IESSEEG: AN OPEN EEG DATASET TOWARDS BETTER UNDERSTANDING OF INFANTILE EPILEPTIC SPASMS SYNDROME

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce IESSEEG, the first open electroencephalography (EEG) dataset dedicated to infantile epileptic spasms syndrome (IESS). IESSEEG comprises approximately 3,172 hours of EEG from 100 pediatric participants, together with diagnosis and treatment-response labels, clinician annotations, Burden of Amplitudes and Epileptiform Discharges (BASED) scores, and clinical metadata. We establish shared benchmarks for diagnosis and treatment-response prediction, comparing traditional quantitative EEG (qEEG) methods with eight EEG foundation models. Both traditional qEEG methods and foundation models achieve strong performance on the diagnosis task, while treatment-response prediction remains more challenging and benefits less consistently from fine-tuning. Further analyses suggest that combining qEEG features, learned representations, and clinical metadata can improve treatment-response prediction. By making these data available with shared benchmarks, IESSEEG enables the community to investigate EEG characteristics associated with IESS and develop more effective methods for diagnosis and treatment-response prediction.

1 INTRODUCTION

Infantile epileptic spasms syndrome (IESS) is an epilepsy syndrome of infancy characterized by epileptic spasms, which often occur in clusters. Onset is usually between 1 and 24 months of age, most often between 3 and 12 months. Affected children frequently experience developmental slowing or regression (Zuberi et al., 2022). Electroencephalography (EEG) is used to establish the diagnosis and assess treatment response (Ramantani et al., 2022). The rarity of IESS limits the size of research cohorts, while its diverse causes complicate the development and validation of EEG biomarkers (Romero Milà et al., 2022). The lack of publicly available IESS EEG data has further limited the development and comparison of computational methods.

We introduce IESSEEG, the first open EEG dataset dedicated to IESS. It contains 150 original long EEG recordings totaling approximately 3,172 hours from 50 cases studied before and after treatment and 50 controls. The collection also includes clinician-selected EEG segments from sleep and wakefulness, as well as 30-minute simulated routine EEG segments with clinician diagnosis labels. Recording-level diagnosis and treatment-response labels, clinical metadata, and Burden of Amplitudes and Epileptiform Discharges (BASED) scores support research on diagnosis and treatment response within the same cohort.

An important motivation for computational EEG analysis in IESS is the need for more consistent clinical assessment. Clinicians rely on visual interpretation, and although the BASED score provides defined scoring criteria, its application remains subject to inter-rater variability (Mytinger et al., 2021; 2025; Swarz et al., 2023). Disagreement about abnormalities such as hypsarrhythmia can adversely affect clinical care (Hussain et al., 2015). Quantitative EEG (qEEG) methods offer numerical measures of EEG characteristics to complement visual assessment. Studies using these features have demonstrated their potential for both IESS diagnosis and treatment-response prediction (Smith et al., 2021; Rajaraman et al., 2024).

EEG foundation models offer another approach by learning representations from large EEG collections that transfer to downstream tasks (Yang et al., 2023; Jiang et al., 2024; Wang et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Their performance on clinical EEG classification and recent application to seizure prediction in IESS motivate evaluating their potential for diagnosis and treatment-response prediction (Zheng et al., 2025). A shared benchmark is needed to assess these applications alongside established qEEG methods and investigate whether learned representations capture additional predictive information.

We evaluate traditional qEEG and foundation models for IESS diagnosis and for response prediction from pre-treatment or post-treatment EEG, including the effect of fine-tuning on foundation-model performance. For response prediction, we further ask whether whole-recording qEEG provides information that supplements short-window representations. Building on the association reported by Rajaraman et al. (2024), we also examine the additional predictive value of lead time, the interval from disease onset to the first EEG.

Our contributions are summarized below.

1. **The first open EEG dataset for IESS.** We release IESSEEG, comprising original long EEG recordings and derived EEG segments with clinical labels, annotations, and participant metadata. It enables the research community to develop and compare methods for IESS diagnosis and treatment-response prediction on the same clinically characterized cohort.
2. **A benchmark for IESS diagnosis and treatment-response prediction.** We benchmark traditional qEEG methods and eight EEG foundation models on IESSEEG for diagnosis and response prediction from pre-treatment and post-treatment EEG. We evaluate foundation models with frozen encoders and full fine-tuning, establishing reference results for future methods.
3. **Insights into treatment-response prediction.** Exploratory analyses suggest that whole-recording qEEG features and clinical metadata provide complementary information to foundation-model representations for pre-treatment response prediction. These findings motivate further development of models that capture EEG characteristics over longer temporal contexts and integrate clinical metadata.

2 Related Work

BASED score. The BASED score is a visual EEG grading system developed for infantile spasms. It grades interictal abnormalities according to background amplitude and epileptiform activity to assess EEG severity and electrographic remission (Mytinger et al., 2021). However, scoring requires time-consuming visual review by clinicians and remains subject to inter-rater variability (Fan et al., 2022; Mytinger et al., 2025; Swarz et al., 2023). To reduce this workload, Fan et al. (2022) developed a convolutional network for automated BASED scoring, but the model was not publicly released.

EEG datasets and clinical benchmarks. Public scalp EEG datasets support research ranging from detecting abnormal activity to predicting clinical outcomes. The Temple University Hospital EEG corpus and its subsets provide benchmarks for abnormality classification and seizure detection (Obeid & Picone, 2016; Shah et al., 2018). More specifically, for pediatric epilepsy, the Children's Hospital Boston–Massachusetts Institute of Technology (CHB-MIT) database contains recordings from patients with intractable seizures, with reported ages ranging from 1.5 to 22 years, and provides seizure onset and offset annotations (Guttag, 2010). For newborns, the Helsinki neonatal seizure dataset contains EEG from 79 neonates evaluated for suspected seizures at 35–45 weeks postmenstrual age, with seizures independently annotated by three clinicians (Stevenson et al., 2019). The Cork neonatal EEG dataset focuses on hypoxic-ischaemic encephalopathy and provides background-severity grades for EEG recorded from 53 newborns within the first 48 hours after birth (O'Toole et al., 2023). For neurological outcome prediction, the International Cardiac Arrest Research consortium (I-CARE) database links continuous EEG from adults who remained comatose after cardiac arrest to their subsequent neurological outcomes (Amorim et al., 2023). However, IESS research has lacked an open EEG dataset and a shared benchmark. IESSEEG addresses this gap by providing an open EEG dataset and a common benchmark for IESS diagnosis and treatment-response prediction.

Computational methods for IESS diagnosis and treatment response. Computational methods have been developed to support IESS diagnosis and treatment-response prediction. Using pre-treatment interictal EEG, Smith et al. (2021) evaluated amplitude, spectral content, entropy, long-range temporal correlations, and functional connectivity as features for IESS diagnosis. A logistic regression model combining selected features achieved an area under the receiver operating characteristic curve

(AUROC) of 0.96 in an age-matched analysis. On the same pre-treatment EEG collection, Miyakoshi et al. (2021) evaluated phase–amplitude coupling for diagnosis, reporting AUROCs above 0.98 during sleep and 0.845 during wakefulness after automated artifact removal. Lu et al. (2023) developed a convolutional neural network that achieved diagnostic accuracies of 95% with sleep EEG and 91% with awake EEG. For treatment response, Rajaraman et al. (2024) analyzed qEEG features from the 400 clinician-selected EEG segments from sleep and wakefulness recorded before and after treatment from the 50 cases in IESSEEG and reported leave-one-out AUROCs of 0.69 for prediction from pre-treatment EEG and 0.91 for distinguishing responders from non-responders using post-treatment EEG. However, the diagnostic performance of some features varied between sleep and wakefulness (Smith et al., 2021), and the accuracy of treatment-response prediction from pre-treatment EEG remained limited (Rajaraman et al., 2024).

Pretrained EEG representations. Recent EEG foundation models learn representations from large scalp EEG collections that can be adapted to multiple downstream tasks (Yang et al., 2023; Jiang et al., 2024; Wang et al., 2025; 2024). Evaluations of Large Brain Model (LaBraM) and Criss-Cross Brain Foundation Model (CBraMod) have demonstrated transfer to abnormality classification, seizure-related tasks, and other EEG classification problems (Jiang et al., 2024; Wang et al., 2025). LUNA and REVE accommodate different electrode configurations (Döner et al., 2025; El Ouahidi et al., 2025), while CodeBrain and CSBrain use multi-scale architectures to learn EEG representations (Ma et al., 2026; Zhou et al., 2025). In the Fractional Amplitude-Phase Expressor (FAPEX) study, Zheng et al. (2025) evaluated pretrained models for seizure prediction in a private cohort of 17 patients with IESS, reporting median AUROCs of 0.762 for CBraMod and 0.714 for EEGPT. These promising seizure-prediction results encourage further investigation of foundation models for IESS diagnosis and treatment-response prediction.

