VLFEEDBACK-EEG: Neural Signals as Implicit Feedback for Vision-Language Model Alignment

Angela Lopez-Cardona [*1,2] **Sebastian Idesis** [*1] **Mireia Masias Bruns** [1] **Matteo Mazzini** [2] **Joemon M. Jose** [3] **Sergi Abadal** [2] **Ioannis Arapakis** [1]

Abstract

We introduce and publicly release VLFEEDBACK-EEG, the first dataset pairing EEG recordings from 21 participants with Large Vision-Language Models (LVLMs) preference annotations, enabling future research on physiological signals as implicit feedback for multimodal alignment. Our analysis suggests that EEG features under pairwise evaluation contain preference information, aligning more closely with human judgments than with AI-generated annotations. An exploratory fusion with pre-trained reward models does not yet surpass RM-only baselines, but real EEG consistently outperforms noise controls, suggesting that physiological signals provide complementary information for preference modelling. We release VLFEEDBACK-EEG publicly to support future research on cognitively-grounded multimodal alignment.

1. Introduction

LVLMs (Murad et al., 2025; OpenAI, 2024), which extend Large Language Models (LLMs) beyond purely textual inputs to incorporate visual information, have been gaining popularity partly due to their capacity to enable richer cross-modal understanding and generation (Yin et al., 2024). However, many of the core challenges that affect LLMs remain, and often become more pronounced with the addition of a visual modality. Issues such as human alignment, hallucination control, and robust multimodal integration are unresolved challenges to date (Casper et al., 2023). Although current LVLMs achieve strong performance on standardized benchmarks, these benchmarks are typically misaligned with real users' attentional focus and intentions. As a result, responses can be imprecise or unhelpful in everyday contexts (Yan et al., 2024), with hallucinations being particularly severe (Yin et al., 2024). Addressing these challenges centrally requires robust alignment methods: while Reinforcement Learning from Human Feedback (RLHF) has become the dominant approach, its reliance on explicit human annotation introduces annotator biases, and, as we show, may not fully capture the cognitive signals underlying human preference.

Bridging the gap between the human brain and language processing has become an important area of research (Murad et al., 2025; Lopez-Cardona et al., 2025), with cognitive signals showing promise in enriching datasets with user-derived behavioural information (Lopez-Cardona et al., 2025; Ding et al., 2022). For example, recent studies have been leveraging Eye-tracking (ET) as implicit feedback to enhance LLMs (Lopez-Cardona et al., 2024; Kiegeland et al., 2024; Galliamov et al., 2025). Less explored are neuroimaging techniques, which may provide high throughput neural information. For instance, Ye et al. (2025) employed functional Magnetic Resonance Imaging (fMRI) to improve language generation. Electroencephalography (EEG), on the other hand, stands out as a non-invasive and portable method capable of capturing electrical activity from the scalp with high temporal resolution, making it suitable for studying real-time neural dynamics, despite typical sources of noise (Babu et al., 2025). There is a growing convergence between LLMs and EEG research, enabling new directions in neural decoding, brain-computer interfaces (BCIs), and affective computing (Babu et al., 2025). These data have also been combined with ET to improve Natural Language Processing (NLP) performance in works such as (Zhang et al., 2024; Ren & Xiong, 2021), with initial approaches for language generation emerging (Mishra et al., 2025; Tao et al., 2025). However, despite recent advancements, it remains unclear how EEG can be leveraged to study human responses in the context of LVLMs, leaving a gap in our understanding of how neural signals can best support multimodal model alignment.

[*]Equal contribution Accepted at the RLxF: Reinforcement Learning from World Feedback Workshop, ICML 2026. [1]Telefónica Research, Barcelona, Spain [2]Universitat Politècnica de Catalunya, Barcelona, Spain [3]University of Glasgow, Glasgow, United Kingdom. Correspondence to: Angela Lopez-Cardona <angela.lopezcardona@telefonica.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

To address this gap, we collect and release VLFEEDBACK-EEG, the first EEG dataset recorded during LVLM preference annotation. Our main contributions are:

- **A novel dataset:** VLFEEDBACK-EEG comprises EEG recordings from 21 participants as they read and evaluated model-generated responses from VLFEEDBACK (Li et al., 2023; 2024a), yielding the first resource of its kind for cognitively-grounded multimodal alignment (Section 3).
- **Preference decoding:** Via classifier training (Section 4), we show that EEG features carry preference information, aligning more closely with human judgments than with AI-generated annotations.
- **Reward model fusion:** An exploratory fusion with pretrained Reward Models (RMs) (Section 5) shows that real EEG consistently outperforms noise controls, providing a proof-of-concept for physiological signals as complementary preference feedback.

2. Related work

2.1. LVLMs

Existing LVLMs extend pre-trained LLMs by incorporating visual encoders through adapter modules (Zhu et al., 2025), enabling language models to interpret and generate grounded descriptions of visual inputs (Zhou et al., 2024). Despite recent advances, key challenges remain, such as visual hallucinations and modality degradation (Zhou et al., 2024; Li et al., 2024b).

Most LVLMs follow a two-stage training pipeline: (1) *modality alignment* via large-scale, image-text pre-training, and (2) *instruction tuning* to improve reasoning and preference alignment (Ouyang et al., 2022). For the second stage, RLHF (Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) are widely adopted, relying on preference triplets (x, y_w, y_l). To reduce annotation cost, Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022) replaces human feedback with LLM-generated supervision and is employed in VLFEEDBACK (Li et al., 2023), the dataset used in this work.

Reward modelling. Central to both RLHF and RLAIF is the RM, which provides the scalar signal that drives preference optimisation; its quality directly bounds how well the resulting LVLM aligns with human intent. Multimodal RMs for LVLMs follow three main approaches: *model-as-judge*, where a model is prompted to score responses directly (Yasunaga et al., 2025; Xiong et al., 2025); *reward-head regression*, where a lightweight head on top of latent representations outputs a scalar (Zang et al., 2025); and *critique-then-score*, where the model generates an intermediate reasoning step before scoring (Zhang et al., 2025b;a). These are evaluated on benchmarks such as VL-RewardBench (Li et al., 2025) and Multimodal RewardBench (Yasunaga et al., 2025). We use LLaVA-Critic (Xiong et al., 2025) and MM-RLHF-Reward (Zhang et al., 2025b) as baselines in our fusion experiments (Section 5).

2.2. EEG in NLP

Recent research has demonstrated that EEG can be directly leveraged to enhance NLP systems. For instance, using the ZuCo dataset (Hollenstein et al., 2018), Ren & Xiong (2021) introduced CogAlign, a multi-task neural network able to align model text representations with cognitive processing signals across various NLP tasks. Using an adversarially trained shared encoder, CogAlign transfers cognitive knowledge to datasets without recorded signals. Building on this, Zhang et al. (2024) proposed CogAware, which learns cross-domain features while preserving modality-specific characteristics, and later Harada & Oseki (2024) tested it with the Cognitive Information Bottleneck (CIB), a method that extracts only the task-relevant information from cognitive processing signals.

A parallel line of work has focused on directly decoding EEG into text. DeWave (Duan et al., 2023) introduced discrete codebook encoding of EEG waves aligned with a pretrained BART decoder. EEG2TEXT (Liu et al., 2024) combines convolutional transformers with a multi-view transformer to capture spatially distributed brain activity. Thought2Text (Mishra et al., 2025) and the R1 Translator (Murad et al., 2025) pair dedicated EEG encoders: a ChannelNet CNN and a BiLSTM, respectively, with pretrained language decoders, while SEE (Tao et al., 2025) addresses cross-modal domain gaps via a semantic matching module. More recently, SENSE (Murhekar et al., 2026) decouples decoding into a lightweight keyword extraction module and zero-shot LLM prompting. These encoder–decoder designs directly inform our fusion architecture (Section 5), where we similarly project EEG into an embedding space compatible with a pretrained model.

However, Jo et al. (2025) showed that several such models achieve comparable performance on random noise inputs and that teacher forcing inflates evaluation metrics by at least threefold, questioning whether current systems perform genuine neural decoding. Together with limited cross-subject generalisation, these findings suggest that EEG may be more reliably leveraged as a complementary cognitive signal than as a standalone input, consistent with the approach explored in this work.

2.3. Cognitive signal datasets

There are few available EEG datasets curated for NLP applications. Some combine ET and EEG, such as ZuCo (Hollen-

stein et al., 2018), ZuCo 2.0 (Hollenstein et al., 2020), and the more recent RaCCooNS (Frank & Aumeistere, 2024), which contains signals of participants tasked with reading narrative sentences in Dutch. Other datasets rely exclusively on EEG, such as CVPR2017 (Palazzo et al., 2021) or MOABB (Jayaram & Barachant, 2018). More recent examples include ChineseEEG-2 (Chen et al., 2025), collected during reading aloud and passive listening tasks in Chinese, and DERCo (Quach et al., 2024), which contains EEG recordings from 22 native English speakers reading narrative texts. While prior work has explored EEG during natural reading or language generation, none has examined the neural correlates of human preference judgments in multimodal alignment tasks. To the best of our knowledge, we are the first to record EEG while participants read and annotate LVLM prompts and responses in the human alignment framework.

