Thought2Text: Text Generation from EEG Signal using Large Language Models (LLMs)

Abhijit Mishra*, **Shreya Shukla***, **Jose Torres**, **Jacek Gwizdka**, **Shounak Roychowdhury**
School of Information, University of Texas at Austin
{abhijitmishra, shreya.shukla, jtorres1221, jacekg, shounak.roychowdhury}@utexas.edu

Abstract

Decoding and expressing brain activity in a comprehensible form is a challenging frontier in AI. This paper presents *Thought2Text*, which uses instruction-tuned Large Language Models (LLMs) fine-tuned with EEG data to achieve this goal. The approach involves three stages: (1) training an EEG encoder for visual feature extraction, (2) fine-tuning LLMs on image and text data, enabling multimodal description generation, and (3) further fine-tuning on EEG embeddings to generate text directly from EEG during inference. Experiments on a public EEG dataset collected for six subjects with image stimuli and text captions demonstrate the efficacy of multimodal LLMs (LLaMA-v3, Mistral-v0.3, Qwen2.5), validated using traditional language generation evaluation metrics, as well as *fluency* and *adequacy* measures. This approach marks a significant advancement towards portable, low-cost "thoughts-to-text" technology with potential applications in both neuroscience and natural language processing.

1 Introduction

Brain-Computer Interface (BCI) systems, combined with portable and wearable noninvasive *Electroencephalographic* (EEG) devices, enable direct interfacing between the brain and external devices (He et al., 2015). Advances in generating images and natural language using EEG (Speier et al., 2016; Benchetrit et al., 2023; Défossez et al., 2023) hold promise for developing BCIs in various domains, including assistive communication (*e.g.,* for ALS and stroke patients), mixed reality (AR/VR) experience enhancement, mental health diagnosis, and gaming. Recent strides in NLP driven by powerful Large Language Models (LLMs) such as OpenAI GPT-4. (Achiam et al., 2023), Google Gemini (Team et al., 2023), Meta-LLaMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), and Microsoft Phi (Gunasekar et al., 2023) have enabled multimodal integration, facilitating language generation from images (Liu et al., 2024b) and speech (Fathullah et al., 2024). Our research focuses on a multimodal solution to decode brain signals directly into text, using EEG signals. We choose EEG because of its affordability compared to mainstream alternatives such as Functional Magnetic Resonance Imaging (fMRI) (Tang et al., 2023), which are costlier and require complex setup. For generating text, we leverage LLMs, which enable flexible, high-quality text generation across various modalities, such as images (Liu et al., 2024b), audio (Rubenstein et al., 2023; Fathullah et al., 2024), and more and often outperform current open vocabulary task-specific methods (Shi et al., 2024).

Our approach for generating textual descriptions from EEG signals involves three key steps: (a) capturing language-agnostic EEG signals via visual stimuli, (b) encoding these signals into embeddings using a deep multichannel neural encoder, and (c) fine-tuning language models by projecting image and EEG embeddings into a token embedding space to generate responses. These responses are compared with gold standard image descriptions to compute the training loss. During inference, only EEG signals and a generic textual prompt are used as inputs to the LLMs to generate responses. Our method requires images, EEG data and descriptions for training, while inference is bimodal, using only EEG to generate text.

For experiments, we use a public 128-channel EEG dataset from six participants viewing visual stimuli. The image descriptions are generated by GPT-4-Omni (Achiam et al., 2023) and quality-checked by human annotators, providing the text modality necessary to build EEG-to-text generation systems. Although the goal is the generation of text from EEG, we use a dataset with visual stimuli for their language-agnostic nature, avoiding the potential complexities associated with

*Equal Contribution

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 3747–3759

reading and language processing (see Section 3 for details). We fine-tune large language models (LLMs) using these descriptions, leveraging pre-trained instruction-based language models such as MISTRAL-V3 (Jiang et al., 2023), LLAMA-V3 (Touvron et al., 2023), and QWEN2.5 (Bai et al., 2023a). Evaluation with standard generation metrics (Sharma et al., 2017) and GPT-4-based assessments confirmed the effectiveness of our approach.

Our paper's key contributions include:

- Integration of brain signals with instruction-tuned LLMs.
- Fine-tuning models on EEG signals captured for visual stimuli, leveraging its language-agnostic nature to enhance LLM interaction.
- Validation of the efficacy of the model in a popular public dataset that contains EEG signals captured using affordable devices.

The code and a link to the processed dataset can be found at `https://github.com/abhijitmishra/Thought2Text`.

2 Related Work

Integrating behavioral signals such as eye movement and brain signals into NLP and computer vision tasks (Mishra and Bhattacharyya, 2018; Sharma and Meena, 2024) has seen significant progress. Key datasets include *ZuCo 2.0* (Hollenstein et al., 2019), which captures EEG and eye-gaze during natural language reading, and *MOABB* (Jayaram and Barachant, 2018), offering over 120,000 EEG samples from 400+ subjects from various BCI tasks such as motor imagery, visual evoked potentials, and cognitive load. Datasets such as *MindBigData* (Vivancos and Cuesta, 2022) and *CVPR2017* (Spampinato et al., 2017) provide substantial EEG data collected from participants' responses to handwritten and open vocabulary object-based image stimuli respectively. These datasets have facilitated research on the classification of EEG data (Spampinato et al., 2017; Palazzo et al., 2020; Khaleghi et al., 2023) and the generation of images from EEG signals using GANs (Goodfellow et al., 2020) and latent diffusion models (Rombach et al., 2022; Bai et al., 2023b; Lan et al., 2023; Tirupattur et al., 2018).