3 The IESSEEG Dataset

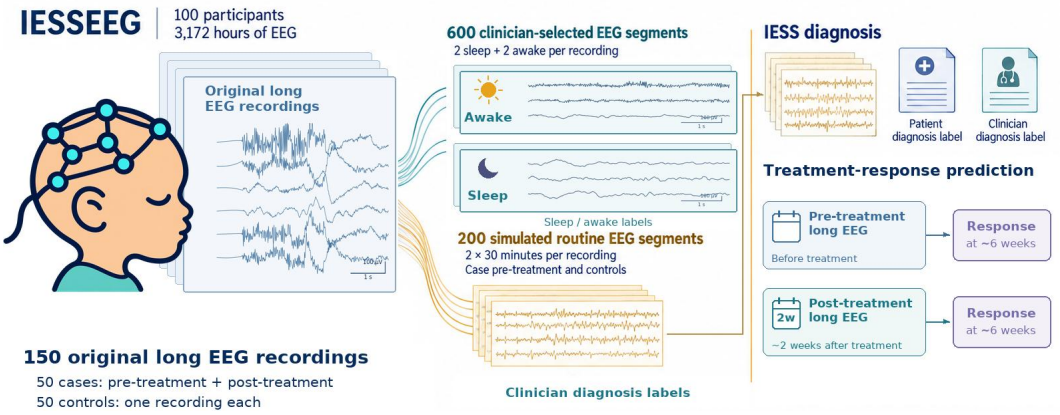


Figure 1: Overview of IESSEEG, its EEG subsets, and benchmark tasks.

3.1 Participants

IESSEEG comprises 100 pediatric participants: 50 cases and 50 controls. The cohort was selected from examinations performed at Mattel Children's Hospital, University of California, Los Angeles (UCLA), between February 2014 and July 2018.

Cases were identified through clinical assessment and confirmation of epileptic spasms during the full diagnostic examination; hypsarrhythmia was not required for inclusion. Controls were evaluated for suspected seizures, but their presenting events were judged non-epileptic and their neurological assessments were normal. Cases include both newly presenting and previously treated patients, and controls are not fully age matched.

Individual participant demographics are provided in Appendix A.

Table 1: Participant characteristics. Age at EEG refers to pre-treatment for cases and the source recording for controls. Continuous values are median (interquartile range).

Characteristic	Cases	Controls
Participants	50	50
Age at EEG, years	0.93 (0.64–1.73)	1.52 (0.59–9.06)
Age at EEG, range in years	0.37–8.31	0.10–17.73
Age at spasm onset, months	7.54 (4.89–12.38)	—
Treatment responders	28/50	—

3.2 EEG ACQUISITION AND SEGMENT SELECTION

EEG acquisition. Each case contributes a pre-treatment and a post-treatment long EEG recording; each control contributes one long recording, giving 150 original long EEG recordings in total. The post-treatment EEG was obtained at the first follow-up, approximately two weeks after treatment initiation.

Recordings were acquired with Nihon Kohden systems using the international 10–20 electrode placement system. They contain 19 scalp EEG channels, with A1 and A2 available for reference construction. Additional channels vary across files.

Segment selection. For each case, a clinician selected two sleep and two awake segments from the pre-treatment recording and another two sleep and two awake segments from the post-treatment recording. For each control, the clinician selected two sleep and two awake segments from the single recording. This yielded 600 *clinician-selected EEG segments*, with eight per case and four per control. The 400 case segments in this collection were previously analyzed by Rajaraman et al. (2024). A randomization algorithm specified one time point for each state within the first four hours and another within the last four hours of the recording. The clinician began at the specified time and, if necessary, moved forward until the requested state could be identified confidently. The clinician distinguished sleep from wakefulness using EEG features such as sleep architecture, posterior dominant rhythm, wakefulness-related artifacts, and slow roving eye movements. The target duration was 20 minutes per segment, with actual durations varying across recordings.

Additionally, we constructed a diagnosis evaluation set to simulate the short recording duration of routine outpatient EEG examinations (Burkholder et al., 2016). We sampled two 30-minute segments at random positions in each of the 50 case pre-treatment and 50 control recordings, yielding 200 additional segments. We refer to these as *simulated routine EEG segments*. Sleep and awake labels are available for clinician-selected EEG segments; simulated routine EEG segments have no state labels.

Table 2 summarizes the counts and durations of the long recordings and derived segments.

Table 2: Counts and total durations of the EEG recordings and segments. Durations are measured from EDF headers. Clinician-selected EEG segments and simulated routine EEG segments are derived from the original long EEG recordings.

(a) Number of recordings or segments

Group	Original long EEG recordings	Clinician-selected EEG segments	Simulated routine EEG segments
Case pre-treatment	50	200	100
Case post-treatment	50	200	0
Control	50	200	100
Total	150	600	200

(b) Total duration (hours)

Group	Original long EEG recordings	Clinician-selected EEG segments	Simulated routine EEG segments
Case pre-treatment	1,134.98	78.64	50.03
Case post-treatment	1,021.30	83.63	—
Control	1,016.20	88.26	50.03
Total	3,172.47	250.53	100.06

3.3 CLINICAL ANNOTATIONS

Diagnosis. For all original long EEG recordings and derived segments, the *patient diagnosis label* indicates whether the source participant is a case or a control. The 200 simulated routine EEG segments additionally have a *clinician diagnosis label*, indicating whether the segment was judged consistent with IESS. Clinicians assigned these labels using clinical expertise and visual EEG interpretation, based only on that segment and without access to the patient's diagnosis.

Treatment response. Following Rajaraman et al. (2024), response required cessation of epileptic spasms and resolution of hypsarrhythmia, when present before treatment, at the first follow-up approximately two weeks after treatment initiation, with no recurrence of spasms for at least one month thereafter. The resulting *treatment-response label* indicates response assessed approximately six weeks after treatment initiation. Among the 50 cases, 28 were responders and 22 were non-responders.

BASED scores. Seven clinicians, four with expertise in pediatric EEG and three trainees, independently assessed 20 original long EEG recordings from 20 different patients: 10 pre-treatment and 10 post-treatment recordings. Each clinician assigned a 2021 BASED score to every recording, yielding 140 individual ratings (Swarz et al., 2023).

3.4 DATA AVAILABILITY

IESSEEG v1.1 is available on Hugging Face[1]. The release includes 150 original long EEG recordings, 600 clinician-selected EEG segments, and 200 simulated routine EEG segments, along with associated annotations, labels, and metadata, including participant demographics, sleep and awake labels, diagnosis and treatment-response labels, clinician judgments, and BASED scores. EEG files are provided in European Data Format (EDF), organized according to the EEG extension of the Brain Imaging Data Structure (BIDS). The accompanying code is available on GitHub[2].

4 THE IESSEEG BENCHMARK

4.1 TASKS

The benchmark covers IESS diagnosis and treatment-response prediction from EEG acquired before or after treatment.

4.1.1 DIAGNOSIS TASK

IESS diagnosis. Methods predict the probability of IESS from a 30-minute simulated routine EEG segment. For methods that require training, we use clinician-selected EEG segments from cases' pre-treatment recordings and from control recordings, with patient diagnosis labels as supervision. Testing uses the 200 simulated routine EEG segments from held-out patients. Each segment receives one prediction, which is evaluated separately against the patient diagnosis label and the clinician diagnosis label defined in Section 3.

4.1.2 RESPONSE PREDICTION TASK

Pre-treatment response prediction. Methods predict treatment response from the original pre-treatment long EEG recordings.

Post-treatment response prediction. Methods predict the same outcome from the original post-treatment long EEG recordings.

Both tasks include the same 50 cases and use the treatment-response labels defined in Section 3. Methods that require training use the corresponding recordings from training patients, with treatment-response labels as supervision. At test time, each recording from a held-out patient receives one response probability, evaluated against that patient's treatment-response label.

[1] https://huggingface.co/datasets/Capur/IESSEEG

[2] https://github.com/IESSEEG/IESSEEG

4.2 METHODS

4.2.1 TRADITIONAL QUANTITATIVE EEG METHODS

We fit logistic regression classifiers to task-specific qEEG features and evaluate their predicted probabilities. For IESS diagnosis, we evaluate two feature sets adapted from the measures studied by Smith et al. (2021). The *qEEG triplet* contains beta-band detrended fluctuation analysis (DFA) intercept, beta-band Shannon entropy, and delta-band phase lag index (PLI) connectivity. The 12-feature *Expanded qEEG* set adds peak-to-peak amplitude averaged across channels and at Cz, log power in the delta, theta, alpha, and beta bands, 95% spectral edge frequency, beta-band permutation entropy at Cz, and beta-band DFA exponent to the triplet.