3. Dataset creation

We recorded EEG from 21 participants as they read and evaluated pairs of responses to Visual Question Answering (VQA) prompts drawn from the VLFEEDBACK dataset. In what follows, we describe the experimental setup in Subsection 3.1, data collection protocol in Subsection 3.2, and data preprocessing in Subsection 3.4.

3.1. Stimulus

VLFeedback (Li et al., 2023; 2024a) dataset served as stimulus source, a large multi-modal preference dataset used to align LVLMs. VLFeedback contains 80k multi-modal instructions over 65k unique images, collected from diverse vision-language sources. Twelve LVLMs (including, among others, GPT-4 (OpenAI, 2024), LLaVA (Liu et al., 2023), Qwen-VL-Chat (Bai et al., 2023), IDEFICS (Laurençon et al., 2023), InstructBLIP (Dai et al., 2024)) were used to generate responses, with GPT-4 acting as an automatic annotator scoring outputs for helpfulness, visual faithfulness, and ethical soundness. From these scores, 20k preference pairs (preferred vs. rejected responses) were built and expanded into pairwise comparisons for preference-based fine-tuning. For this study, responses were filtered by length (10–750 characters), and 30 prompt–response triplets were randomly selected for experimentation.

3.2. Participants and procedure

Twenty-one participants (*mean age* = 30.4 ± 4.15 years; 4 female) were employed in this study. All reported normal or corrected-to-normal vision, no history of neurological or psychiatric disorders, and no use of relevant medication. Participants had at least an upper-intermediate level of English proficiency. Written informed consent was obtained from all participants prior to the experiment. Participants also completed a brief demographic questionnaire upon arrival and received monetary compensation for their time. Before starting, participants were given a detailed overview of the study objectives and task sequence. The experimenter ensured that the instructions were fully understood before proceeding. Stimulus presentation and behavioural response collection were implemented in PsychoPy (Peirce et al., 2019).

Each trial followed the sequence illustrated in Figure 1. Initially, participants viewed an instruction screen outlining the evaluation criteria, a given VQA prompt, followed by a fixation cross and the first model-generated response, where they were asked to rate it according to helpfulness, visual faithfulness, and ethical criteria (aligned with those used by GPT-4 when annotating the VLFEEDBACK dataset, Figure 6). After confirming the rating, the second response appeared and was evaluated in the same way. Finally, participants selected which of the two responses they preferred.

All participants viewed the same set of 30 VQA, each associated with an image and two candidate responses. The order of questions and response pairs was randomized for every participant to minimize order effects. EEG data were continuously recorded throughout the session, with particular focus on the response presentation windows corresponding to active evaluation (Figure 1). The entire experiment lasted approximately 45 minutes per participant, including setup and debriefing. One example of this combination of question, image, and responses is shown in Figure 7.

3.3. Apparatus

The stimulus was presented on a 27-inch monitor (23.53 $\times$ 13.24 inches). EEG was recorded using a gel-based 32-channel actiCHamp system (Brain Products GmbH, Germany), with a sampling rate of 250 Hz. A high-impedance amplifier was used to ensure optimal signal quality and minimise noise. Electrode placement followed the international 10-20 system, standard in EEG acquisition. The ground and reference electrodes were positioned at Fpz and Cz, respectively. Impedances were kept below 10 kΩ, following standard procedures. Electrode preparation took approximately 15 minutes per participant. A complete list of electrodes is illustrated in Figure 4. Before recording, the scalp was cleaned and lightly abraded to reduce impedance. Visual inspection ensured that all electrodes were properly attached and impedance levels remained within acceptable thresholds throughout the session.

3.4. Data-preprocessing and feature extraction

Figure 2 shows the preprocessing and feature extraction pipeline. Artifact correction was carried out using Independent Component Analysis (ICA), allowing for the removal of ocular and muscle-related components. EEG channels contaminated with persistent noise were interpolated us-

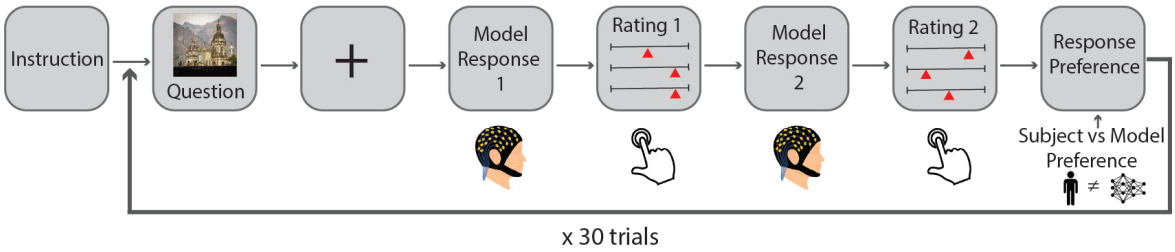


Figure 1. Task procedure: For each trial, participants viewed an image and question, rated two sequential responses, and chose one.

ing spherical splines when necessary. Band-pass filtering (0.5–40 Hz) was applied, and for frequency-related metrics, this was followed by the computation of the Power Spectral Density (PSD), obtained as the squared magnitude of the Fast Fourier Transform (FFT). Welch's PSD and band-limited measures were then computed over the canonical frequency bands (Cochran et al., 1967; Zheng & Lu, 2015): δ (0.5–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), and γ (30–40 Hz). To avoid confounding EEG features with differences in response duration, all response epochs were cropped to the same fixed-length window before feature extraction. Therefore, entropy, spectral, and connectivity features were computed from EEG segments of equal duration across preferred and rejected responses. A list of EEG features was derived, with further details provided in Appendix A.1.

- **Mean Absolute Amplitude**: Average absolute signal magnitude (Pillay et al., 2020).
- **Shannon Entropy**: Measure of signal complexity, quantifying unpredictability in the EEG time series (Phung et al., 2014).
- **Conditional Entropy**: Captures dependency between successive signal states, indicating temporal structure in brain activity (Keshmiri, 2021).
- **Alpha Asymmetry**: Difference in alpha-band power between left (F3) and right (F4) frontal electrodes, commonly linked to affective and motivational states (Smith et al., 2017).
- **Band Power**: Average spectral power within canonical frequency ranges (δ, θ, α, β, γ) (Newson & Thiagarajan, 2019).
- **PSD**: Integral of the power spectral density across the full frequency range, summarising total EEG power (Wang et al., 2015).
- **Phase-Locking Value (PLV)**: Average strength of phase synchronisation across channel pairs within each frequency band, representing functional connectivity (Roach & Mathalon, 2008).

Not all metrics were interpreted as channel-wise features. Phase-locking value reflects synchronization between pairs of channels, while alpha asymmetry is computed from predefined left–right electrode comparisons. Therefore, these features were treated as connectivity- or contrast-level measures rather than as independent features for each individual channel.

4. EEG-Based Preference Decoding

Next, we evaluated whether EEG features carry information about human preference by training multiple classifiers predicting different labels. The original VLFEEDBACK dataset was annotated by GPT-4 following an RLAIF-style preference protocol (*model preference*), whereas our labels were collected from participants (i.e., *user preference*). When aggregating participant responses, we refer to the majority label as *user preference averaged*. Participants rated and explicitly chose between responses according to the same criteria. There is moderate inter-annotator agreement ($\kappa = 0.453$) and a model accuracy of 76.7% against the user majority. Additional corpus statistics are reported in Appendix B.1.

4.1. Classification setup

We tested several classifiers, including logistic regression (Hosmer et al., 2013), naive Bayes, support vector machine (SVM) (Cortes & Vapnik, 1995), random forests (Breiman, 2001), gradient boosting (Friedman, 2001), and k-nearest neighbours (Cover & Hart, 1967). Generalisation to unseen participants was estimated via nested, subject-independent, 5-fold cross-validation. In each outer fold, all samples from a held-out set of subjects formed the test partition, while models were trained on the remaining subjects. Within each outer training set, a grid search over hyperparameters was conducted using inner subject-independent 3-fold cross-validation. Preprocessing steps, including median imputation and standardisation, were fitted exclusively on the training folds. Models selected in the inner loop were retrained on the full outer training set and evaluated on the held-out subjects.

