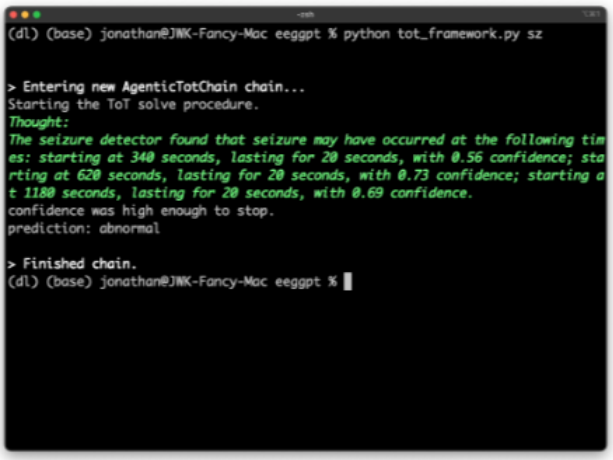


Figure 6: Framework's analysis of EEG file known to contain seizure

두 번째 데모는 발작이 없지만 비정상으로 알려진 EEG 파일을 프레임워크에 전달함. 해결 절차를 진행하면서 seizure detector가 어떠한 발작도 발견하지 못하는 것을 확인함. → qEEG feature comparison tool 사용함. → 프레임워크는 해당 파일이 정상 EEG의 기준 집합과 유사하다고 판단하며, qEEG 도구 자체에서 제공한 신뢰도 측정값은 0.54임. → 낮은 신뢰도에 따라 프레임워크는 이 파일을 분류하기 위해 더 많은 정보가 필요하다고 결정하고 automated spike detection tool을 사용함. → 이 도구는 실제로 스파이크를 감지하며 이 파일이 비정상임을 나타냄. 이 시점에서 프레임워크는 충분한 정보와 신뢰도가 확보되었다고 판단하고 최종 분류를 진행하며 실행을 중단함.

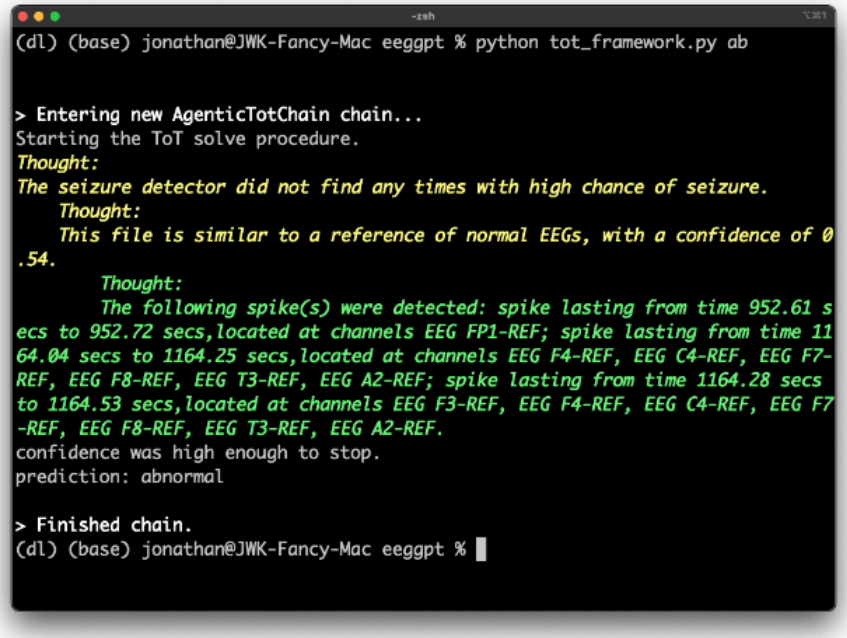


Figure 7: Framework's analysis of abnormal EEG file known to be seizure-free

## Discussion & Future Work

1. **EEG-GPT 성능**
   - EEG-GPT는 EEG 해석 및 분류 작업에서 기존 머신러닝 방법을 능가하는 성과를 보임.
   - 적은 양의 훈련 데이터에서도 기존 딥러닝 모델과 유사한 성능을 발휘.
   - EEG-GPT의 few-shot 및 zero-shot 능력은 대규모 데이터셋에 포함된 정상 EEG 데이터 덕분일 가능성이 있음.
2. **임상 적용의 어려움과 해결책**

- 기존 머신러닝 모델은 "블랙박스" 문제로 인해 임상 적용이 어려움.
- LLM은 논리적 사고 흐름을 구현할 수 있지만, "환각(hallucination)" 정보 생성 문제 존재.
- 이를 해결하기 위해 인간의 개입 및 단계별 검증(step-wise verifiability)이 필요.

3. **미래 연구 방향**
   - 임상 EEG 특성 공간을 탐색하여 LLM 성능 최적화 연구 필요.
   - LLM이 적은 훈련 데이터에서도 효과적이므로 희귀질환 적용 가능성 탐색.
   - LLM의 논리적 추론 평가 및 "환각" 오류 검출 연구 필수.
   - EEG 사례 자동 인식 기능 추가 및 특화 소프트웨어 도입 필요.
4. **LLM의 EEG 해석 잠재력**
   - EEG 해석과 분류에서 유망한 접근 방식임.
   - 그러나 환각 문제를 해결하기 위한 체계적 검증이 반드시 필요.
   - 임상 적용을 위해 높은 신뢰성과 해석 가능성을 보장하는 연구가 지속되어야 함.