For treatment-response prediction, we use the *response qEEG* feature set adapted from Rajaraman et al. (2024). It contains beta-band DFA intercept, beta-band Shannon entropy, and delta-band PLI connectivity, with response-specific definitions given in Appendix B.

4.2.2 EEG FOUNDATION MODELS

We evaluate Biosignal Transformer (BIOT), LaBraM, CBraMod, EEGPT, LUNA, REVE, CodeBrain, and CSBrain for IESS diagnosis and both response prediction tasks (Yang et al., 2023; Jiang et al., 2024; Wang et al., 2025; 2024; Döner et al., 2025; El Ouahidi et al., 2025; Ma et al., 2026; Zhou et al., 2025). These models encode short EEG windows lasting seconds to tens of seconds. We divide longer EEG inputs into windows and encode each window separately, following other clinical EEG benchmarks (Kastrati et al., 2025).

For diagnosis, every training window receives its source participant's patient diagnosis label. For both response prediction tasks, every training window receives its source participant's treatment-response label. Assigning recording-level labels to individual windows follows established approaches to clinical EEG classification (van Leeuwen et al., 2019; Zhu et al., 2023). We evaluate each pretrained model using two training strategies with this window-level supervision. With a frozen encoder, the pretrained weights remain fixed while we train a classification head under the supervision of the assigned window labels. For full fine-tuning, we initialize the encoder with the pretrained weights and train it jointly with the classification head under the supervision of the same window labels.

Window prediction and aggregation. At inference, we divide each input EEG into windows and pass each window through the encoder and classification head to obtain a predicted probability. We then average these probabilities across all windows to obtain the prediction for the input EEG, following Zhu et al. (2023). The same procedure applies to frozen and fully fine-tuned models.

Details of the qEEG features, preprocessing procedures for both method families, and foundation-model configurations are provided in Appendix B.

4.3 TRAINING AND VALIDATION

Training objective. For EEG foundation models, we minimize cross-entropy against the assigned patient diagnosis label or treatment-response label.

Cross-validation. We use fixed, stratified five-fold cross-validation with patient-disjoint folds. The fold assignments are provided with IESSEEG. Training and validation details are described in Appendix B.4.

4.4 EVALUATION

Metrics. We report balanced accuracy and the mean AUROC across the five test folds. In the main benchmark, balanced accuracy uses a fixed probability threshold of 0.5 for both diagnosis and response prediction. In the main benchmark tables, values are percentages and brackets denote 95% confidence intervals. AUROC confidence intervals follow the cross-validated influence-function method of LeDell et al. (2015); balanced-accuracy intervals use a delta-method influence function. Both calculations group observations by patient (Appendix B.6).

5 RESULTS

5.1 IESS DIAGNOSIS

Both traditional qEEG methods and EEG foundation models achieve high AUROC for IESS diagnosis (Table 3). The qEEG triplet achieves AUROCs of 91.4% against patient diagnosis labels and 90.7% against clinician diagnosis labels, increasing to 93.1% and 94.7% with Expanded qEEG.

Foundation models perform well even with frozen encoders. Full fine-tuning further improves AUROC for all eight models against patient diagnosis labels and seven of eight against clinician diagnosis labels. After fine-tuning, AUROC ranges from 93.9% to 97.6% and from 91.4% to 97.4%, respectively. LaBraM achieves the highest AUROC under both label definitions in both training settings. Among fully fine-tuned models, LaBraM also achieves the highest balanced accuracy against patient diagnosis labels (93.5%), while CSBrain achieves the highest against clinician diagnosis labels (92.7%).

Full diagnosis results are provided in Appendix C.1, Tables C1 and C2.

Table 3: Model performance on the IESS diagnosis task using simulated routine EEG segments.

	Patient diagnosis label		Clinician diagnosis label	
Method	AUROC	Balanced accuracy	AUROC	Balanced accuracy
Traditional qEEG				
qEEG triplet	91.4 [87.3–95.5]	85.5 [80.1–90.9]	90.7 [85.9–95.4]	83.5 [77.9–89.2]
Expanded qEEG	93.1 [88.8–97.4]	85.5 [79.7–91.3]	94.7 [91.3–98.1]	86.2 [80.7–91.7]
Frozen encoder				
BIOT	87.6 [81.5–93.7]	81.5 [75.5–87.5]	88.4 [82.9–94.0]	81.5 [75.7–87.3]
LaBraM	94.8 [91.3–98.3]	86.5 [80.6–92.4]	97.3 [95.3–99.4]	89.2 [83.9–94.4]
CBraMod	90.7 [85.3–96.1]	84.5 [78.7–90.3]	92.2 [87.8–96.6]	85.1 [79.6–90.6]
EEGPT	94.7 [90.9–98.4]	87.5 [81.7–93.3]	93.2 [88.9–97.5]	84.7 [78.5–90.8]
LUNA	87.5 [81.0–93.9]	84.0 [78.1–89.9]	88.1 [81.6–94.6]	84.7 [78.8–90.5]
REVE	92.8 [88.2–97.3]	87.5 [82.1–92.9]	91.1 [86.3–96.0]	85.5 [79.9–91.1]
CodeBrain	94.6 [91.0–98.1]	88.0 [82.2–93.8]	95.1 [91.9–98.4]	89.9 [84.7–95.1]
CSBrain	87.9 [82.1–93.7]	80.5 [74.4–86.6]	90.2 [84.8–95.6]	81.1 [75.2–86.9]
Full fine-tuning				
BIOT	97.2 [94.9–99.4]	90.5 [86.0–95.0]	95.0 [91.7–98.2]	86.6 [81.2–92.0]
LaBraM	**97.6** [95.0–100.0]	**93.5** [89.2–97.8]	**97.4** [94.9–99.9]	92.5 [88.2–96.7]
CBraMod	96.3 [92.7–99.9]	91.0 [86.1–95.9]	95.2 [91.4–99.0]	90.0 [85.4–94.6]
EEGPT	95.4 [91.8–99.0]	91.0 [86.3–95.7]	92.6 [88.1–97.1]	87.0 [81.4–92.6]
LUNA	93.9 [89.0–98.8]	87.0 [81.5–92.5]	92.4 [87.2–97.7]	85.8 [80.0–91.6]
REVE	94.8 [90.8–98.7]	88.5 [83.0–94.0]	91.4 [86.6–96.3]	84.7 [78.6–90.8]
CodeBrain	96.1 [92.9–99.3]	90.5 [85.6–95.4]	96.2 [93.0–99.5]	90.5 [85.9–95.2]
CSBrain	96.5 [93.6–99.4]	91.0 [86.4–95.6]	96.4 [93.2–99.6]	**92.7** [88.8–96.7]

5.2 TREATMENT-RESPONSE PREDICTION

Treatment-response prediction shows lower AUROC and greater variation across methods than IESS diagnosis (Table 4). qEEG achieves AUROCs of 62.3% from pre-treatment EEG and 77.0% from post-treatment EEG.

Full fine-tuning does not consistently improve AUROC, with increases for four of eight models using pre-treatment EEG and six of eight using post-treatment EEG. Frozen BIOT achieves the highest pre-treatment AUROC (73.0%), while fully fine-tuned CBraMod achieves the highest post-treatment AUROC (79.4%). Frozen LaBraM and frozen CSBrain achieve the highest balanced accuracy for pre-treatment and post-treatment prediction, respectively (69.8% and 66.6%).

Full response prediction results are provided in Appendix C.1, Tables C3 and C4.

Table 4: Model performance on the pre-treatment and post-treatment response prediction tasks using original long EEG recordings.