We considered two labelling regimes. In *pointwise* mode,

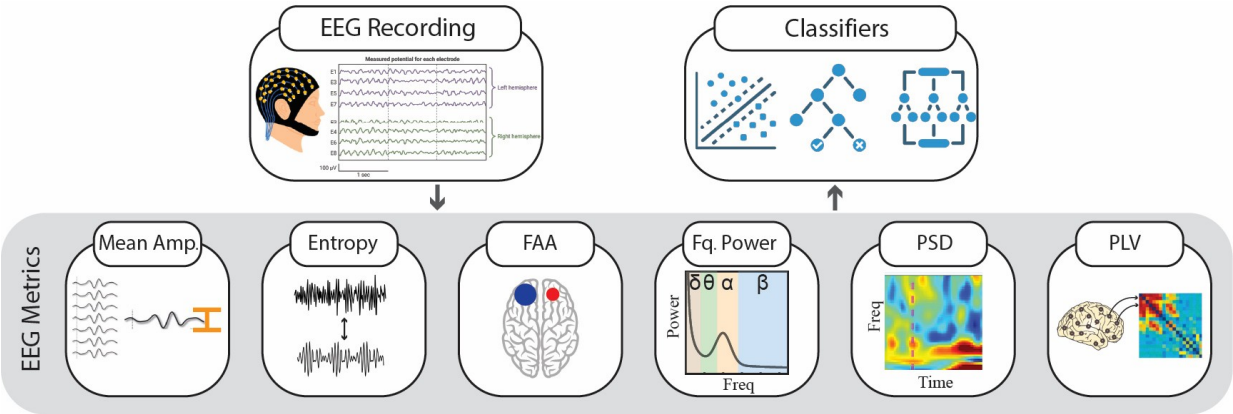


Figure 2. EEG preprocessing pipeline. We first record the raw EEG signal and apply a band-pass filter, followed by an FFT transformation to obtain its frequency-domain representation. Using the resulting time–frequency decomposition, we extract features from the standard frequency bands, which are then used as inputs to the classifier.

each response is represented as a separate row and the classifier predicts whether it is preferred or rejected independently. In *pairwise* mode, each row corresponds to a dyad: features from both responses are concatenated, and a single label encodes which response was chosen.

Reported performance is the mean accuracy across folds. Statistical significance was assessed using one-tailed t-tests comparing each classifier's mean accuracy against chance (50%), with the alternative hypothesis that performance exceeds chance. Effect sizes were calculated using Cohen's d, and 95% confidence intervals were computed for mean accuracy estimates. Results are shown in Table 1 and Table 2, where * denotes $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; values in brackets are 95% CIs and Cohen's d in parentheses. More detailed results are provided in Appendix A.3.

4.2. Classification results

Table 1 reports classification accuracy using all EEG features (Subsection 3.4). Under Pointwise labels, performance remains near chance across all classifiers and labels. However, the pairwise regime yields consistent improvements: when averaging across subjects, all classifiers significantly exceed chance, with KNN reaching 59.63% ($p < 0.001$) and Gradient Boosting 59.40% ($p < 0.01$). This suggests that relative differences between paired EEG responses are more discriminative than absolute feature values.

4.3. Exploratory analysis

To further investigate which EEG metrics drive differentiation, we fitted a linear mixed-effects model per subject and metric, using channel and trial as random intercepts, with p-values corrected using the Benjamini–Hochberg FDR procedure. This analysis revealed that entropy-based metrics differentiated preferred from non-preferred responses more strongly than conventional band-power features (Appendix A.2), consistent with prior work showing EEG complexity measures can characterise cognitive-state differences more effectively in certain tasks (Zarjam et al., 2013; Kosciessa et al., 2020; Lau et al., 2022). These properties may be more sensitive to the cognitive processes underlying preference formation, such as semantic evaluation, uncertainty resolution, and comparative judgment. A per-feature classifier across 31 channels confirmed this pattern (Table 5 and Table 6). Motivated by this finding, we ran the same classification pipeline using only the entropy features from all 30 channels.

Table 2 shows results when using entropy-based features (Shannon and conditional entropy across all channels for each response). Entropy features match or improve upon the full tabular set. Under pairwise labels, Gradient Boosting reaches 60.78% ($p < 0.01$) and Logistic Regression 60.23% ($p < 0.001$) on subject-averaged preferences. Individual-label performance also improves relative to the tabular set, with several classifiers reaching significance on model-level preferences.

5. EEG-Augmented Reward Modelling

Having established that EEG features carry a modest but statistically significant preference signal (Section 4), we now explore whether this signal can complement existing RMs within the human-alignment framework, following the spirit of prior work that integrates EEG into NLP tasks (Zhang et al., 2024; Ren & Xiong, 2021) and ET also on reward modelling (Lopez-Cardona et al., 2024) .

5.1. Setup

Reward model baselines. We evaluate two pretrained multimodal RMs as baseline predictors of preference: LLaVA-Critic (Xiong et al., 2025) and MM-RLHF-Reward (Zhang et al., 2025b). For each trial, the RM pro-

Table 1. Classifier accuracy (%) by preference; **all features EEG**; **Bold** indicates the highest value in each column. Underline indicates the highest value in each row

	POINTWISE LABELS			PAIRWISE LABELS		
Classifier	**Model**	**Subj.**	**Subj. avg**	**Model**	**Subj.**	**Subj. avg**
Gradient Boosting	50.004	51.541	50.558	49.897	53.560	59.401**
KNN	50.887	49.539	50.892	**54.742**	53.359	**59.630*****
Linear SVM	47.614	50.205	52.711*	50.291	51.837	55.015*
Logistic Regression	48.864	49.706	51.026	52.218	52.608	56.051**
Naive Bayes	51.654	**51.957**	**52.814**	51.251*	**54.537***	57.980**
Random Forest	**51.941**	48.957	52.527	52.802	52.031	58.049**

Table 2. Classifier accuracy (%) by preference; **segments entropy**; **Bold** indicates the highest value in each column. Underline indicates the highest value in each row.

	POINTWISE LABELS			PAIRWISE LABELS		
Classifier	**Model**	**Subj.**	**Subj. avg**	**Model**	**Subj.**	**Subj. avg**
Gradient Boosting	51.383	50.789	**55.461****	55.478*	**54.685**	**60.776****
KNN	52.399**	50.278	55.342	53.950*	52.373	59.863**
Linear SVM	49.924	50.360	53.364*	52.973*	48.790	60.386**
Logistic Regression	51.053	49.594	53.084*	54.303**	51.869	60.232***
Naive Bayes	**52.495***	**51.221**	54.870**	51.044	51.409	56.012**
Random Forest	49.643	48.567	54.253*	**56.231****	52.215	59.836*

duces scores s_a and s_b for the two responses. A binary RM preference is formed as $\mathrm{rm_pref} = \mathbf{1}[s_a > s_b]$. Per-fold accuracy compares this prediction to the human label.

Fusion architecture. Our architecture follows the encoder, fusion paradigm widely adopted in recent EEG-language work: a dedicated EEG encoder produces a fixed-dimensional representation that is then aligned with a frozen pretrained model. This two-stage design mirrors the pipelines used in EEG-to-text systems, where convolutional, recurrent or transformer encoders project raw signals into an embedding space compatible with pretrained language models (Duan et al., 2023; Mishra et al., 2025; Liu et al., 2024; Murad et al., 2025).

In our case, the pretrained backbone is a multimodal RM rather than a language generator. We explore two branches for encoding the EEG signal. In the raw-signal branch (M1 in Figure 3), EEG segments are encoded by a 1D CNN into an embedding z_{eeg}. In the feature branch (M2 in Figure 3), the features described in Subsection 3.4 or the per-channel entropy features from Subsection 4.3 are passed through a small MLP to produce z_{eeg}, providing a lower-variance alternative when labelled data are scarce. Both branches feed into a shared gated fusion module that learns a residual correction on top of the frozen RM score. Concretely, z_{eeg} is first linearly projected to the RM embedding dimension, and the fused representation is formed as in Equation (1).

$$x = [e_{\mathrm{rm}}; z_{\mathrm{eeg}}; e_{\mathrm{rm}} \odot z_{\mathrm{eeg}}; s_{\mathrm{rm}}] \tag{1}$$

$$\hat{s} = s_{\mathrm{rm}} + f_g(x) \cdot f_\Delta(x) \tag{2}$$

where e_{rm} is the frozen RM embedding, z_{eeg} is the EEG representation, s_{rm} is the base RM scalar score, f_g is a learned MLP with sigmoid output that produces the gate value, and f_Δ is a learned MLP that produces the residual correction (Equation (2)). The gated residual formulation ensures that when the EEG signal is uninformative, the gate $f_g(x)$ can collapse to zero and the model gracefully falls back to the base RM score. The model is trained end-to-end with a pairwise ranking loss (Equation (3)).

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \sigma\left(\hat{s}_i^+ - \hat{s}_i^-\right) \tag{3}$$

where $\hat{s}^+$ and $\hat{s}^-$ denote the predicted scores for the chosen and rejected responses within each pair, following standard practice in RM training (Ouyang et al., 2022). To verify that the fusion model genuinely exploits the EEG data, we evaluate each trained model under three conditions: (*i*) **normal** – real EEG data; (*ii*) **no-gate** – RM score alone, bypassing the fusion head entirely; and (*iii*) **random-features** – the EEG data z_{eeg} is replaced with Gaussian noise.