Additionally, multimodal datasets like *The Alice Dataset* (Bhattasali et al., 2020), which includes EEG and fMRI recordings from participants listening to a story, provide two measurable modalities: audio stimulus and the corresponding text. Another recent multimodal dataset, *EIT-1M* (Zheng et al., 2024), contains one million EEG-Image-Text pairs, collected as participants viewed visual-textual stimuli. At the time of writing, a partial version of the *EIT-1M* dataset was released, containing data for only one subject.

Generating language from EEG signals remains an elusive challenge. The most closely related work is *EEG2TEXT* (Liu et al., 2024a), which utilizes EEG pretraining and a multi-view transformer to decode EEG signals into text. Another approach similar to ours, using multiple modalities, is presented in (Ikegawa et al., 2024), where intracranial EEG (iEEG) signals were recorded from patients watching videos, and each video frame was used to generate images and text using CLIP vision-language model (Radford et al., 2021). Unlike our approach, this study involved implanting electrodes in patients. Furthermore, these works do not leverage large language models (LLMs) with prompt engineering for generating prompt-specific responses. We believe our method of fine-tuning LLMs using non-invasive EEG input is the first of its kind.

3 Dataset and the need for Visual Stimuli

Building a system that generates text from neural activity naturally requires a dataset of paired $< eeg,\ text >$ examples. However, using textual stimuli presents inherent challenges. Reading is a learned skill that requires the decoding of symbols into sounds and meanings, syntactic parsing, and sequential integration in time. In contrast, visual perception is more innate, natural, and image processing is more parallel (Dehaene, 2009; Townsend, 1990). Furthermore, using EEG data collected on textual stimuli introduces additional complexities of language processing, such as determining brain activity windows for specific words, managing retention of word context post-onset, and managing the overlap of contexts when words are shown in different time frames (Wehbe et al., 2014; Murphy et al., 2022). Additionally, vocabulary size presents a challenge: while EEG-to-text systems perform well in closed-vocabulary settings, open-vocabulary decoding becomes inefficient as vocabulary size increases (Martin et al., 2018; Wang and Ji, 2022; Liu et al., 2024a). The core challenge

Annotator	Percentage of Correct Captions
Annotator 1	98%
Annotator 2	96%
Agreement Score	93%

Table 1: GPT-4 Captions Validation Results from Amazon Mechanical Turk

in thought-to-text systems thus lies in two key aspects: *(a) harnessing language-agnostic neural signals with minimal interference from linguistic processing* and *(b) using these signals to generate text in a target language*, potentially by specifying an instruction or prompt.

This leads us to experiment with datasets where EEG signals are recorded in response to visual stimuli rather than textual input. For *(a)*, visual stimuli provide several advantages: By relying on images, it circumvents the complexities of language processing and also elicits brain responses to salient image features, making them more suited to capture neural activity in a language-agnostic manner. For *(b)*, generating text from visual stimuli-evoked EEG remains challenging, as most EEG datasets collected with visual stimuli (e.g., *CVPR2017* dataset (Spampinato et al., 2017), *MOABB* (Jayaram and Barachant, 2018), *MindBigData* (Vivancos and Cuesta, 2022)) lack an associated text component, which is essential for evaluating whether the generated text from EEG signals accurately captures the perceived salient features of the images. To address this, robust image captioning tools such as GPT-4 Omni (Achiam et al., 2023) can be employed to generate captions that effectively capture the salient imformation of visual stimuli, thus creating a tri-modal dataset with $< eeg,\ text,\ image >$ tuples. Text generation can further be guided by specifying a prompt or instruction to tailor the output to a particular language or context. This approach enriches the dataset and provides a more comprehensive foundation for thought-to-text research.

For our experiments, we utilize the *CVPR2017* dataset (Spampinato et al., 2017), which contains preprocessed EEG data from six participants, each viewing 50 images across 40 diverse object categories[1] (such as vehicles, musical instruments, *etc*). Each EEG recording, corresponding to one participant and one visual stimulus, consists of 128 channels recorded for 0.5 seconds at a sampling rate of 1 kHz. Data are represented as a 128×N matrix, where N, approximately 500, represents the number of samples per channel in each segment. According to Spampinato et al. (2017), the EEG signals were pre-processed by first applying a second-order Butterworth bandpass filter between 5 Hz and 95 Hz and a notch filter at 50 Hz to remove power line noise. In addition, since the exact duration of the EEG signals can vary, the first 20 samples (20 ms) were discarded to reduce interference from the previous image, and then standardizing the signal to a common length of 440 samples, accounting for segments with N < 500. The dataset provides pre-filtered signals across three frequency ranges: 14-70Hz, 5-95Hz, and 55-95Hz. In line with previous research (Palazzo et al., 2020), we selected the 55-95 Hz range, as it has shown the most reliable results. We used the training, validation, and test split from the original paper for overall and subject-wise analysis.

Since the *CVPR2017* dataset lacks textual descriptions, we used GPT-4 to generate brief one-line captions for each image as they are efficient in capturing salient information without introducing extraneous details. Here, the aim is to map EEG data to concise captions to simplify the alignment task, making it feasible despite the noise and variability inherent in EEG signals. To ensure quality, we validated these auto-generated captions using human annotators from *Amazon Mechanical Turk*, who rated their fluency and adequacy on a binary scale. Table 1 shows the validation results, including annotator agreement on acceptable captions, highlighting the reliability of the generated captions. We release the descriptions under the same license as the original data.

4 Method

Our approach uses a three-stage process to generate coherent text from EEG signals. The overall workflow is illustrated in Figure 1.

4.1 Stage1: Training EEG Encoder for Embedding Extraction

The first stage of our approach focuses on developing an encoder that extracts meaningful embeddings (H_{eeg}) from multi-channel EEG signals. Since the target thoughts are short and pertain to the most salient features of an image, we design the encoder with two objectives: (a) aligning the

[1] we use the terms *object* and *class* interchangeably to refer to an object category.