	Pre-treatment EEG		Post-treatment EEG	
Method	AUROC	Balanced accuracy	AUROC	Balanced accuracy
Traditional qEEG				
Response qEEG	62.3 [46.7–77.9]	61.2 [47.7–74.7]	77.0 [63.1–90.9]	65.3 [52.0–78.5]
Frozen encoder				
BIOT	**73.0** [59.7–86.3]	55.2 [42.8–67.5]	57.8 [43.5–72.0]	61.4 [48.2–74.5]
LaBraM	70.6 [55.4–85.7]	**69.8** [57.0–82.6]	68.6 [54.1–83.1]	64.8 [51.6–77.9]
CBraMod	54.0 [38.7–69.3]	51.8 [37.8–65.7]	75.2 [62.2–88.3]	65.6 [53.3–77.8]
EEGPT	58.6 [42.3–74.9]	56.7 [43.1–70.2]	58.1 [42.4–73.9]	51.8 [37.8–65.7]
LUNA	60.5 [48.8–72.3]	57.6 [43.8–71.4]	58.8 [42.6–75.0]	55.8 [42.0–69.7]
REVE	59.2 [42.7–75.6]	50.0 [36.0–64.0]	59.5 [44.5–74.4]	61.2 [47.7–74.7]
CodeBrain	51.5 [36.2–66.8]	57.8 [44.7–70.9]	65.4 [49.7–81.1]	58.0 [44.9–71.0]
CSBrain	55.0 [40.1–69.8]	51.0 [37.2–64.7]	63.8 [46.8–80.9]	**66.6** [53.7–79.4]
Full fine-tuning				
BIOT	67.9 [53.0–82.9]	54.1 [40.1–68.0]	64.7 [49.7–79.8]	56.2 [42.9–69.4]
LaBraM	68.1 [52.9–83.3]	63.0 [49.6–76.3]	68.7 [54.1–83.2]	60.7 [47.3–74.1]
CBraMod	55.7 [40.1–71.2]	50.3 [36.9–63.7]	**79.4** [66.3–92.6]	64.8 [51.6–77.9]
EEGPT	67.3 [51.2–83.4]	62.5 [49.3–75.7]	62.9 [47.8–78.0]	53.7 [43.8–63.6]
LUNA	59.9 [44.1–75.6]	54.4 [41.0–67.8]	44.8 [30.5–59.0]	47.7 [33.8–61.7]
REVE	57.5 [40.6–74.3]	47.1 [35.0–59.2]	59.6 [44.5–74.8]	55.7 [42.8–68.5]
CodeBrain	59.1 [43.2–75.1]	63.5 [50.0–76.9]	73.7 [59.7–87.6]	57.5 [44.9–70.1]
CSBrain	61.4 [46.2–76.6]	54.1 [40.1–68.0]	59.7 [43.8–75.6]	55.8 [42.0–69.7]

6 Discussion and Analysis

6.1 Sensitivity and Specificity of IESS Diagnosis

The high diagnostic AUROCs in Table 3 show that both traditional qEEG and foundation models can distinguish IESS in this benchmark. Clinical use, however, also requires reliable decisions at a chosen operating threshold. Missed cases can delay treatment, while false-positive predictions can lead to unnecessary investigations and additional clinical review. The consequences of delayed treatment are particularly relevant in IESS, for which clinical guidelines recommend initiating treatment promptly after diagnosis (Ramantani et al., 2022).

At the benchmark threshold of 0.5, fully fine-tuned LaBraM achieves 91.0% sensitivity and 96.0% specificity against patient diagnosis labels, corresponding to nine missed case segments and four false-positive control segments. Fully fine-tuned CSBrain achieves higher specificity (99.0%) with lower sensitivity (83.0%). The complete results in Appendix C.1 show the remaining errors for each method. Improving sensitivity while maintaining high specificity remains an important goal for models intended to assist IESS diagnosis.

6.2 Whole-Recording qEEG for Response Prediction

Treatment-response prediction is more difficult than diagnosis in our benchmark, particularly before treatment (Table 4). The foundation models evaluated here process short EEG windows independently and average their prediction probabilities, without explicitly modeling temporal relationships across windows. Traditional qEEG provides another way to characterize a recording, including DFA measurements of amplitude-envelope fluctuations associated with response (Rajaraman et al., 2024). We therefore examine whether qEEG computed over a complete recording provides predictive information that supplements the learned representations.

For this analysis, we construct a dataset of 500 segments by randomly sampling ten continuous 30-minute segments from each of the 50 pre-treatment long recordings. For each segment, we average the window embeddings from the frozen foundation models BIOT, LaBraM, or CBraMod and reduce

the resulting vector to five dimensions using partial least squares (PLS), fitted with the training patients' treatment-response labels (Wold et al., 2001). We then concatenate the reduced embedding with qEEG features computed from the complete source recording. Logistic regression then predicts response from the reduced embeddings, qEEG, or their concatenation. Experimental details and complete results are provided in Appendix C.2.

Combining whole-recording qEEG with segment embeddings yields higher AUROC than either feature set alone for all three foundation models (Table 5). These results suggest that the two feature sets contain complementary information for pre-treatment response prediction.

These findings motivate extending EEG foundation models to capture information over longer temporal contexts. This could involve architectures that model long EEG sequences or training objectives that encourage representations to capture EEG properties over longer timescales. For example, whole-recording qEEG features could provide auxiliary supervision during fine-tuning.

Table 5: Pre-treatment response AUROC (%) for sampled 30-minute segments. E denotes the five-dimensional representations derived from frozen foundation models, Q denotes whole-recording qEEG, and L denotes log-transformed lead time.

Foundation model	E	Q	Q+L	E+Q	E+L	E+Q+L
BIOT	68.1	62.3	69.6	70.2	71.9	73.3
LaBraM	67.6	62.3	69.6	72.1	71.7	74.3
CBraMod	57.7	62.3	69.6	66.6	68.3	70.9

6.3 Clinical Metadata for Response Prediction

Clinical history provides information beyond the EEG itself. Lead time, the interval from disease onset to the first EEG, is included in the IESSEEG metadata. Following Rajaraman et al. (2024), who identified shorter lead time as a predictor of treatment response in this cohort, we evaluate lead time alongside the EEG features. In our analysis, we append log-transformed lead time to the same logistic classifiers to assess its additional predictive value.

Adding lead time increases AUROC for all three foundation models, both with and without qEEG (Table 5). The combination of LaBraM representations, whole-recording qEEG, and lead time achieves the highest AUROC (74.3%), compared with 72.1% without lead time. These results suggest that clinical history can complement EEG representations and motivate incorporating clinical metadata into response prediction models.

7 Conclusion

We introduced IESSEEG, the first open EEG dataset dedicated to IESS, with original long EEG recordings and clinical labels for benchmarking diagnosis and treatment-response prediction. The benchmark compares traditional qEEG methods with eight EEG foundation models on a shared cohort. Both approaches achieve high diagnostic AUROC, and full fine-tuning improves most foundation models. Treatment-response prediction remains more difficult, with less consistent benefits from fine-tuning. Our exploratory pre-treatment analyses show higher AUROC when segment representations are combined with whole-recording qEEG and lead time, suggesting that whole-recording qEEG and clinical history provide useful complementary information. These findings motivate foundation models that learn from longer EEG context and incorporate clinical metadata. By releasing the recordings, annotations, and benchmark code, we provide a common basis for the community to develop and evaluate methods for analyzing IESS EEG to support diagnosis and treatment-response prediction.

Ethics Statement

The retrospective study received institutional review board approval under protocol 17-000026, with waivers of informed consent and research authorization under the Health Insurance Portability and

Accountability Act (HIPAA). EEG files were de-identified for research use. The models have not been prospectively validated for clinical decisions.

Reproducibility Statement

The dataset and benchmark code are available through the links in Section 3.4. The code includes fixed patient splits and instructions for running the experiments. Method details and additional experimental settings are provided in Appendices B and C.

Artificial Intelligence Assistance

Artificial intelligence (AI) tools were used for literature retrieval, research ideation, drafting and editing text, and assistance with code and analyses.

References

Edilberto Amorim et al. I-CARE: International cardiac arrest REsearch consortium database (version 2.1). PhysioNet, 2023. https://physionet.org/content/i-care/.

David B. Burkholder, Jeffrey W. Britton, Vijayalakshmi Rajasekaran, Rachel R. Fabris, Perumpillichira J. Cherian, Kristen M. Kelly-Williams, Elson L. So, Katherine C. Nickels, Lily C. Wong-Kisiel, Terrence D. Lagerlund, Gregory D. Cascino, Gregory A. Worrell, and Elaine C. Wirrell. Routine vs extended outpatient EEG for the detection of interictal epileptiform discharges. *Neurology*, 86(16):1524–1530, 2016. doi: 10.1212/WNL.0000000000002592.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.52202/085713-2377. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/66969a9e6bd7a26dfeccea7227178ca7-Abstract-Conference.html.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0760. URL https://papers.neurips.cc/paper_files/paper/2025/hash/20a917f77773ac0fa8bea2bdd6606b66-Abstract-Conference.html.