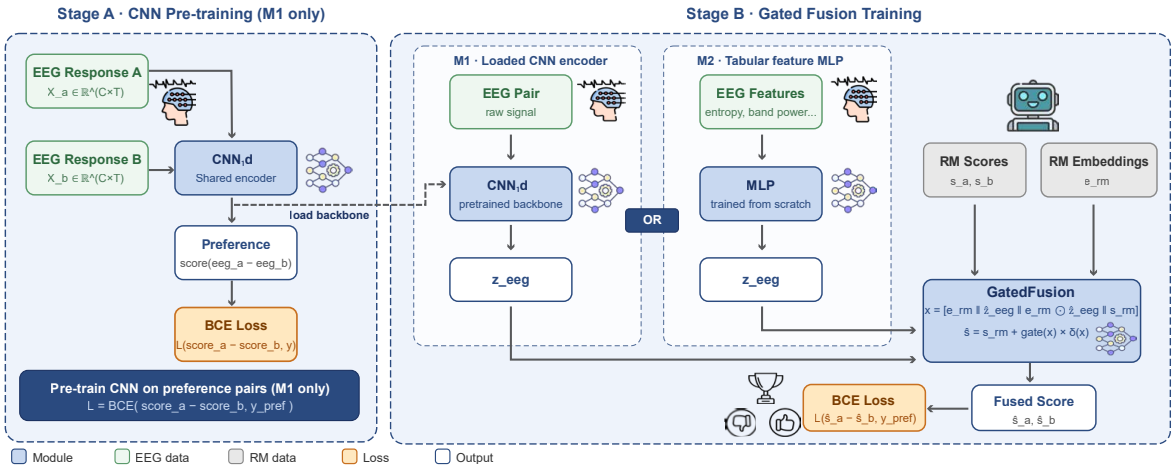


Figure 3. Two-stage fusion pipeline. Stage A (M1 only) shows the CNN pre-training on pairwise EEG preference labels, whose backbone weights are then loaded into Stage B. Stage B illustrates both encoding methods side by side: M1 uses the pretrained CNN backbone and M2 passes tabular EEG features through an MLP trained from scratch; both feed into a shared gated fusion module to produce the final preference score.

Data splits and evaluation. We use two evaluation protocols to assess different aspects of generalisation. In the *trial split*, held-out trials define the test set while all participants may contribute to both training and test pairs across folds, evaluating generalisation to new trials but not strict cross-subject transfer. In the *dual split*, both trials and subjects are held out simultaneously, providing evidence also of cross-subject generalisation at the cost of a reduced test set ($N \approx 66$ pairs/fold vs. $N \approx 200$). Both protocols use 3-fold cross-validation and are repeated across 6 random seeds to account for optimisation and partition variance. For each evaluation, we assess the RM on exactly the same test sets as each method, resulting in varying N depending on the data discarded.

5.2. Results

Table 3 reports test accuracy (mean $\pm$ std across folds and seeds) for all fusion pipelines and RM-only baselines under *trial split* evaluation. The raw EEG encoder underperforms its RM baseline with high variance, consistent with overfitting at this data scale. The tabular features pipeline narrows the gap, particularly with LLaVA-Critic, while the entropy pipeline achieves the closest performance to its RM baseline and consistently outperforms its random-feature ablation across both RMs, providing evidence that the fusion module captures genuine preference-relevant information rather than fitting to noise. Table 4 reports results under the stricter *dual split*, where both trials and subjects are held out. Performance drops across all pipelines, confirming cross-subject generalisation as the primary bottleneck. The entropy pipeline remains the most competitive, maintaining a gap of approximately 3 percentage points to its RM baseline under both RMs, and retaining its real-vs-random separation, unlike the raw encoder, which loses this separation under the stricter protocol.

Table 3. Test accuracy of EEG fusion pipelines against RM-only baselines under *trial split* evaluation (mean $\pm$ std, 6 seeds $\times$ 3 folds). (rnd.) = random EEG.

Pipeline	LLaVA-Critic	MM-RLHF
Raw EEG	0.635 ± 0.027	0.654 ± 0.028
Raw EEG (rnd.)	0.630 ± 0.027	0.631 ± 0.033
Reward model†	0.679 ± 0.001	0.745 ± 0.000
EEG features	0.665 ± 0.024	0.699 ± 0.012
EEG features (rnd.)	0.641 ± 0.015	0.691 ± 0.022
Reward model‡	0.688 ± 0.003	0.760 ± 0.005
EEG entropy	0.686 ± 0.040	0.720 ± 0.049
EEG entropy (rnd.)	0.658 ± 0.035	0.694 ± 0.027
Reward model§	0.712 ± 0.024	0.752 ± 0.022

†RM on raw-EEG test pairs, N=201. ‡RM on tabular-feature test pairs, N=199. §RM on segments test pairs, N=174.

6. Discussion

Classification significantly exceeds chance under pairwise labels and user-averaged preferences, with stronger differentiation against participants' judgments than GPT-4's annotations, suggesting that EEG captures aspects of human preference that RLAIF-based labels do not fully reflect. Entropy-based features were more discriminative than band-power measures, consistent with prior work on EEG complexity measures for cognitive-state differentiation.

Regarding fusion, the entropy pipeline emerges as the most robust branch, outperforming its random-feature ablation across both evaluation protocols and both RMs. This real-vs-random separation persists even under the stricter dual

Table 4. Test accuracy of EEG fusion pipelines against RM-only baselines under *dual split* evaluation (mean ± std, 6 seeds × 3 folds). (rnd.) = random EEG.

Pipeline	LLaVA-Critic	MM-RLHF
Raw EEG	0.611 ± 0.036	0.629 ± 0.038
Raw EEG (rnd.)	0.627 ± 0.028	0.646 ± 0.033
Reward model†	0.657 ± 0.016	0.734 ± 0.014
EEG features	0.636 ± 0.059	0.649 ± 0.087
EEG features (rnd.)	0.611 ± 0.028	0.650 ± 0.039
Reward model‡	0.653 ± 0.022	0.733 ± 0.031
EEG entropy	0.663 ± 0.033	0.699 ± 0.025
EEG entropy (rnd.)	0.662 ± 0.034	0.687 ± 0.034
Reward model§	0.693 ± 0.015	0.733 ± 0.020

†RM on raw-EEG test pairs, N=67. ‡RM on tabular-feature test pairs, N=66. §RM on segments test pairs, N=58.

split, unlike the raw encoder, which loses it, suggesting that engineered entropy features provide a more generalisable EEG representation than end-to-end learned encoders at this data scale, consistent with the classification findings in Section 4. This is notable given recent concerns that EEG models achieve comparable performance on random noise (Jo et al., 2025); our ablations provide positive evidence to the contrary for the entropy pipeline.

Under the stricter dual split, performance drops consistently across all pipelines, confirming that cross-subject generalisation is the primary bottleneck. The test set also shrinks from ∼200 to ∼66 pairs per fold, confounding generalisation with reduced statistical power. With only ∼630 preference pairs, the EEG branch is prone to overfitting on subject-specific neural noise rather than learning generalisable preference features — as evidenced by the raw encoder losing its real-vs-random separation under this protocol. The remaining gap to RM-only baselines is plausibly attributable to data scale rather than a fundamental incompatibility, as the RM baselines themselves vary by several percentage points across test partitions, suggesting statistical power is a significant confound. Whether more data would close the gap entirely remains open, and the release of VLFEEDBACK-EEG is intended to enable the larger-scale studies needed to resolve this question.

7. Conclusions

We introduce VLFEEDBACK-EEG, providing a first proof-of-concept that physiological signals can serve as implicit, annotation-free feedback for LVLM alignment. Classification results confirm that neural activity carries a modest but real preference signal, and that this signal tracks human judgments more closely than AI-generated labels. This suggests that RLAIF-based annotation does not fully substitute for the cognitive signals underlying genuine human preference, motivating the exploration of complementary world feedback sources.

Fusion with pretrained RMs did not improve over RM-only baselines at the current scale. However, the entropy pipeline consistently outperforms its random ablation across both evaluation protocols, providing evidence that the signal is genuine rather than artifactual. Cross-subject generalisation remains the primary bottleneck, and larger participant pools are the most direct path forward, a challenge that extends beyond this work, as recent large-scale benchmarking across 36 EEG tasks similarly identifies it as one of the field's primary open problems (Banville et al., 2026).

Future work could further explore richer EEG feature spaces and time-frequency decompositions, as well as alternative encoder architectures. Another promising direction is the integration of EEG foundation models (Ouahidi et al., 2025), which have recently emerged as general-purpose encoders pretrained on large corpora of brain recordings. These could be leveraged both for preference classification (Section 4) and as stronger EEG encoders within the fusion architecture (Section 5), potentially improving reward signal quality. However, recent large-scale benchmarking evidence suggests that foundation models only marginally outperform task-specific models across a wide range of EEG decoding tasks, even at scale (Banville et al., 2026), tempering expectations for near-term gains from this direction alone.

Finally, beyond direct reward scoring, passively recorded physiological signals represent a promising class of world feedback, scalable, implicit, and grounded in real cognitive responses, that could complement annotation-based alignment pipelines. More immediately, EEG may be deployed to weight annotation confidence or flag trials where neural responses diverge from assigned labels, a narrower use case that the current data already partially supports.

References

Babu, N., Mathew, J., and Vinod, A. P. Large Language Models for EEG: A Comprehensive Survey and Taxonomy, June 2025. URL http://arxiv.org/abs/2506.06353. arXiv:2506.06353 [eess].