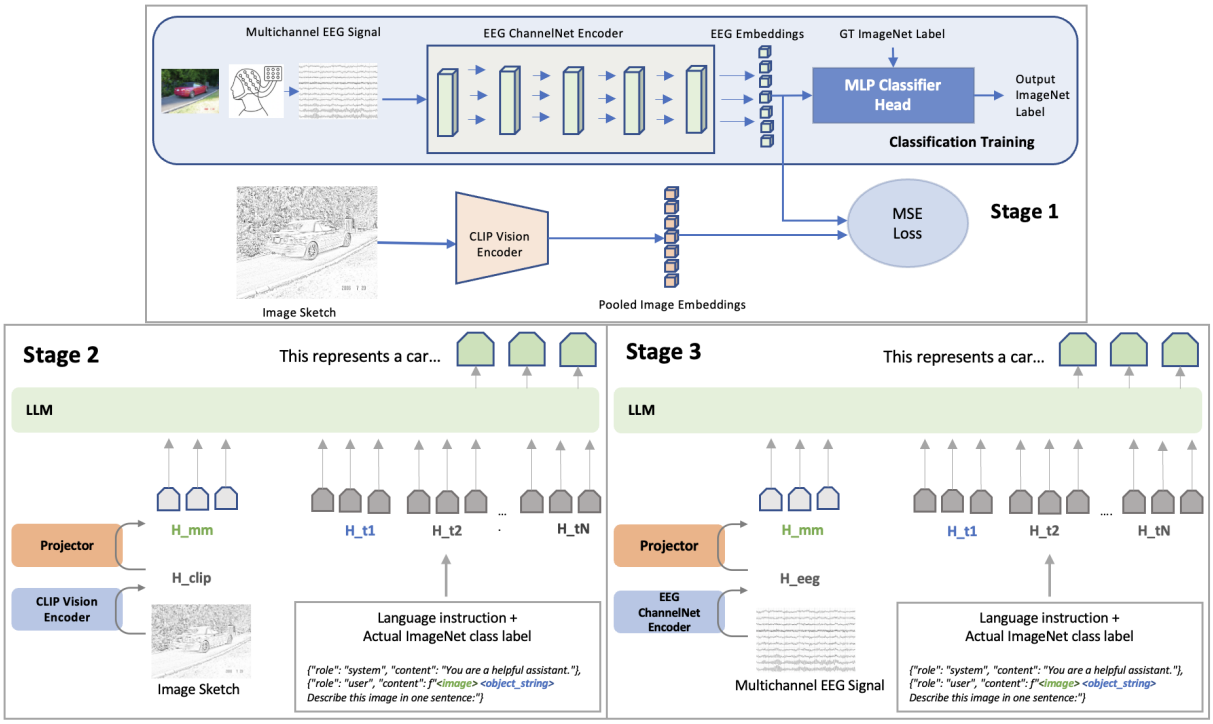


Figure 1: Multi-stage training process for *Thought2Text*. ***Stage 1:*** EEG ChannelNet encoder is trained using MSE loss (aligning EEG embeddings with CLIP model image embeddings) and CE loss (for EEG classification). ***Stage 2:*** LLMs are fine-tuned using image descriptions and object labels, with only the projector trained while the LLM and CLIP encoder remain frozen. ***Stage 3:*** Similar to stage 2, but the projector is trained with EEG embeddings. LLM and EEG encoder remain frozen.

EEG embeddings with those derived from image stimuli using a pretrained visual encoder, and (b) predicting the most salient object (*e.g.,* piano) from the EEG embeddings.

As shown in Figure 1, we use a multichannel EEG encoder inspired by *ChannelNet* (Palazzo et al., 2020), a deep convolutional neural network model, to convert EEG signals into multidimensional embeddings (H_{eeg}). These embeddings are further processed by an MLP classifier to predict object labels (y_{obj}), which correspond to objects present in the image stimuli. The label-set matches ImageNet's labels, available as part of the dataset, representing the most salient object in each image.

The model is trained by minimizing two losses: (A) a categorical cross-entropy loss (CE) between the predicted and ground-truth object labels, and (B) a mean squared error (MSE) between the EEG embeddings (H_{eeg}) and pooled image embeddings (H_{clip}) from a pretrained CLIP model (Radford et al., 2021), which captures semantically rich image representations. CLIP's embeddings offer robust transfer learning capabilities, making them ideal for aligning with EEG data. Performance metrics for EEG-to-image classification using trained MLP Classifier

To better align EEG embeddings with visual stimuli, we simplify images by removing non-central details like color, converting them into sketches using techniques such as *Gaussian Blur* and *Canny* filters. Although this step is empirical and optional, using sketches help focus the alignment on the core features of the object as discussed in Fine-Grained Sketch-Based Image Retrieval (Luo et al., 2023) and Interactive Sketch Question Answering (Lei et al., 2024).

The general loss function ($\mathcal{L}$) balances MSE and CE, weighted by a hyper-parameter α (set to 0.5):

$$\mathcal{L} = (1-\alpha)\cdot \text{MSE}(\mathbf{H}_{eeg}, \mathbf{H}_{clip}) + \alpha \cdot \text{CE}(\mathbf{y}_{obj}, \hat{\mathbf{y}}_{obj}) \quad (1)$$

We predict both H_{eeg} and y_{obj} for three reasons: (a) aligning EEG embeddings with image embeddings allows us to leverage multimodal vision-language models later, adapting pretrained models for EEG-based text generation, (b) joint optimization ensures the embeddings emphasize salient objects in the images, and (c) object labels, combined

with EEG embeddings, can later be fed into multimodal language models to guide more accurate generation. Given the noisy nature of EEG signals, including object labels in the prompts helps keep the model grounded in the salient object and reduces the likelihood of hallucinations.