Yuying Fan, Duo Chen, Hua Wang, Yijie Pan, Xueping Peng, Xueyan Liu, and Yunhui Liu. Automatic based scoring on scalp eeg in children with infantile spasms using convolutional neural network. *Frontiers in Molecular Biosciences*, 9:931688, 2022.

John Guttag. Chb-mit scalp eeg database. https://physionet.org/content/chbmit/1.0.0/, jun 2010. URL https://physionet.org/content/chbmit/1.0.0/.

Shaun A Hussain, Grace Kwong, John J Millichap, John R Mytinger, Nicole Ryan, Joyce H Matsumoto, Joyce Y Wu, Jason T Lerner, and Raman Sankar. Hypsarrhythmia assessment exhibits poor interrater reliability: a threat to clinical trial validity. *Epilepsia*, 56(1):77–81, 2015.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47393e8594c82ce8fd83adc672cf9872-Abstract-Conference.html.

Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. EEG-Bench: A benchmark for EEG foundation models in clinical applications, 2025. URL https://arxiv.org/abs/2512.08959.

Erin LeDell, Maya Petersen, and Mark van der Laan. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic Journal of Statistics*, 9(1): 1583–1607, 2015. doi: 10.1214/15-EJS1035.

Mingjian Lu, Yipeng Zhang, Atsuro Diada, Shingo Oana, Rajsekar R Rajaraman, Hiroki Nariai, Vwani Roychowdhury, and Shaun A Hussain. Application of an EEG-based deep learning model to discriminate children with epileptic spasms from normal controls. *medRxiv*, pp. 2023.06.30.23292096, 2023. doi: 10.1101/2023.06.30.23292096.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=msJgEkjwh5.

Makoto Miyakoshi, Hiroki Nariai, Rajsekar R. Rajaraman, Danilo Bernardo, Daniel W. Shrey, Beth A. Lopour, Myung Shin Sim, Richard J. Staba, and Shaun A. Hussain. Automated preprocessing and phase-amplitude coupling analysis of scalp EEG discriminates infantile spasms from controls during wakefulness. *Epilepsy Research*, 178:106809, 2021. doi: 10.1016/j.eplepsyres.2021.106809.

John R Mytinger, Jorge Vidaurre, Melissa Moore-Clingenpeel, Joseph R Stanek, and Dara V Albert. A reliable interictal EEG grading scale for children with infantile spasms – the 2021 BASED score. *Epilepsy Research*, 173:106631, 2021.

John R Mytinger, Dara V F Albert, Shawn C Aylward, et al. A multicenter training and interrater reliability study of the BASED score for infantile epileptic spasms syndrome. *Journal of Clinical Neurophysiology*, 42(3):230–234, 2025. doi: 10.1097/WNP.0000000000001101.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

John M. O'Toole et al. Neonatal EEG graded for severity of background abnormalities in hypoxic-ischaemic encephalopathy. *Scientific Data*, 10:129, 2023.

Rajsekar R Rajaraman, Rachel J Smith, Shingo Oana, Atsuro Daida, Daniel W Shrey, Hiroki Nariai, Beth A Lopour, and Shaun A Hussain. Computational eeg attributes predict response to therapy for epileptic spasms. *Clinical Neurophysiology*, 163:39–46, 2024.

Georgia Ramantani, Bigna K. Bölsterli, Michael Alber, Joerg Klepper, Rudolf Korinthenberg, Gerhard Kurlemann, Daniel Tibussek, Markus Wolff, and Bernhard Schmitt. Treatment of infantile spasm syndrome: Update from the interdisciplinary guideline committee coordinated by the german-speaking society of neuropediatrics. *Neuropediatrics*, 53(6):389–401, 2022. doi: 10.1055/a-1909-2977.

Blanca Romero Milà, Kavyakantha Remakanthakurup Sindhu, John R Mytinger, Daniel W Shrey, and Beth A Lopour. Eeg biomarkers for the diagnosis and treatment of infantile spasms. *Frontiers in Neurology*, 13:960454, 2022.

Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12:83, 2018.

Rachel J Smith, Derek K Hu, Daniel W Shrey, Rajsekar Rajaraman, Shaun A Hussain, and Beth A Lopour. Computational characteristics of interictal eeg as objective markers of epileptic spasms. *Epilepsy Research*, 176:106704, 2021.

Nathan J Stevenson, Karoliina Tapani, Leena Lauronen, and Sampsa Vanhatalo. A dataset of neonatal eeg recordings with seizure annotations. *Scientific Data*, 6(1):190039, 2019. doi: 10.1038/sdata.2019.39.

Jeff Swarz, Shingo Oana, Nassim Naderi, Benjamin Edmonds, Atsuro Daida, Rajsekar Rajaraman, Hiroki Nariai, and Shaun Hussain. Reliability of the 2021 BASED score and hypsarrhythmia for assessment of infantile spasms. American Epilepsy Society Annual Meeting, abstract 2.138, 2023. URL https://aesnet.org/abstractslisting/reliability-of-the-2021-based-score-and-hypsarrhythmia-for-assessment-of-infantile-spasms.

K. G. van Leeuwen, H. Sun, M. Tabaeizadeh, A. F. Struck, M. J. A. M. van Putten, and M. B. Westover. Detecting abnormal electroencephalograms using deep convolutional networks. *Clinical Neurophysiology*, 130(1):77–84, 2019. doi: 10.1016/j.clinph.2018.10.012. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6309707/.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbbd6d915cb90be21c1254a82d45cedd-Abstract-Conference.html.

Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. doi: 10.1016/S0169-7439(01)00155-1.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

Ruizhe Zheng, Lingyan Mao, Dingding Han, Tian Luo, Yi Wang, Jing Ding, and Yuguo Yu. FAPEX: Fractional amplitude-phase expressor for robust cross-subject seizure prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/b7de1556bc4be01966c6bc56fb2da6e0-Abstract-Conference.html.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2915. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7e199ad8ae40eb19b2980f61f659cb07-Abstract-Conference.html.

Yixuan Zhu, Luke J. W. Canham, and David Western. Scope and arbitration in machine learning clinical EEG classification. In *2023 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2023. doi: 10.1109/SPMB59478.2023.10372635. URL https://isip.piconepress.com/conferences/ieee_spmb/2023/papers/l02_01.pdf.

Sameer M. Zuberi, Elaine Wirrell, Elissa Yozawitz, Jo M. Wilmshurst, Nicola Specchio, Kate Riney, Ronit Pressler, Stephane Auvin, Pauline Samia, Edouard Hirsch, Santiago Galicchio, Chahnez Triki, O. Carter Snead, Samuel Wiebe, J. Helen Cross, Paolo Tinuper, Ingrid E. Scheffer, Emilio Perucca, Solomon L. Moshé, and Rima Nabbout. ILAE classification and definition of epilepsy syndromes with onset in neonates and infants: Position statement by the ILAE task force on nosology and definitions. *Epilepsia*, 63(6):1349–1397, 2022. doi: 10.1111/epi.17239.

A PARTICIPANT DEMOGRAPHICS

Table A1: Demographic and treatment-response metadata for all 100 participants. Age at EEG refers to the pre-treatment recording for cases and the source recording for controls. Onset age refers to the onset of infantile spasms. Immediate response indicates whether the patient was clinically considered a responder at the time of the post-treatment EEG, approximately two weeks (around 16 days) after treatment initiation, regardless of subsequent relapse. Response indicates whether the patient responded at approximately six weeks after treatment initiation, as defined in the main text. NaN indicates not applicable.