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, October 2023. URL http://arxiv.org/abs/2308.12966. arXiv:2308.12966 [cs].

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J.,

Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Banville, H., d'Ascoli, S., Dahan, S., Rapin, J., Careil, M., Benchetrit, Y., Lévy, J., Panchavati, S., Ratouchniak, A., Mingfang, Zhang, Cascardi, E., Begany, K., Brooks, T., and King, J.-R. NeuralBench: A Unifying Framework to Benchmark NeuroAI Models, May 2026. URL http://arxiv.org/abs/2605.08495. arXiv:2605.08495 [cs.LG] version: 1.

Breiman, L. Random Forests. *Machine Learning*, 45(1): 5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023.

Chen, S., Li, B., He, C., Li, D., Wu, M., Shen, X., Wang, S., Wei, X., Wang, X., Wu, H., and Liu, Q. ChineseEEG-2: An EEG Dataset for Multimodal Semantic Alignment and Neural Decoding during Reading and Listening, August 2025. URL http://arxiv.org/abs/2508.04240. arXiv:2508.04240 [eess].

Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., Maling, G. C., Nelson, D. E., Rader, C. M., and Welch, P. D. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674, 1967.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, January 1967. ISSN 1557-9654. doi: 10.1109/TIT.1967.1053964. URL https://ieeexplore.ieee.org/document/1053964.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 49250–49267, Red Hook, NY, USA, May 2024. Curran Associates Inc.

Ding, X., Chen, B., Du, L., Qin, B., and Liu, T. CogBERT: Cognition-Guided Pre-trained Language Models. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H. (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3210–3225, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.284.

Duan, Y., Zhou, C., Wang, Z., Wang, Y.-K., and Lin, C.-t. DeWave: Discrete Encoding of EEG Waves for EEG to Text Translation. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=WaLI8slhLw.

Frank, S. L. and Aumeistere, A. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657, June 2024. ISSN 1574-0218. doi: 10.1007/s10579-023-09684-x. URL https://doi.org/10.1007/s10579-023-09684-x.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, October 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1013203451. URL https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full. Publisher: Institute of Mathematical Statistics.

Galliamov, K., Titov, I., and Pershin, I. Enhancing RLHF with Human Gaze Modeling. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30625–30631, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.1559/.

Harada, Y. and Oseki, Y. Cognitive Information Bottleneck: Extracting Minimal Sufficient Cognitive Language Processing Signals. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3480–3489, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.309.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1):180291, December 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.291. URL https://www.nature.com/articles/sdata2018291. Number: 1 Publisher: Nature Publishing Group.

Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 138–146, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.18.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, 3rd edition, 2013. ISBN 9780470582473. doi: 10.1002/9781118548387.

Jayaram, V. and Barachant, A. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, December 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aadea0.

Jo, H., Yang, Y., Han, J., Duan, Y., Xiong, H., and Lee, W. H. Evaluating EEG-to-text models through noise-based performance analysis. *Scientific Reports*, 16(1): 350, December 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-29587-x. URL https://www.nature.com/articles/s41598-025-29587-x.

Keshmiri, S. Conditional entropy: a potential digital marker for stress. *Entropy*, 23(3):286, 2021.

Kiegeland, S., Reich, D. R., Cotterell, R., Jäger, L. A., and Wilcox, E. The Pupil Becomes the Master: Eye-Tracking Feedback for Tuning LLMs. In *ICML 2024 Workshop on LLMs and Cognition*, July 2024. URL https://openreview.net/forum?id=8oLUcBgKua.

Kosciessa, J. Q., Kloosterman, N. A., and Garrett, D. D. Standard multiscale entropy reflects neural dynamics at mismatched temporal scales: What's signal irregularity got to do with it? *PLOS Computational Biology*, 16(5): e1007885, May 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007885. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007885.

Lau, Z. J., Pham, T., Chen, S. H. A., and Makowski, D. Brain entropy, fractal dimensions and predictability: A review of complexity measures for EEG in healthy and neuropsychiatric populations. *The European Journal of Neuroscience*, 56(7):5047–5069, October 2022. ISSN 0953-816X. doi: 10.1111/ejn.15800. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9826422/.

Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A. M., Kiela, D., Cord, M., and Sanh, V. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 71683–71702, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., and Kong, L. Silkie: Preference Distillation for Large Visual Language Models, December 2023. URL http://arxiv.org/abs/2312.10665. arXiv:2312.10665.

Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., Kong, L., and Liu, Q. VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.358. URL https://aclanthology.org/2024.emnlp-main.358/.

Li, L., Wei, Y., Xie, Z., Yang, X., Song, Y., Wang, P., An, C., Liu, T., Li, S., Lin, B. Y., Kong, L., and Liu, Q. VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24657–24668, June 2025. doi: 10.1109/CVPR52734.2025.02296. URL https://ieeexplore.ieee.org/document/11092887. ISSN: 2575-7075.

Li, S., Lin, R., and Pei, S. Multi-modal Preference Alignment Remedies Degradation of Visual Instruction Tuning on Language Models. In Ku, L.-W., Martins, A.,

and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14188–14200, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.765.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Liu, H., Hajialigol, D., Antony, B., Han, A., and Wang, X. EEG2Text: Open Vocabulary EEG-to-Text Translation with Multi-View Transformer. pp. 1824–1833. IEEE Computer Society, December 2024. ISBN 979-8-3503-6248-0. doi: 10.1109/BigData62323.2024.10825980. URL https://www.computer.org/csdl/proceedings-article/bigdata/2024/10825980/23yl5pp1fri.

Lopez-Cardona, A., Segura, C., Karatzoglou, A., Abadal, S., and Arapakis, I. Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=uZgK0tcPqd&referrer=%5Bthe%20profile%20of%20%C3%81ngela%20L%C3%B3pez-Cardona%5D(%2Fprofile%3Fid%3D~%C3%81ngela_L%C3%B3pez-Cardona1).

Lopez-Cardona, A., Idesis, S., and Arapakis, I. Integrating Cognitive Processing Signals into Language Models: A Review of Advances, Applications and Future Directions, April 2025. URL http://arxiv.org/abs/2504.06843. arXiv:2504.06843 [cs].

Mishra, A., Shukla, S., Torres, J., Gwizdka, J., and Roychowdhury, S. Thought2Text: Text Generation from EEG Signal using Large Language Models (LLMs). In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3747–3759, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.207. URL https://aclanthology.org/2025.findings-naacl.207/.

Murad, S. A., Dahal, A., and Rahimi, N. EEG-to-Text Translation: A Model for Deciphering Human Brain Activity, May 2025. URL http://arxiv.org/abs/2505.13936. arXiv:2505.13936 [cs].

Murhekar, A., Liu, C., Mishra, A., Roychowdhury, S., and Gwizdka, J. SENSE: Efficient EEG-to-Text via Privacy-Preserving Semantic Retrieval, 2026. URL https://arxiv.org/abs/2603.17109. Version Number: 1.

Newson, J. J. and Thiagarajan, T. C. Eeg frequency bands in psychiatric disorders: a review of resting state studies. *Frontiers in human neuroscience*, 12:521, 2019.

OpenAI. GPT-4V(ision) system card, February 2024. URL https://openai.com/index/gpt-4v-system-card/.

Ouahidi, Y. E., Lys, J., Thölke, P., Farrugia, N., Pasdeloup, B., Gripon, V., Jerbi, K., and Lioi, G. REVE: A Foundation Model for EEG - Adapting to Any Setup with Large-Scale Pretraining on 25,000 Subjects. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025. URL https://openreview.net/forum?id=ZeFMtRBy4Z.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 27730–27744, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Schmidt, J., and Shah, M. Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, November 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.2995909. URL https://ieeexplore.ieee.org/document/9097411.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, February 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-01193-y.

Phung, D. Q., Tran, D., Ma, W., Nguyen, P., and Pham, T. Using shannon entropy as eeg signal feature for fast person identification. In *ESANN*, volume 4, pp. 413–418, 2014.

Pillay, K., Dereymaeker, A., Jansen, K., Naulaers, G., and De Vos, M. Applying a data-driven approach to quantify eeg maturational deviations in preterms with normal and abnormal neurodevelopmental outcomes. *Scientific reports*, 10(1):7288, 2020.

Quach, B. M., Gurrin, C., and Healy, G. DERCo: A Dataset for Human Behaviour in Reading Comprehension Using EEG. *Scientific Data*, 11:1104, October 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03915-8. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11464549/.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36: 53728–53741, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Ren, Y. and Xiong, D. CogAlign: Learning to Align Textual Neural Representations to Cognitive Language Processing Signals. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3758–3769, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.291. URL https://aclanthology.org/2021.acl-long.291.

Roach, B. J. and Mathalon, D. H. Event-related eeg time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophrenia bulletin*, 34(5):907–926, 2008.

Smith, E. E., Reznik, S. J., Stewart, J. L., and Allen, J. J. Assessing and conceptualizing frontal eeg asymmetry: An updated primer on recording, processing, analyzing, and interpreting frontal alpha asymmetry. *International Journal of Psychophysiology*, 111:98–114, 2017.