4.2 Stage2: Priming LLMs with Image Embeddings

To enable LLMs to process multimodal inputs, such as EEG and visual embeddings, we designed a projector inspired by recent advancements in vision-language models (Zhang et al., 2024) and is a fully connected feed-forward layer. Since LLMs are inherently text-based and cannot natively accept non-text embeddings, the projector transforms embeddings from the vision and EEG models into the token embedding space of the LLM. This ensures that the projected embeddings have the same dimensionality as the LLM's token embeddings.

The projector is a simple feed-forward layer that maps the embeddings to the LLM token space. These transformed embeddings are then concatenated with token embeddings extracted from the input prompt, allowing the LLM to process both text and external modality embeddings seamlessly. In our main setting, the input prompt is structured as follows:

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

Input prompt

In stage 2, we first embed the tokens from the extended prompt, including the object labels, using the LLM's token embedding layer. For this, text tokens are input as token IDs that were converted to dense vectors via a lookup in the LLM's *embed_tokens* layer. The token embeddings, $H_{t_1}, H_{t_2}, ..., H_{t_N}$, are then augmented with multimodal embeddings. Specifically, the embedding for the $< image >$ token is replaced by the projected multimodal embedding H_{mm}, and the token-embeddings for the $< object\_string >$ is replaced by the embeddings of the ground-truth object label. The multimodal embedding H_{mm} is computed by projecting pooled embeddings from CLIP, H_{clip}, into the LLM's token embedding space using the following transformation:

$$\mathbf{H}_{mm} = \mathbf{W}_{mm} \cdot \mathbf{H}_{clip} + \mathbf{b}_{mm} \quad (2)$$

Here, $\mathbf{W}_{mm}$ and $\mathbf{b}_{mm}$ represent the projector parameters. Once H_{mm} is computed, it is prepended to the token embeddings from the input prompt, enabling the LLM to process multimodal information.

During training, labels are created by right-shifting the tokens in the prompt, aligning them with the ground-truth image descriptions. Special tokens such as beginning of sentences (BoS), end of sentences (EoS) padding tokens (PAD), system and user and assistant message indicator tokens are used considering LLM specific tokenizers. The input prompts are also converted into LLM specific chat templates. The LLM is then fine-tuned using a standard cross-entropy loss between the predicted tokens and the actual tokens from the ground-truth descriptions.

4.3 Stage3: Tuning LLMs with EEG Embeddings

This stage closely resembles stage 2, with the distinction that instead of H_{clip}, we utilize H_{eeg}, extracted from the EEG encoder trained in stage 1, to compute multimodal embeddings H_{mm}. During this stage, the projector parameters $\mathbf{W}_{mm}$ and $\mathbf{b}_{mm}$ are further tuned. We would like to highlight that throughout this and the previous stage, only the projector is trained, while the LLM and EEG encoders remain frozen to mitigate parameter instability caused by EEG noise.

4.4 Inference

During inference, the EEG ChannelNet encoder processes EEG signals to generate EEG embeddings H_{eeg}. The multimodal projector, trained in stage 3, transforms H_{eeg} into multimodal embeddings H_{mm}. H_{eeg} is also fed into the MLP Classifier, trained in stage 1, to predict the object label. This predicted label is appended to a generic input prompt, similar to the one mentioned in Section 4.2, and the token embeddings are computed for the combined input. Finally, the token embeddings from the projector and the language prompt are concatenated and fed into the LLM to generate descriptions. Notably, no images are used during inference, making the process strictly bimodal.

5 Experimental Details

In this section we highlight the dataset, model details, and evaluation procedure.

5.1 Dataset

We utilize the open-source *CVPR2017* dataset (Spampinato et al., 2017), licensed for academic research, featuring EEG signals from six subjects viewing 50 images across 40 ImageNet classes (Deng et al., 2009), totaling 2000 images. The data is split into training (7959), evaluation (1994), and test (1987) examples. Additional details can be found under Section 3.

5.2 Model Details

We use *ChannelNet* (Palazzo et al., 2020) for the EEG encoder, modifying the final linear layer to produce 512-dimensional embeddings to match the output of the CLIP vision encoder (*openai/clip-vit-base-patch32*). The EEG encoder is trained with a batch size of 16 for 100 epochs, using the AdamW optimizer and a learning rate of $1e^{-4}$. For fine-tuning the LLM, we use a similar setup: a batch size of 16, training for 5 epochs per stage, and employing gradient accumulation and checkpointing. The learning rate for the LLM fine-tuning is kept at $2e^{-5}$. All implementations are carried out using *PyTorch* and Huggingface's *transformers* library.

We evaluate three LLMs: LLAMA-V3 (*meta-llama/Meta-Llama-3-8B-Instruct*), MISTRAL-V0.3 (*mistralai/Mistral-7B-Instruct-v0.3*), and QWEN2.5-7B (*Qwen/Qwen2.5-7B-Instruct*), selected for their efficiency on consumer-grade GPUs such as the NVIDIA RTX 4060Ti. The multimodal embedding H_{mm} is projected to match the token embedding dimensions required by each LLM. All models are permissively licensed for academic research, and training takes approximately 8 GPU hours per LLM training cycle. During inference, we use a batch size of 1, with generation parameters such as `top_k`, `top_p`, and temperature set to their default values for each LLM.

To demonstrate the effectiveness of our approach, we compare it against the following baselines: (a) ONLY_OBJ: where LLMs generate a description based solely on the predicted object without any additional input (e.g., if the object is "car," the LLM generates a description of the word "car" following the prompt in Section 4.2); (b) ONLY_OBJ + RAND_EMB: where we pass a random embedding alongside the predicted object labels to the LLMs; (c) NO_STAGE2: where the priming step described in Section 4 is skipped; and (d) ONLY_EEG: where only the EEG embeddings from Stage 1 are used as input, ignoring the object labels. Our proposed Thought2Text solution, which incorporates all stages and all inputs, is labeled **ALL** in the experiments.