Participant	Age at EEG (years)	Onset age (months)	Onset to first EEG (months)	Immediate response	Response
UCB020	0.90	8.40	2.07	No	No
UCB035	0.80	3.00	6.08	Yes	Yes
UCB006	1.30	10.70	4.83	Yes	Yes
UCB018	0.40	1.80	2.56	No	No
UCB029	0.60	7.20	0.56	Yes	Yes
UCB045	0.60	7.10	0.00	Yes	Yes
UCB026	0.70	5.50	2.53	Yes	Yes
UCB004	6.30	73.20	2.30	No	No
UCB007	1.70	5.50	15.01	No	No
UCB036	0.60	7.70	0.00	Yes	Yes
UCB002	6.10	68.90	4.57	Yes	Yes
UCB021	1.10	13.60	0.13	Yes	Yes
UCB016	1.60	12.90	5.98	Yes	Yes
UCB037	2.00	24.40	0.00	Yes	Yes
UCB028	0.90	7.00	4.17	Yes	Yes
UCB010	0.50	5.40	0.56	No	No
UCB050	0.40	5.10	0.13	Yes	Yes
UCB015	0.70	8.00	0.07	Yes	Yes
UCB022	1.30	8.50	7.00	Yes	Yes
UCB025	1.40	3.40	12.88	No	No
UCB008	3.60	3.90	39.75	No	No
UCB043	0.70	8.00	0.10	Yes	Yes
UCB030	5.70	65.50	2.46	Yes	Yes
UCB032	2.00	17.20	6.67	No	No
UCB001	8.30	3.00	96.66	Yes	No
UCB019	2.90	33.00	1.28	Yes	Yes
UCB048	1.30	3.00	12.39	Yes	Yes
UCB044	0.50	4.80	0.66	Yes	Yes
UCB038	0.90	8.00	2.99	Yes	No
UCB047	0.60	4.00	3.25	Yes	Yes
UCB033	1.70	14.90	5.95	No	No
UCB049	0.40	4.50	0.03	Yes	Yes
UCB040	0.90	7.50	2.89	No	No
UCB014	0.60	7.60	0.00	Yes	Yes
UCB011	0.50	3.00	2.40	No	No
UCB009	0.90	10.20	0.59	Yes	No
UCB003	3.10	34.20	3.61	No	No
UCB042	1.70	1.00	19.78	Yes	No
UCB027	0.90	8.20	2.04	No	No
UCB031	1.70	20.40	0.23	No	No
UCB005	1.10	9.80	3.09	No	No
UCB024	0.40	3.10	2.17	Yes	Yes
UCB017	0.50	6.10	0.03	Yes	Yes
UCB046	2.60	23.50	7.89	Yes	Yes
UCB039	0.90	6.00	5.16	No	No
UCB023	0.70	7.30	0.69	Yes	Yes
UCB034	1.40	16.10	1.22	Yes	Yes

Continued on next page

Individual participant metadata. Continued.

Participant	Age at EEG (years)	Onset age (months)	Onset to first EEG (months)	Immediate response	Response
UCB041	0.40	5.20	0.00	No	No
UCB012	2.00	9.00	15.44	No	No
UCB013	1.00	2.00	9.79	Yes	Yes
UCB056	0.60	NaN	NaN	NaN	NaN
UCB089	1.60	NaN	NaN	NaN	NaN
UCB098	16.40	NaN	NaN	NaN	NaN
UCB083	4.60	NaN	NaN	NaN	NaN
UCB101	0.60	NaN	NaN	NaN	NaN
UCB080	7.10	NaN	NaN	NaN	NaN
UCB100	10.00	NaN	NaN	NaN	NaN
UCB057	0.60	NaN	NaN	NaN	NaN
UCB053	1.00	NaN	NaN	NaN	NaN
UCB059	0.80	NaN	NaN	NaN	NaN
UCB102	0.60	NaN	NaN	NaN	NaN
UCB052	1.00	NaN	NaN	NaN	NaN
UCB054	1.50	NaN	NaN	NaN	NaN
UCB074	0.80	NaN	NaN	NaN	NaN
UCB078	6.30	NaN	NaN	NaN	NaN
UCB079	6.80	NaN	NaN	NaN	NaN
UCB086	4.40	NaN	NaN	NaN	NaN
UCB097	9.40	NaN	NaN	NaN	NaN
UCB094	3.10	NaN	NaN	NaN	NaN
UCB093	11.30	NaN	NaN	NaN	NaN
UCB072	0.10	NaN	NaN	NaN	NaN
UCB064	0.60	NaN	NaN	NaN	NaN
UCB058	0.30	NaN	NaN	NaN	NaN
UCB071	0.50	NaN	NaN	NaN	NaN
UCB051	0.70	NaN	NaN	NaN	NaN
UCB087	8.20	NaN	NaN	NaN	NaN
UCB062	0.80	NaN	NaN	NaN	NaN
UCB055	0.70	NaN	NaN	NaN	NaN
UCB061	0.50	NaN	NaN	NaN	NaN
UCB077	5.60	NaN	NaN	NaN	NaN
UCB088	12.40	NaN	NaN	NaN	NaN
UCB085	15.40	NaN	NaN	NaN	NaN
UCB082	13.40	NaN	NaN	NaN	NaN
UCB095	4.80	NaN	NaN	NaN	NaN
UCB069	0.40	NaN	NaN	NaN	NaN
UCB084	11.00	NaN	NaN	NaN	NaN
UCB070	0.60	NaN	NaN	NaN	NaN
UCB073	0.70	NaN	NaN	NaN	NaN
UCB076	3.80	NaN	NaN	NaN	NaN
UCB096	17.70	NaN	NaN	NaN	NaN
UCB068	0.40	NaN	NaN	NaN	NaN
UCB081	8.10	NaN	NaN	NaN	NaN
UCB065	0.30	NaN	NaN	NaN	NaN
UCB067	0.30	NaN	NaN	NaN	NaN
UCB099	17.10	NaN	NaN	NaN	NaN
UCB091	16.80	NaN	NaN	NaN	NaN
UCB066	0.70	NaN	NaN	NaN	NaN
UCB075	0.20	NaN	NaN	NaN	NaN
UCB092	12.90	NaN	NaN	NaN	NaN
UCB090	15.40	NaN	NaN	NaN	NaN

B Implementation Details

This appendix describes the quantitative EEG (qEEG) features, foundation-model inputs, training procedures, and evaluation used in the benchmark. Diagnosis and response use separate feature sets and training configurations.

B.1 Diagnostic qEEG

The diagnostic qEEG triplet adapts the beta-band detrended fluctuation analysis (DFA) intercept, beta-band Shannon entropy, and delta-band phase lag index (PLI) features used by Smith et al. (2021).

Feature computation. Signals are analyzed at 200 Hz in microvolts. Entropy and DFA use a linked-ear reference and a 14–30 Hz beta band; PLI uses a common-average reference and a 1–4 Hz delta band. Shannon entropy is computed per channel using Freedman–Diaconis histogram bins and then averaged across the 19 scalp channels. DFA uses the beta-amplitude envelope, with 20 logarithmically spaced scales from one second to one tenth of the retained signal duration, capped at 120 seconds. The intercept is averaged across channels. These frequency bands, references, entropy binning, and DFA scale limits follow the published analysis. Band-specific filters use a SciPy least-squares finite impulse response design and forward–backward filtering; the tap count is increased to the nearest odd value when required by SciPy.

Connectivity and scalar reduction. Following the surrogate-testing procedure in Smith et al. (2021), PLI is calculated in eight-second artifact-free epochs. Our implementation generates 100 surrogates per epoch by independently randomizing Fourier phases while retaining each channel's Fourier magnitudes. An observed channel-pair PLI is retained if it exceeds the surrogate distribution's 95th percentile; otherwise, it is set to zero. Retained values keep their original magnitude. No multiple-comparison correction is applied within this mask. We average the masked matrices over epochs, then average the 171 unique off-diagonal entries to obtain one scalar connectivity feature. Zeroed connections remain in this mean.

Artifact screening. The source study used automated extreme-value detection and impedance-check removal. A linked-ear signal is filtered with a first-order high-pass filter at approximately 1.59 Hz and a third-order low-pass filter at 40 Hz. The amplitude threshold is the larger of 7.5 channel standard deviations and 200 μV, measured about the channel mean. Detected samples receive a 0.9-second buffer on either side. Samples with unchanged values in more than eight channels are also marked. We exclude an entire one-second block if it contains a marked sample. The absolute threshold floor and whole-second exclusion are implementation choices beyond the source paper's stated 7.5-standard-deviation rule. PLI epochs remain within contiguous clean intervals; entropy and DFA use concatenated clean seconds. Inputs with at most ten clean seconds receive missing feature values.

Benchmark adaptation. The published diagnostic model used sleep features averaged across two sleep EEG segments per participant. Here, each clinician-selected EEG segment supplies a training example, and each simulated routine EEG segment supplies a test example; training includes both sleep and awake segments. We fit an L2-regularized logistic classifier within the fixed patient partitions and select regularization using validation data. We do not reuse the published regression coefficients. Thus, the benchmark evaluates the diagnostic feature set under a different fitting and evaluation design.

Expanded qEEG. We combine 12 features from the families studied by Smith et al. (2021) for an additional diagnostic classifier. The complete set comprises (1) the channel mean of median one-second peak-to-peak amplitude; (2) median one-second peak-to-peak amplitude at Cz; (3–6) channel-averaged log power in the delta, theta, alpha, and beta bands; (7) the channel-averaged 95% spectral edge frequency; (8) channel-averaged beta Shannon entropy; (9) beta permutation entropy at Cz; (10) channel-averaged beta DFA intercept; (11) channel-averaged beta DFA exponent; and (12) the scalar delta PLI feature defined above. These are computed in the same implementation, with one feature vector per EEG segment.