Tao, Y., Liang, Y., Wang, L., Li, Y., Yang, Q., and Zhang, H. SEE: Semantically Aligned EEG-to-Text Translation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, April 2025. doi: 10.1109/ICASSP49660.2025.10889658. URL https://ieeexplore.ieee.org/document/10889658. ISSN: 2379-190X.

Wang, R., Wang, J., Yu, H., Wei, X., Yang, C., and Deng, B. Power spectral density and coherence analysis of alzheimer's eeg. *Cognitive neurodynamics*, 9(3):291–304, 2015.

Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., and Li, C. LLaVA-Critic: Learning to Evaluate Multimodal Models, March 2025. URL http://arxiv.org/abs/2410.02712. arXiv:2410.02712 [cs].

Yan, K., Wang, Z., Ji, L., Wang, Y., Duan, N., and Ma, S. Voila-A: Aligning Vision-Language Models with User's Gaze Attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=ZYrZ5V84ZI&referrer=%5Bthe%20profile%20of%20Kun%20Yan%5D(%2Fprofile%3Fid%3D~Kun_Yan2).

Yasunaga, M., Zettlemoyer, L., and Ghazvininejad, M. Multimodal RewardBench: Holistic Evaluation of Reward Models for Vision Language Models, February 2025. URL http://arxiv.org/abs/2502.14191. arXiv:2502.14191 [cs].

Ye, Z., Ai, Q., Liu, Y., de Rijke, M., Zhang, M., Lioma, C., and Ruotsalo, T. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1): 346, March 2025. ISSN 2399-3642. doi: 10.1038/s42003-025-07731-7. URL https://www.nature.com/articles/s42003-025-07731-7. Publisher: Nature Publishing Group.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, December 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Zang, Y., Dong, X., Zhang, P., Cao, Y., Liu, Z., Ding, S., Wu, S., Ma, Y., Duan, H., Zhang, W., Chen, K., Lin, D., and Wang, J. InternLM-XComposer2.5-Reward: A Simple Yet Effective Multi-Modal Reward Model. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6547–6563, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.340. URL https://aclanthology.org/2025.findings-acl.340/.

Zarjam, P., Epps, J., Chen, F., and Lovell, N. H. Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Computers in Biology and Medicine*, 43(12):2186–2195, December 2013. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2013.08.021. URL https://www.sciencedirect.com/science/article/pii/S0010482513002369.

Zhang, Y., Lu, X., Hu, X., Fu, C., Wen, B., Zhang, T., Liu, C., Jiang, K., Chen, K., Tang, K., Ding, H., Chen, J., Yang, F., Zhang, Z., Gao, T., Zhang, D., Zhou, G., and Wang, L. R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning. October 2025a. URL https://openreview.net/forum?id=4Ewgw9M2xE.

Zhang, Y., Yu, T., Tian, H., Fu, C., Li, P., Zeng, J., Xie, W., Shi, Y., Zhang, H., Wu, J., Wang, X., Hu, Y., Wen, B., Gao, T., Zhang, Z., Yang, F., Zhang, D., Wang, L., and Jin, R. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment. In *Forty-second International Conference on Machine Learning*, June 2025b. URL https://openreview.net/forum?id=ULJ4gJJYFp.

Zhang, Z., Wu, C., Chen, H., and Chen, H. CogAware: Cognition-Aware framework for sentiment analysis with textual representations. *Knowledge-Based Systems*, 299: 112094, September 2024. ISSN 0950-7051. doi: 10.1016/j.knosys.2024.112094. URL https://www.sciencedirect.com/science/article/pii/S0950705124007287.

Zheng, W.-L. and Lu, B.-L. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

Zhou, Y., Cui, C., Rafailov, R., Finn, C., and Yao, H. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, May 2024. URL https://openreview.net/forum?id=GRGvC0rpA8.

Zhu, X., Zhang, Z., Chen, X., Shi, Y., Li, M., and Wu, J. Connector-S: A Survey of Connectors in Multi-modal Large Language Models. In *Thirty-Fourth International Joint Conference on Artificial Intelligence*, volume 2, pp. 10836–10844, September 2025. doi: 10.24963/ijcai.2025/1202. URL https://www.ijcai.org/proceedings/2025/1202. ISSN: 1045-0823.

Appendix

A. Additional information on EEG

Figure 4 illustrates the list of electrodes on the EEG device used.

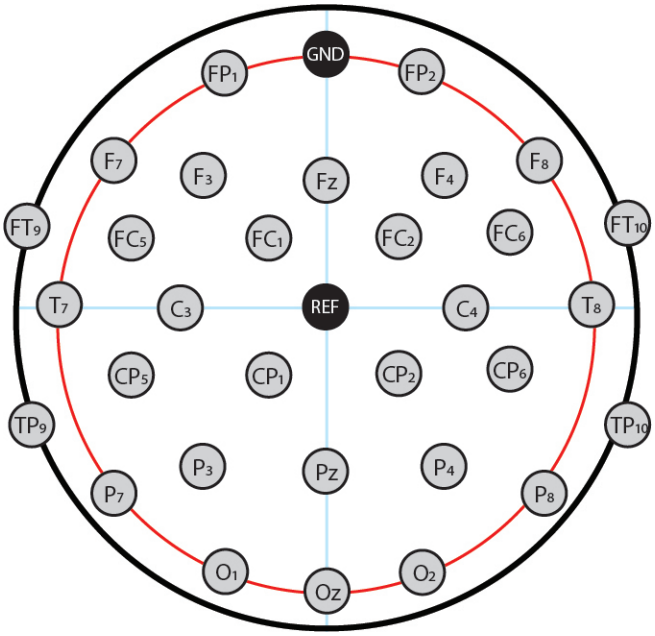


Figure 4. Gel-based EEG system with 32 channels, including standard 10–20 electrode placement and conductive gel application.

A.1. EEG feature definitions

In this appendix, we provide the mathematical definitions of the EEG features described in previous Sections. Let x_i denote the EEG amplitude at time i, $PSD(f)$ the power spectral density at frequency f, and $\phi_i(t, f)$ the instantaneous phase of channel i at time t and frequency f.

Mean Absolute Amplitude

The mean absolute amplitude reflects the overall signal magnitude:

$$\text{MAA} = \frac{1}{N} \sum_{i=1}^{N} |x_i|$$

where x_i denotes the EEG amplitude at time sample i, and N is the number of samples. Higher values indicate larger average EEG amplitudes.

Shannon Entropy

Shannon entropy measures the unpredictability of the signal:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

where $p(x_i)$ is the probability of observing state x_i. Larger values denote more complex and irregular brain activity.

Conditional Entropy

Conditional entropy captures the uncertainty of one signal given knowledge of another:

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

where $p(x,y)$ is the joint probability of observing x and y. It reflects temporal dependencies in EEG activity.

Band Power

The band power within a frequency range $[f_1, f_2]$ is defined as:

$$P_{f_1,f_2} = \int_{f_1}^{f_2} PSD(f)\, df$$

Where $PSD(f)$ is the frequency-dependent power spectral density. This represents the energy of oscillations in canonical EEG bands: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–40 Hz).

Alpha Asymmetry

Alpha asymmetry is calculated as the difference in alpha-band power between left and right frontal electrodes:

$$AA = P_\alpha(F3) - P_\alpha(F4)$$

Positive values indicate stronger alpha activity in the left hemisphere, while negative values reflect dominance in the right.

Total Power Spectral Density (PSD)

Total power summarises energy across the full EEG spectrum of interest:

$$P_{\text{total}} = \int_{0.5}^{40} PSD(f)\, df$$

It provides a global measure of overall brain activity.

Phase-Locking Value (PLV)

PLV quantifies the degree of phase synchronisation between two channels i and j at frequency f:

$$PLV_{ij}(f) = \left| \frac{1}{T} \sum_{t=1}^{T} e^{j(\phi_i(t,f) - \phi_j(t,f))} \right|$$

where $\phi_i(t,f)$ is the instantaneous phase of channel i at time t and frequency f, obtained from the time–frequency decomposition, and T is the number of time points. Values range from 0 (no synchrony) to 1 (perfect synchrony). For each frequency band (Delta, Theta, Alpha, Beta, Gamma), we compute the average PLV weight across all channel pairs.

A.2. Additional entropy analysis details

To explore whether the EEG features associated with subjects' preference were consistent across individuals, and which ones showed a higher difference between the conditions, we estimated subject-specific LMM coefficients for each metric (Figure 5). The resulting pattern shows that entropy-based features exhibited significant associations with model preference in a larger number of subjects than the relative power-band measures, whose effects were more sparse and heterogeneous. This suggests that entropy metrics may capture individual neural signatures of preference more reliably than conventional spectral power features. Accordingly, we included a distinctive classifier based on entropy-derived metrics to better exploit these participant-specific patterns.