5.3 Evaluation

We use standard NLG metrics such as BLEU, METEOR, ROUGE (Sharma et al., 2017), and BERTScore (Zhang et al., 2019). Furthermore, we assess the quality of the generation using GPT-4, following (Liu et al., 2024b).GPT-4 measures two aspects: *fluency*, for grammar , and *adequacy*, for accuracy in conveying meaning, both rated on a scale of 1-5, with 5 denoting the highest quality.

6 Results

Table 2 presents a comprehensive comparison of metrics across different models and setups. From these results, it is evident that the complete approach, denoted as **ALL**, consistently outperforms other setups across all evaluation metrics. As expected, chance-based baselines like ONLY_OBJ and OBJ + RAND_EMB exhibit poor performance. With our proposed methodology, the LLaMA-v3-8B_ALL model achieved a **BLEU-N (N=1)** score of 25.5%, Mistral-7B_ALL scored 26%, and Qwen2.5-7B_ALL reached 22.7%—-all significantly higher than their respective chance scores under the chance setups. In particular, some models, mainly the LLaMa variants, show increased sensitivity to random input, leading to further reduction in the scores with chance setups.

We will first discuss the inferences from the results obtained with other alternative setups before delving into a subject-wise analysis.

6.1 Comparison with Stage 2 Omission (NO_STAGE2)

The **ROUGE-N (N=1)** score for LLaMA3-8B's ALL variant is 30.0%, whereas the NO_STAGE2 variant achieves only 26.9%, indicating a significant improvement when Stage 2 is added. The same trend is observed for other metrics like BLEU, METEOR, and BERT Score, and is observed across other models as well. In terms of adequacy as measured by GPT-4, the ALL variant stands out except for Qwen model which shows a tendency to copy

LLM	ROUGE-N N=1	ROUGE-N N=2	ROUGE-L	BLEU-N N=1	BLEU-N N=4	MET-EOR	BERT Score	GPT-4 Flu.	GPT-4 Ade.
LLaMA3-8B$_{\text{ONLY_OBJ}}$	9.8	1.5	8.5	7.3	1.3	12.6	0.84	3.44	1.30
LLaMA3-8B$_{\text{OBJ+RAND_EMB}}$	3.8	0.4	3.3	2.8	0.4	5.9	0.84	4.72	1.08
LLaMa3-8B$_{\text{ONLY_EEG}}$	28.9	7.3	26.2	24.1	5.2	23.7	0.89	4.80	1.49
LLaMA3-8B$_{\text{NO_STAGE2}}$	26.9	6.1	23.9	22.6	4.3	23.7	0.88	4.83	1.41
LLaMA3-8B$_{\text{ALL}}$	**30.0**	**8.1**	**26.6**	**25.5**	**5.5**	**26.3**	**0.89**	**4.82**	**1.58**
Mistral-7B-v0.3$_{\text{ONLY_OBJ}}$	17.6	3.4	14.8	14.5	2.5	23.2	0.86	4.46	1.52
Mistral-7B-v0.3$_{\text{OBJ+RAND_EMB}}$	17.9	3.6	15.1	15.7	2.9	22.8	0.87	4.89	1.55
Mistral-7B-v0.3$_{\text{ONLY_EEG}}$	26.7	5.3	23.5	23.3	4.2	22.0	0.88	4.82	1.25
Mistral-7B-v0.3$_{\text{NO_STAGE2}}$	29.2	7.3	26.5	24.1	5.0	24.0	0.89	4.77	1.60
Mistral-7B-v0.3$_{\text{ALL}}$	**30.6**	**8.8**	**28.0**	**26.0**	**6.1**	**26.2**	**0.89**	**4.79**	**1.65**
Qwen2.5-7B$_{\text{ONLY_OBJ}}$	17.6	2.8	14.5	14.8	2.4	21.0	0.85	3.91	1.47
Qwen2.5-7B$_{\text{OBJ+RAND_EMB}}$	1.7	0.1	1.6	1.3	0.3	6.4	0.84	4.73	1.01
Qwen2.5-7B$_{\text{ONLY_EEG}}$	25.2	3.6	21.5	21.9	3.2	20.2	0.88	4.77	1.10
Qwen2.5-7B$_{\text{NO_STAGE2}}$	24.4	4.1	20.9	20.7	3.3	20.2	0.88	4.66	1.24
Qwen2.5-7B$_{\text{ALL}}$	**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 (*Flu.* is fluency and *Ade.* is adequacy) 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).

the object and produce shorter sentences in the case of ONLY_OBJ which is positively rated by GPT-4. Overall, the table demonstrates that incorporating Stage 2 (as detailed in Section 4) – which aligns EEG embeddings with image embeddings using a CLIP-based supervision strategy – contributes to higher-quality text generation.