B.2 TREATMENT-RESPONSE QEEG

The response feature set follows Rajaraman et al. (2024) and comprises beta-band DFA intercept, beta-band Shannon entropy, and delta-band PLI connectivity. It uses the linked-ear and common-average references, band-specific filters, and artifact mask described in Appendix B.1. The DFA calculation is shared with the diagnosis implementation. Response entropy uses 350 equally spaced histogram-bin centers and is averaged across the 19 scalp channels.

For response connectivity, we average raw PLI across contiguous clean eight-second epochs, then calculate the percentage of the 171 channel pairs whose mean PLI exceeds 0.20. This calculation does not apply the surrogate mask used by the diagnostic connectivity feature. In the main response benchmark, each complete long recording supplies one vector of these three features.

Classifier fitting. Each diagnostic and response feature set is mapped to a probability by L2-regularized logistic regression. Missing values are imputed with training medians, and features are standardized using training means and standard deviations. The candidate inverse regularization strengths are $\{0.01, 0.1, 1, 10, 100\}$ for diagnosis and $\{0.01, 0.1, 1, 10\}$ for response. Validation AUROC selects the regularization strength, retaining the first candidate in ascending order in a tie. Diagnosis uses the L-BFGS solver without class weighting; response uses the LIBLINEAR solver with inverse-frequency class weights. Both allow at most 3,000 iterations. Classifiers are fitted on three training folds, selected using the validation fold, and evaluated on the test fold without refitting on validation patients.

B.3 EEG FOUNDATION MODELS

Table B1 lists the inputs used for both frozen encoders and full fine-tuning in the main benchmark. The exploratory analyses in Appendix C use 10-second LaBraM embeddings, as specified there.

Table B1: Foundation-model inputs in the main benchmark. Window lengths are in seconds; sampling rates are those presented to the model.

Model	Diagnosis	Response	Rate (Hz)	Channels
BIOT	30	30	200	18 bipolar
LaBraM	10	16	200	19 scalp
CBraMod	30	30	200	19 scalp
EEGPT	4	4	250	19 scalp
LUNA	30	30	256	22 bipolar
REVE	30	30	200	19 scalp
CodeBrain	30	30	200	19 scalp
CSBrain	30	30	200	19 scalp

Input processing. BIOT normalizes each bipolar channel by its window's 95th percentile absolute amplitude. CBraMod and LaBraM use 19 scalp channels with a common-average reference, a 60-Hz notch, and band limits of 0.3–75 Hz and 0.1–75 Hz, respectively. Signals are divided by 100 μV and arranged as one-second patches. EEGPT resamples to 250 Hz and applies exponential moving standardization. LUNA uses a 22-channel bipolar montage, resampling to 256 Hz and channel-wise standardization. REVE receives standardized signals and electrode coordinates. CodeBrain and CSBrain use 19-channel, one-second patches at 200 Hz, with amplitudes divided by 100 μV.

For response, windows are sampled directly from the original long recordings. Except for BIOT and LUNA, input processing includes a common-average reference, a 60-Hz notch, and a 0.3–75 Hz bandpass, with LaBraM's lower cutoff instead set to 0.1 Hz. Filtering is restricted to the sampled window. EEGPT uses an exponential update factor of 0.001 and a numerical floor of 10^{-4}. LUNA and REVE are standardized separately within each window and channel.

Response representations and heads. BIOT supplies its native pooled representation. CBraMod, LaBraM, EEGPT, LUNA, CodeBrain, and CSBrain average their output tokens; REVE uses its attention pooling layer. The resulting dimensions are 256 for BIOT and LUNA, 512 for EEGPT and REVE, and 200 for the remaining models. Frozen response encoders use a layer-normalization (LayerNorm) layer and a linear binary head. Full fine-tuning uses the same head structure, with

dropout before the linear layer for REVE (0.3), CodeBrain (0.4), and CSBrain (0.1). The encoder and this head are updated together.

B.4 Training Details

Patient partitions and validation. Diagnosis folds are stratified by patient diagnosis label; both response tasks share a separate partition stratified by treatment response. All methods within a task use the same five outer folds. Each patient's recordings and derived windows remain in the same outer fold, including pre-treatment and post-treatment EEG. Each diagnosis test fold contains ten cases and ten controls; each response test fold contains ten cases. For qEEG, frozen diagnosis encoders, and all response models, fold f supplies test patients, fold $(f + 1) \bmod 5$ supplies validation patients, and the remaining three folds supply training patients. Fully fine-tuned diagnosis models retain their model-specific internal validation procedures within the four non-test folds. The outer test patients are excluded from fitting and model selection. Each outer fold uses its own fitted head and, for full fine-tuning, its own encoder checkpoint.

Frozen diagnosis encoders. All eight frozen encoders use a LayerNorm and linear classification head. Training-window means and standard deviations normalize the embeddings, which are clipped to $[-10, 10]$. The head minimizes binary cross-entropy with a positive-class weight equal to the ratio of negative to positive training windows. Adaptive moment estimation with decoupled weight decay (AdamW) uses a learning rate of 3×10^{-4}, weight decay of 10^{-3}, and batches of up to 512 windows. Gradient norms are clipped at 5. Each epoch visits all training windows in shuffled order. Validation probabilities are averaged within each clinician-selected EEG segment before computing AUROC. Training ends after ten epochs without an improvement greater than 10^{-5}, with a maximum of 100 epochs. The selected head is evaluated without refitting on validation patients. The random seed is $20260924 + f$ for outer fold f.

Fully fine-tuned diagnosis models. Table B2 gives the model-specific optimization settings. All models are initialized with pretrained weights and trained for at most 50 epochs. BIOT uses its pooled binary head; CBraMod flattens the channel and time representations and uses a three-layer classifier with hidden widths 6,000 and 200, exponential linear unit activations, and dropout 0.1. LaBraM uses a linear classifier on mean-pooled tokens. EEGPT uses its montage-projection and classification layers, LUNA uses its classification head, REVE uses attention pooling and a linear head, and CodeBrain and CSBrain use mean pooling, dropout 0.1, and a linear head. BIOT and CBraMod minimize binary cross-entropy; the remaining models use two-class cross-entropy, with label smoothing of 0.1 for LaBraM.

Table B2: Full fine-tuning settings for diagnosis. Learning rates are peak values. Selection refers to the validation criterion; loss and accuracy are computed over windows.

Model	Optimizer	Learning rate	Weight decay	Batch	Warmup	Selection
BIOT	Adam	3×10^{-5}	0.05	64	0	AUROC
CBraMod	AdamW	3×10^{-5}	0.05	64	0	Loss
LaBraM	AdamW	5×10^{-4}	0.05	64	5	Accuracy
EEGPT	AdamW	5×10^{-4}	0.05	64	5	Loss
LUNA	AdamW	5×10^{-4}	0.05	64	5	Loss
REVE	AdamW	10^{-4}	0.05	16	5	Loss
CodeBrain	AdamW	10^{-4}	5×10^{-4}	32	0	Loss
CSBrain	AdamW	10^{-4}	0.05	32	0	Loss

BIOT uses a constant learning rate and early stopping with patience five. The other models use cosine schedules over 50 epochs. The minimum rate is 10^{-6} except for LUNA, which uses 2.5×10^{-6}. Layer-wise learning-rate decay is 0.65 for LaBraM and 0.75 for EEGPT, LUNA, and REVE. LaBraM also uses drop path 0.1. EEGPT, LUNA, REVE, CodeBrain, and CSBrain stop after eight epochs without lower validation loss; CBraMod and LaBraM retain the best checkpoint over all 50 epochs. Gradient clipping is 1 for CBraMod, EEGPT, LUNA, and REVE, and 5 for CodeBrain and CSBrain.

Each training epoch samples windows from clinician-selected EEG segments, with 10,000 draws for BIOT, 5,000 for CBraMod, LaBraM, EEGPT, and LUNA, and 2,500 for REVE, CodeBrain, and CSBrain. Validation windows are non-overlapping for BIOT, CBraMod, and LaBraM; their starting

positions are spaced 120 seconds apart for EEGPT and LUNA and 240 seconds apart for REVE, CodeBrain, and CSBrain. The random seed is 3407 for CBraMod and 42 for the other models. Frozen and fully fine-tuned comparisons share the outer test patients, test-window durations, and probability averaging, while their heads and fitting procedures differ.