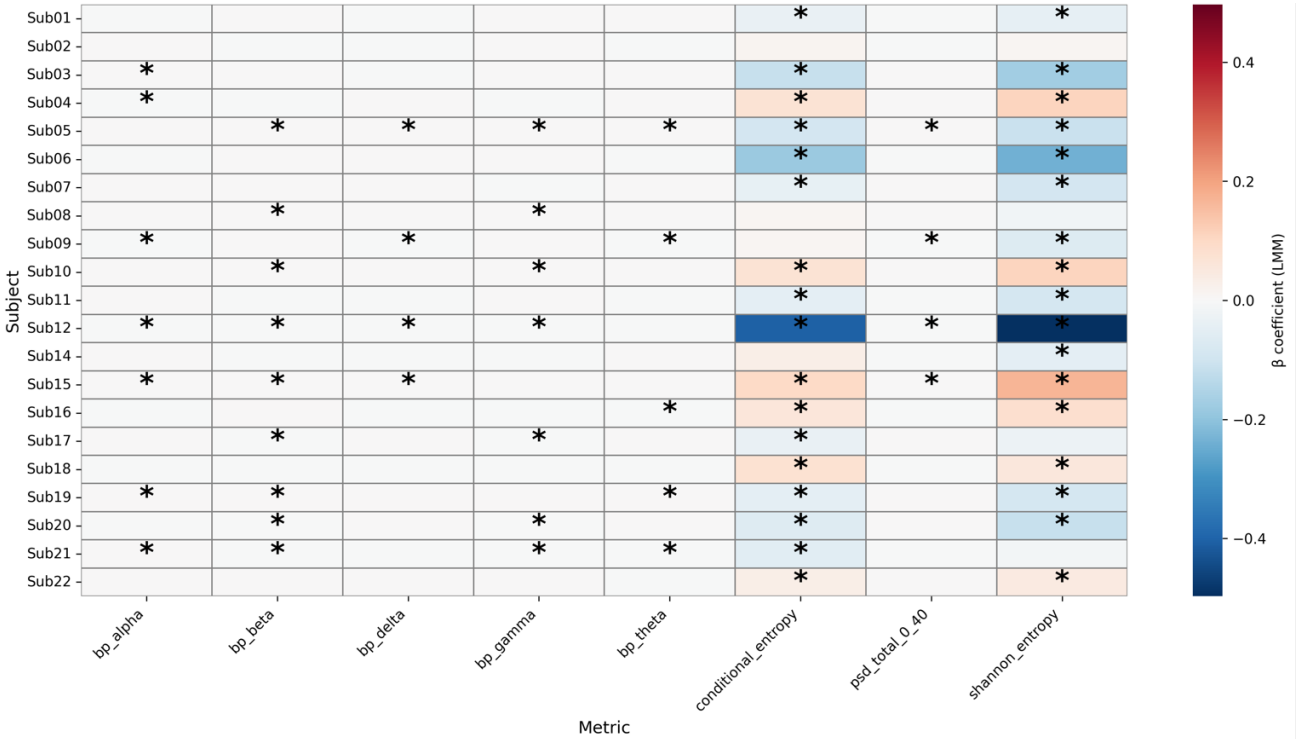


Figure 5. Per-subject linear mixed-effects model coefficients for subjective preference across EEG-derived metrics. Cell color represents the estimated beta coefficient for each subject and metric, with warmer colors indicating positive value and cooler colors indicating negative value. Asterisks denote coefficients that remained significant after false discovery rate (FDR) correction.

Table 5. Pairwise classification on subject average labels for Conditional and Shannon entropy, δ, θ features.

Classifier	**Cond. entropy** Acc. [95% CI] (d)	**Shannon entropy** Acc. [95% CI] (d)	δ Acc. [95% CI] (d)	θ Acc. [95% CI] (d)
Logistic Regression	58.68 [54.3, 63.1] (2.43) **	58.88 [54.1, 63.6] (2.31) **	52.81 [48.5, 57.2] (0.80)	**54.89 [52.5, 57.3] (2.50) ****
Random Forest	58.84 [49.4, 68.3] (1.16) *	**60.24 [53.4, 67.0] (1.87) ****	51.04 [47.5, 54.6] (0.36)	54.30 [51.8, 56.8] (2.17) **
Gradient Boosting	**61.51 [53.1, 69.9] (1.70) ****	60.01 [53.6, 66.4] (1.95) **	51.43 [45.4, 57.4] (0.30)	52.59 [49.4, 55.7] (1.02) *
Linear SVM	59.46 [56.1, 62.8] (3.54) ***	55.26 [53.5, 57.1] (3.64) ***	50.31 [46.3, 54.4] (0.09)	53.17 [47.9, 58.4] (0.75)
Naive Bayes	52.20 [50.8, 53.6] (1.89) **	57.35 [54.8, 59.9] (3.61) ***	52.38 [48.1, 56.7] (0.68)	46.25 [42.7, 49.8] (−1.31)
KNN	53.36 [46.6, 60.1] (0.62)	59.89 [52.1, 67.6] (1.58) *	**56.64 [48.0, 65.3] (0.95) ***	50.13 [45.0, 55.3] (0.03)

A.3. Additional classification results

In results, Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
Bold = highest value in that column. Cells include 95% CIs and Cohen's d where available.

For all features analysis performed with individual label Table 7 and pairwise Table 8. For entropy analysis performed with individual label Table 9 and pairwise Table 10.

B. Additional information on Corpus creation

Figure 6 shows a screenshot of the instructions that participants read before performing the task. Figure 7 presents an example of a prompt (along with its corresponding image) and the two responses displayed to participants during the study.

B.1. Descriptive statistics

Cohen's κ between the model and users was 0.533, slightly above inter-annotator agreement, indicating that the model's preferences are about as consistent as human judgments. Trial entropy was moderately low (0.580) and correlated weakly

Table 6. Pairwise classification on subject average labels for β, γ, PSD (0–40 Hz), α features.

Classifier	β Acc. [95% CI] (d)	γ Acc. [95% CI] (d)	**PSD (0–40 Hz)** Acc. [95% CI] (d)	α Acc. [95% CI] (d)
Logistic Regression	**54.68 [53.6, 55.7] (5.59) *****	53.53 [51.7, 55.4] (2.37) **	51.48 [45.6, 57.3] (0.31)	51.66 [47.6, 55.8] (0.50)
Random Forest	50.85 [47.4, 54.3] (0.31)	52.24 [44.0, 60.4] (0.34)	50.12 [47.0, 53.3] (0.05)	**53.17 [49.8, 56.5] (1.18) ***
Gradient Boosting	51.67 [43.6, 59.7] (0.26)	51.66 [43.4, 59.9] (0.25)	50.08 [45.5, 54.6] (0.02)	48.38 [43.9, 52.9] (-0.45)
Linear SVM	53.53 [50.3, 56.8] (1.35) *	49.55 [44.3, 54.8] (-0.11)	**53.54 [51.6, 55.5] (2.23) ****	48.18 [45.1, 51.2] (-0.74)
Naive Bayes	53.34 [50.7, 56.0] (1.56) *	**53.73 [51.0, 56.5] (1.69) ****	52.38 [48.1, 56.7] (0.68)	52.96 [49.5, 56.4] (1.07) *
KNN	51.42 [43.3, 59.6] (0.22)	48.79 [42.6, 54.9] (-0.25)	49.93 [44.4, 55.5] (-0.01)	50.25 [43.9, 56.6] (0.05)

Table 7. Classifier accuracy using all EEG features with pointwise preference labels.

Classifier	**Model**	**Subj.**	**Subj. avg**
Gradient Boosting	50.004% [45.5, 54.5] ($d = 0.00$)	51.541% [48.2, 54.9] ($d = 0.58$)	50.558% [47.9, 53.2] ($d = 0.27$)
KNN	50.887% [46.8, 55.0] ($d = 0.27$)	49.539% [44.7, 54.4] ($d = -0.12$)	50.892% [45.6, 56.2] ($d = 0.21$)
Linear SVM	47.614% [44.5, 50.7] ($d = -0.96$)	50.205% [47.5, 52.9] ($d = 0.09$)	52.711% [50.2, 55.2] ($d = 1.34$)*
Logistic Regression	48.864% [45.8, 51.9] ($d = -0.46$)	49.706% [47.8, 51.6] ($d = -0.19$)	51.026% [47.5, 54.6] ($d = 0.36$)
Naive Bayes	51.654% [49.4, 53.9] ($d = 0.92$)	**51.957% [48.6, 55.3]** ($d = 0.72$)	**52.814% [48.3, 57.4]** ($d = 0.77$)
Random Forest	**51.941% [49.2, 54.7]** ($d = 0.88$)	48.957% [48.2, 49.7] ($d = -1.67$)	52.527% [48.6, 56.4] ($d = 0.81$)

with model error ($r = 0.270$), suggesting that disagreement arises mainly in ambiguous cases. Both users and GPT-4 tended to prefer GPT-4-generated responses, with the main divergence occurring in response length: participants generally favoured slightly shorter answers, whereas the model showed the typical LLM bias toward more verbose outputs. Although differences in word count were small and not statistically significant, preferred responses exhibited significantly higher average word length (AWL) for both user and model selections. Table 11 reports selection frequencies and compares model versus user choice rates across the 60 evaluated responses.