6.2 Generation performance without object labels in the input

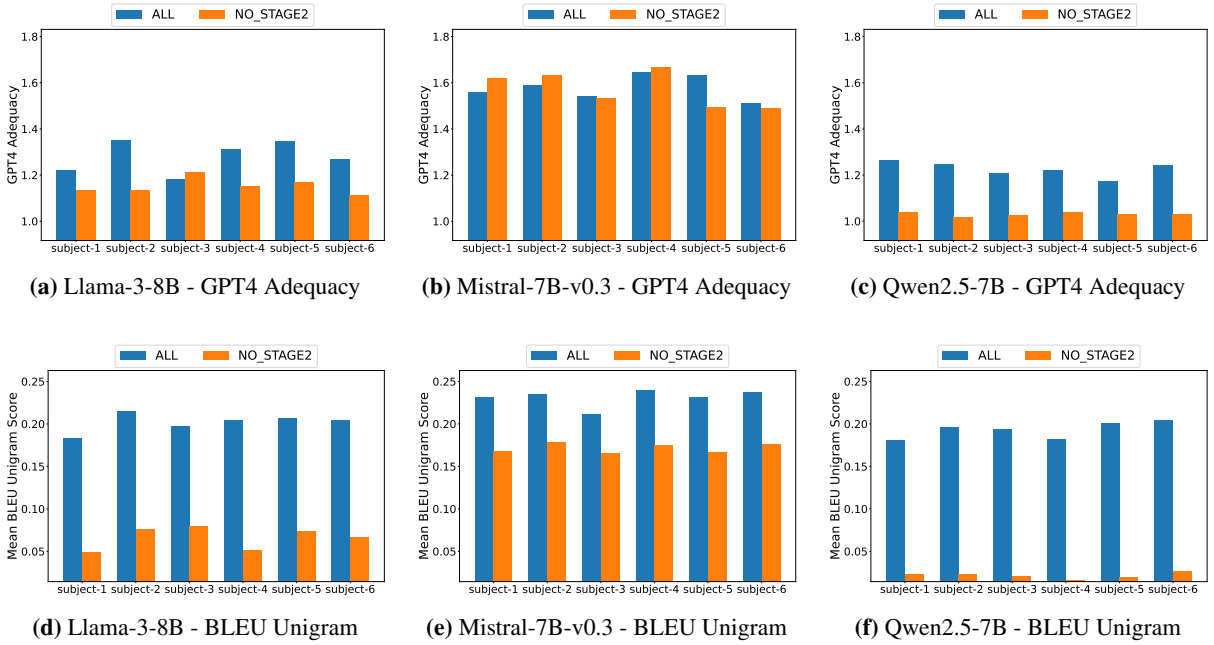
While our initial hypothesis was that EEG embeddings – derived from noisy multichannel EEG data – might not fully capture complex thoughts, thus requiring additional inputs like object labels, it is pleasantly surprising to observe that even without object labels, the models (ONLY_EEG) perform comparably with the best models (ALL). This underscores the effectiveness of aligning EEG embeddings with vision embeddings in stage 1 and pretraining the LLM, particularly the projectors, in stage 2 with vision embeddings.

6.3 Subject-wise Analysis

For this analysis, each subject's EEG data is used to independently train and test the LLMs, simulating a personalized solution. The dataset comprises six subjects, allowing for individual analysis to evaluate the robustness of the approach across different participants. As depicted in Figure 2, in the subject-wise analysis, the advantages of the complete approach (ALL) become even more prominent. Cross-subject and in-subject experiments consistently favor the ALL configuration, with significant improvements in *Adequacy* score when Stage 2 is included, especially for LLaMA3-8B and Qwen2.5-7B models. These improvements, though numerically small, are crucial in the context of EEG data where every bit of alignment and finetuning matters due to its inherently noisy and sparse nature.

The NO_STAGE2 configuration, which omits the essential alignment step between EEG embeddings and image embeddings, consistently demonstrates lower performance across subjects, as illustrated by the BLEU Unigram scores in Figure 2. This validates our hypothesis that direct finetuning without the warm-up provided by Stage 1 and Stage 2 is insufficient for EEG data. The challenge is further exacerbated by the inherent difficulties associated with EEG data, which, even when ethically collected and cleaned, still suffer from limited data availability and significant variability across sessions and subjects.

Subject-wise analysis is essential in practice due to the inherently private and sensitive nature of EEG data. Models like ours, designed for thought-to-text translation, must be developed and deployed within privacy-preserving settings. Creating a personalized solution without access to extensive cross-subject EEG data can be challeng-



(a) Llama-3-8B - GPT4 Adequacy **(b)** Mistral-7B-v0.3 - GPT4 Adequacy **(c)** Qwen2.5-7B - GPT4 Adequacy

(d) Llama-3-8B - BLEU Unigram **(e)** Mistral-7B-v0.3 - BLEU Unigram **(f)** Qwen2.5-7B - BLEU Unigram

Figure 2: Subjectwise analysis comparing 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.

ing. To this end, our multi-stage approach demonstrates that pretraining on non-EEG data (such as images) and fine-tuning on small amounts of subject-specific EEG data opens new avenues for privacy-preserving personalized EEG-LLM model development.

6.4 Qualitative Inspection and Basic Error Analysis

For qualitative inspection, we compared the generated descriptions of images produced by both the GPT-4 and the MISTRAL-7B-V0.3 model (ALL and ONLY_EEG variants). Table 3 in the Appendix section presents notable examples from our assessment of the model's generated descriptions. In many instances, the approach with input of EEG + predicted object label generates highly accurate descriptions, as illustrated in positive examples 2 and 4 for the piano and pumpkin, respectively. However, in cases of misgeneration, we encounter not only significantly inaccurate outputs and hallucinations, as seen in examples 6, 7, and 8, but also instances of genuine confusion. For example, the EEG signal for a flower is misidentified as a mushroom in example 5, leading to incorrect generation. This misidentification indicates potential areas for improvement in object classification.

However, a key advantage of the EEG-only approach is that even when the predicted object label is incorrect, the model can still produce reasonably coherent descriptions, as seen in anecdotal example 1. This highlights the robustness of the EEG embeddings in guiding the language model's generation. Similar observations were made with other models based on LLaMa and Qwen architectures, further validating the consistency and reliability of our approach across various LLM frameworks. We acknowledge that the selected examples are purposefully chosen to illustrate different scenarios.