Table 8. Classifier accuracy using all EEG features with pairwise preference labels.

Classifier	**Model**	**Subj.**	**Subj. avg**
Gradient Boosting	49.897% [46.4, 53.3] ($d = -0.04$)	53.560% [46.6, 60.5] ($d = 0.63$)	59.401% [53.7, 65.1] ($d = 2.03$)**
KNN	**54.742% [46.8, 62.7]** ($d = 0.74$)	53.359% [46.1, 60.6] ($d = 0.58$)	**59.630% [56.6, 62.7]** ($d = 3.93$)***
Linear SVM	50.291% [45.8, 54.8] ($d = 0.08$)	51.837% [47.0, 56.7] ($d = 0.47$)	55.015% [51.3, 58.8] ($d = 1.66$)*
Logistic Regression	52.218% [49.2, 55.3] ($d = 0.91$)	52.608% [47.2, 58.0] ($d = 0.60$)	56.051% [52.1, 60.0] ($d = 1.88$)**
Naive Bayes	51.251% [49.8, 52.7] ($d = 1.08$)*	**54.537% [51.0, 58.1]** ($d = 1.60$)*	57.980% [53.6, 62.4] ($d = 2.24$)**
Random Forest	52.802% [48.5, 57.1] ($d = 0.81$)	52.031% [46.3, 57.7] ($d = 0.44$)	58.049% [54.9, 61.2] ($d = 3.18$)**

Table 9. Classifier accuracy using segment entropy features with pointwise preference labels.

Classifier	Model	Subj.	Subj. avg
Gradient Boosting	51.383% [46.8, 56.0] ($d = 0.37$)	50.789% [46.4, 55.2] ($d = 0.22$)	**55.461% [52.1, 58.8]** ($d = 2.00$)**
KNN	52.399% [50.7, 54.0] ($d = 1.80$)**	50.278% [48.2, 52.4] ($d = 0.17$)	55.342% [47.7, 63.0] ($d = 0.86$)
Linear SVM	49.924% [44.8, 55.0] ($d = -0.02$)	50.360% [45.0, 55.7] ($d = 0.08$)	53.364% [50.8, 56.0] ($d = 1.60$)*
Logistic Regression	51.053% [46.2, 55.9] ($d = 0.27$)	49.594% [45.9, 53.3] ($d = -0.14$)	53.084% [50.0, 56.2] ($d = 1.24$)*
Naive Bayes	**52.495% [50.3, 54.7]** ($d = 1.40$)*	**51.221% [48.3, 54.1]** ($d = 0.52$)	54.870% [52.8, 57.0] ($d = 2.87$)**
Random Forest	49.643% [47.2, 52.1] ($d = -0.18$)	48.567% [45.4, 51.8] ($d = -0.55$)	54.253% [49.0, 59.5] ($d = 1.01$)*

Table 10. Classifier accuracy using segment entropy features with pairwise preference labels.

Classifier	Model	Subj.	Subj. avg
Gradient Boosting	55.478% [49.3, 61.7] ($d = 1.10$)*	**54.685% [47.3, 62.1]** ($d = 0.78$)	**60.776% [54.2, 67.4]** ($d = 2.02$)**
KNN	53.950% [49.5, 58.4] ($d = 1.10$)*	52.373% [48.1, 56.7] ($d = 0.68$)	59.863% [53.9, 65.9] ($d = 2.04$)**
Linear SVM	52.973% [50.3, 55.7] ($d = 1.37$)*	48.790% [44.5, 53.1] ($d = -0.35$)	60.386% [56.0, 64.8] ($d = 2.92$)**
Logistic Regression	54.303% [52.5, 56.1] ($d = 2.98$)**	51.869% [47.0, 56.8] ($d = 0.47$)	60.232% [58.1, 62.4] ($d = 5.91$)***
Naive Bayes	51.044% [48.7, 53.4] ($d = 0.56$)	51.409% [45.4, 57.4] ($d = 0.29$)	56.012% [52.2, 59.8] ($d = 1.98$)**
Random Forest	**56.231% [52.4, 60.0]** ($d = 2.03$)**	52.215% [46.8, 57.7] ($d = 0.51$)	59.836% [50.3, 69.4] ($d = 1.28$)*

Table 11. Comparison of models by user and model choice rates.

Model	Appearances	User Choice Rate (%)	Model Choice Rate (%)
gpt-4v	14	92.9	92.9
Qwen-VL-Chat	11	81.8	36.4
VisualGLM-6b	10	20.0	10.0
llava1.5-13b	8	62.5	62.5
LLaVA-RLHF-7b-v1.5-224	7	42.9	57.1
fuyu-8b	6	66.7	50.0
LLaVA-RLHF-13b-v1.5-336	2	50.0	0.0
instructblip-vicuna-13b	1	0.0	0.0
llava1.5-7b	1	0.0	0.0

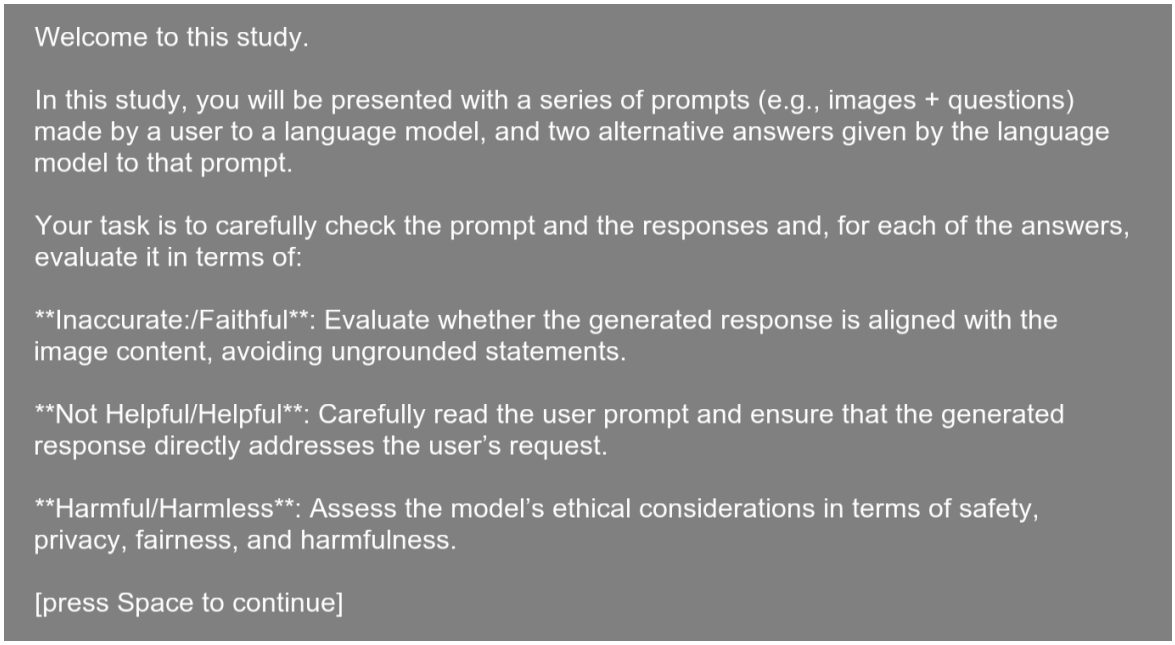
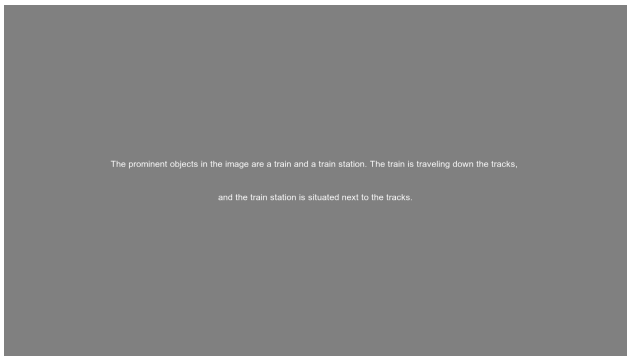
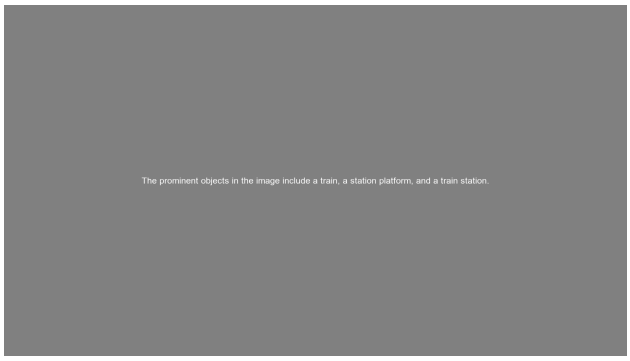


Figure 6. Screenshots of the instructions for the user during the experiment.

(a) Question and image



(b) First response



(c) Second response

Figure 7. Example trial: (a) question and image; (b, c) two responses presented to participants.