7 Conclusion and Future Work

In this paper, we introduced a novel approach to convert EEG signals into text, leveraging instruction-tuned LLMs fine-tuned with EEG data. Our method progresses through three stages: training an EEG encoder for feature extraction, fine-tuning LLMs on multimodal data, and further refining them with EEG embeddings for direct text generation from neural signals. Validation on a public EEG dataset demonstrates the efficacy of popular LLMs in "transforming" thoughts evoked by viewing images into text. We observed significant performance improvements compared to chance evaluation, and our methodology, incorporating all

stages, performed well in both cross-subject and in-subject analyses, as validated through quantitative evaluation. The qualitative evaluation provided further insights into various scenarios involving EEG signals and object labels versus EEG signals alone as inputs for generating text. These evaluations not only reinforce the effectiveness of our methodology for efficient text generation but also underscore the potential of utilizing EEG data alone to achieve the desired results. However, instances of misidentification that result in incorrect outputs reveal opportunities for further improvements in text generation.

Our future work will focus on optimizing the model architecture, leveraging foundational pre-trained EEG models on diverse dataset, improving EEG-text alignment through stage 2 training on large scale image datasets and diverse tasks such as optical character recognition, question answering and summarizing and exploring practical applications in healthcare and assistive technologies, marking a significant stride toward accessible "thoughts-to-text" systems.

Limitations

Extracting fine-grained information from EEG signals presents challenges due to high data-to-noise ratio and low spatial resolution. Despite these difficulties, EEG signals can still identify object categories, which can then be used with a generic prompt to aid in text generation. However, misclassification of similar-shaped objects (see the Appendix, Figure 3), such as mistaking a mushroom for a flower, underscores potential ambiguities in object recognition. In addition, there were instances where the encoder generated unrelated descriptions, such as identifying a coffee maker as a computer or an elephant as a panda. Object classification accuracy varies among subjects (see the Appendix, Table 4) due to individual differences in interpreting images, leading to diverse EEG signal variations and increased prediction variance. One approach to address this is by training personalized models for each subject and assessing their performance. Additionally, implementing methods that enhance the generalizability of predictions across subjects could be explored. Another challenge is data scarcity; Deep Learning models typically require substantial data for training. Hence, acquiring more high-quality, multi-channel EEG data under controlled experimental conditions is crucial to reduce noise. Despite these challenges, our quantitative and qualitative findings demonstrate promising results. We believe that additional training on larger datasets and rigorous controlled experiments will significantly improve performance.

While reading thoughts can be beneficial for individuals with limited ability to communicate, the major risk lies in the potential misuse of BCIs to intrude into thoughts without consent. However, with appropriate measures and regulations, these risks should not hinder advancements in understanding and translating human cognition, as the benefits outweigh the challenges.

Ethics Statement

For this work, we utilized anonymized open-source EEG data, acknowledging the sensitivity of EEG data and the imperative of ethical compliance. All experimental data used in our research were anonymized to protect participant privacy and uphold ethical standards. Additionally, we employed OpenAI's ChatGPT-4 system to enhance writing efficiency by generating LaTeX code, ensuring concise sentences, and aiding in error debugging.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. 2023b. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. 2023. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*.

Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. 2020. The alice datasets: fmri & eeg observations of natural language comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 120–125.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive

brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107.

Stanislas Dehaene. 2009. *Reading in the Brain: The New Science of How We Read*. Penguin.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. 2024. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Bin He, Bryan Baxter, Bradley J Edelman, Christopher C Cline, and W Ye Wenjing. 2015. Noninvasive brain-computer interfaces based on sensorimotor rhythms. *Proceedings of the IEEE*, 103(6):907–925.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2019. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.

Yuya Ikegawa, Ryohei Fukuma, Hidenori Sugano, Satoru Oshino, Naoki Tani, Kentaro Tamura, Yasushi Iimura, Hiroharu Suzuki, Shota Yamamoto, Yuya Fujita, et al. 2024. Text and image generation from intracranial electroencephalography using an embedding space for text and images. *Journal of Neural Engineering*, 21(3):036019.

Vinay Jayaram and Alexandre Barachant. 2018. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of Neural Engineering*, 15(6):066011.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Nastaran Khaleghi, Shaghayegh Hashemi, Sevda Zafarmandi Ardabili, Sobhan Sheykhivand, and Sebelan Danishvar. 2023. Salient arithmetic data extraction from brain activity via an improved deep network. *Sensors*, 23(23):9351.

Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. 2023. Seeing through the brain: Image reconstruction of visual perception from human brain signals. *arXiv e-prints*, pages arXiv–2308.

Zixing Lei, Yiming Zhang, Yuxin Xiong, and Siheng Chen. 2024. Emergent communication in interactive sketch question answering. *Advances in Neural Information Processing Systems*, 36.

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. 2024a. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pretraining and multi-view transformer. *arXiv preprint arXiv:2405.02165*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Qing Luo, Xiang Gao, Bo Jiang, Xueting Yan, Wanyuan Liu, and Junchao Ge. 2023. A review of fine-grained sketch image retrieval based on deep learning. *Mathematical Biosciences and Engineering*, 20(12):21186–21210.

Stephanie Martin, Iñaki Iturrate, José del R Millán, Robert T Knight, and Brian N Pasley. 2018. Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience*, 12:422.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. *Cognitively inspired natural language processing: An investigation based on eye-tracking*. Springer.

Alex Murphy, Bernd Bohnet, Ryan McDonald, and Uta Noppeney. 2022. Decoding part-of-speech from human EEG signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2201–2210, Dublin, Ireland. Association for Computational Linguistics.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Ramnivas Sharma and Hemant Kumar Meena. 2024. Emerging trends in eeg signal processing: A systematic review. *SN Computer Science*, 5(4):1–14.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Hengcan Shi, Son Duy Dao, and Jianfei Cai. 2024. Llmformer: Large language model for open-vocabulary semantic segmentation. *International Journal of Computer Vision*, pages 1–18.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. 2017. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817.

W Speier, C Arnold, and N Pouratian. 2016. Integrating language models into classifiers for bci communication: a review. *Journal of Neural Engineering*, 13(3):031002.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. 2018. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

James T. Townsend. 1990. Serial vs. parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science*, 1(1):46–54.

David Vivancos and Felix Cuesta. 2022. Mindbigdata 2022 a large dataset of brain signals. *arXiv preprint arXiv:2212.14746*.

Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5350–5358.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 233–243.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xu Zheng, Ling Wang, Kanghao Chen, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. 2024. Eit-1m: One million eeg-image-text pairs for human visual-textual recognition and more. *arXiv preprint arXiv:2407.01884*.

8 Appendix

We present supporting anecdotal examples in Table 3, subject-wise classification results in Table 4, and image classification results in Figure 3.

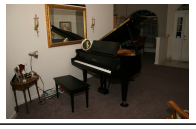
ID	Image	Input	Reference Object	Predicted Object	Reference Description	Thought2Text Description
1		EEG	mushroom	flower	A large yellow mushroom with a brown stem and a brown cap, surrounded by green foliage.	A group of mushrooms growing on a log.
2		EEG + OBJ	piano	piano	A black grand piano in a living 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.

Subject	Accuracy
Subject 1	0.631
Subject 2	0.631
Subject 3	0.544
Subject 4	0.600
Subject 5	0.594
Subject 6	0.531
Overall	0.530

Table 4: Subject-wise Object Classification Accuracy obtained using raw EEG Signals from Stage-1 (%). The train-test splits used to obtain these results are identical to those used in the main experiments.

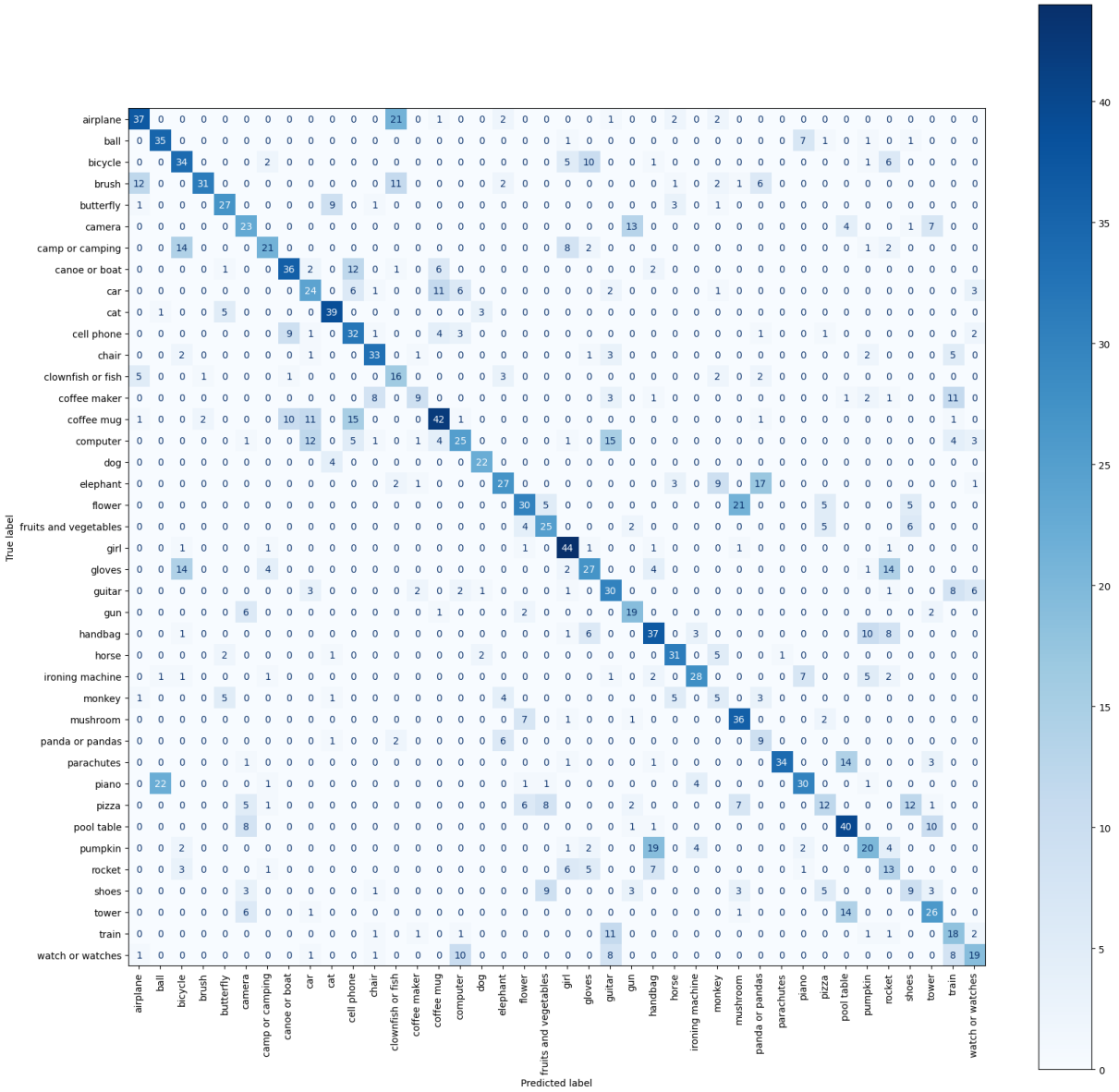


Figure 3: Confusion Matrix for classification results from stage 1. Some misclassifications are flowers being identified as mushrooms, airplane as identified as clownfish or fish, elephant as a panda etc.