Frozen response encoders. For each training or validation recording, we sample 30 continuous 30-minute segments at uniformly distributed valid starting positions. Segments may overlap, and their constituent windows inherit the patient's response label. A LayerNorm and linear window head is trained on frozen embeddings standardized with training statistics and clipped to $[-10, 10]$. AdamW uses a learning rate of 3×10^{-4}, weight decay of 10^{-3}, and gradient clipping at 5. A minibatch contains up to 64 segments and includes their constituent windows. Training-fold response frequencies determine the positive-class weight in binary cross-entropy. Validation window AUROC selects the head, with a maximum of 100 epochs and patience of ten epochs for improvements greater than 10^{-5}. The random seed is $20260924 + f$. Testing then encodes each complete held-out recording as described in Appendix B.5.

Full fine-tuning for response. All eight models update the encoder and classification head using Table B3. Each epoch draws 5,000 windows, approximately balanced across training patients, with a uniformly sampled valid start within each recording. Training minimizes unweighted binary cross-entropy. Validation uses 64 fixed random windows per patient. Its window AUROC selects the checkpoint, with at most 100 epochs and patience of ten epochs for improvements greater than 10^{-4}. Early stopping is applied after warmup. Selected checkpoints are used without refitting on validation patients. The seed is $20260924 + 100f$ for pre-treatment models, with 10,000 added for post-treatment models.

Table B3: Full fine-tuning settings for response. Learning rates are peak values; warmup is measured in epochs. Adam denotes adaptive moment estimation.

Model	Optimizer	Learning rate	Weight decay	Batch	Warmup	Schedule
BIOT	Adam	10^{-3}	10^{-5}	64	0	Constant
CBraMod	AdamW	10^{-4}	0.05	64	0	Cosine
LaBraM	AdamW	5×10^{-4}	0.05	64	5	Cosine
EEGPT	AdamW	5×10^{-4}	0.05	64	5	Cosine
LUNA	AdamW	10^{-4}	0.05	32	5	Cosine
REVE	StableAdamW	10^{-4}	0.01	16	5	Plateau
CodeBrain	AdamW	10^{-5}	5×10^{-5}	32	0	Cosine
CSBrain	AdamW	10^{-4}	0.01	32	0	Cosine

Cosine schedules use a 100-epoch horizon and a minimum learning rate of 10^{-6}, except for LUNA, which uses 5×10^{-6}. Layer-wise learning-rate decay is 0.65 for LaBraM and EEGPT and 0.5 for LUNA. LaBraM and LUNA use drop path 0.1. CodeBrain uses backbone dropout 0.4. REVE uses mixup with coefficient sampled from $\mathrm{Beta}(1, 1)$ and halves its learning rate after five epochs without improved validation AUROC. Its optimizer uses $(\beta_1, \beta_2) = (0.92, 0.999)$ and $\epsilon = 10^{-9}$; the other optimizers use $(0.9, 0.999)$ and 10^{-8}. Gradient norms are clipped at 2 for REVE and 1 for the other models.

B.5 Window Aggregation

For an input represented by K windows $x_1, \ldots, x_K$, an encoder f_θ and binary head g_ϕ produce window probabilities p_k. A single-logit head uses $p_k = \sigma(g_\phi(f_\theta(x_k)))$; a two-class head uses the positive-class softmax probability. The aggregated probability is

$$\widehat{p}(X) = \frac{1}{K} \sum_{k=1}^{K} p_k. \tag{1}$$

All main benchmark comparisons use this rule for both frozen and fully fine-tuned foundation models. No additional classifier is fitted after averaging. Complete non-overlapping windows tile each diagnosis test segment and each original long recording in the response test set, beginning at time zero. An incomplete final window is discarded. Thus, response evaluation covers the full

available recording rather than only the segments sampled for training. Feature-level classifiers used in the exploratory analyses are described separately in Appendix C.

B.6 Evaluation Details

AUROC across folds. Let A_f denote AUROC in test fold f. Following LeDell et al. (2015), we report the equally weighted cross-validated AUROC,

$$\widehat{A} = \frac{1}{5}\sum_{f=1}^{5} A_f. \tag{2}$$

Each A_f compares scores from the same fitted model. Positive labels indicate IESS for diagnosis and response to treatment for response prediction. Equal scores receive half credit. We do not pool scores from different folds when computing AUROC.

Balanced accuracy. Balanced accuracy is the arithmetic mean of sensitivity and specificity, computed from the combined held-out classifications. Unlike AUROC, we do not average balanced accuracy across folds. The pooled calculation equals the mean of fold-wise balanced accuracies when every fold has the same positive and negative counts, as in the evaluation against patient diagnosis labels. Class counts vary across folds for clinician diagnosis labels and treatment-response labels, so the two calculations can differ. For all main benchmark results, predictions are classified using a fixed probability threshold of 0.5. Logistic regression decision scores are converted to probabilities before evaluation.

AUROC confidence intervals. We use the influence-function method of LeDell et al. (2015), including its extension to repeated observations from the same patient. Within fold f, let C_j be the proportion of opposite-class observations correctly ranked relative to observation j, with half credit for ties, and let $\widehat{p}_y$ be the proportion of all evaluation observations with label y. The observation's estimated influence is

$$D_j = \frac{C_j - A_f}{\widehat{p}_{y_j}}. \tag{3}$$

For patient i, we sum these contributions and divide by the mean number of observations per patient, $\overline{m}$, giving $U_i = \overline{m}^{-1}\sum_{j\in i} D_j$. Diagnosis has two segments per patient; recording-level response evaluation has one recording per patient, and the random-segment analyses in Appendix C have ten segments per patient. The metric remains defined over the stated evaluation inputs. The standard error is

$$\widehat{\mathrm{SE}}(\widehat{A}) = \sqrt{\frac{1}{N}\left(\frac{1}{5}\sum_{f=1}^{5}\frac{1}{N_f}\sum_{i\in f} U_i^2\right)}, \tag{4}$$

where N is the number of patients and $N_f = N/5$. For clinician diagnosis labels, each segment retains its own label when calculating D_j, and the two contributions from a patient are combined when calculating U_i.

Balanced-accuracy confidence intervals. We derive the interval separately using the delta method for sensitivity and specificity. Let $h_j \in \{0, 1\}$ be the held-out classification, $\widehat{s}$ sensitivity, $\widehat{c}$ specificity, and $\widehat{p}$ the positive-label proportion. The estimated observation-level influence is

$$D_j^{\mathrm{bal}} = \frac{1}{2}\left[\frac{y_j}{\widehat{p}}(h_j - \widehat{s}) + \frac{1-y_j}{1-\widehat{p}}(1 - h_j - \widehat{c})\right]. \tag{5}$$

We aggregate these contributions within patients as above and estimate the variance of balanced accuracy by $N^{-2}\sum_i (U_i^{\mathrm{bal}})^2$. This derivation applies to the fixed held-out classifications and is separate from the AUROC method of LeDell et al. (2015).

For both metrics, 95% confidence intervals (CIs) are the estimate plus or minus 1.96 standard errors, restricted to $[0, 1]$. These are asymptotic intervals that assume independent patients and sufficient stability of the fitted prediction rules. They do not include variation from rerunning the complete training and model-selection procedure. With 50 patients in each response task, nominal 95% coverage is not guaranteed; LeDell et al. (2015) also report reduced coverage in small samples.

C Additional Analyses

C.1 Additional Classification Metrics

Tables C1–C4 provide full classification results for diagnosis under both label definitions and for pre-treatment and post-treatment response prediction. They use the same held-out predictions and probability threshold of 0.5 as the main tables. Diagnosis is evaluated on 200 simulated routine EEG segments from 100 patients; each response task is evaluated on 50 original long EEG recordings from 50 patients. The positive class is IESS for diagnosis and treatment response for both response tasks. Patient diagnosis labels contain 100 positive and 100 negative segments; clinician diagnosis labels contain 96 positive and 104 negative segments. Each response task contains 28 responders and 22 non-responders.

Metrics. We report AUROC, balanced accuracy, sensitivity, specificity, accuracy, precision, F1, and average precision. Sensitivity is equivalent to recall. Precision and F1 refer to the positive class, with F1 computed as $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, where TP, FP, and FN denote true positives, false positives, and false negatives. AUROC and average precision are averaged across the five test folds; the other metrics are computed from the combined held-out classifications. Average precision is the non-interpolated area summary of the precision–recall curve, computed as the sum of precision weighted by successive increases in recall. All table entries are percentages with 95% confidence intervals. Each table separates the eight metrics into two panels for readability.

Confidence intervals. AUROC intervals follow Appendix B.6. For the other metrics, we use empirical influence functions with patient clustering. For any ratio $\widehat{\theta} = \overline{a}/\overline{b}$, the observation-level influence is $(a_i - \widehat{\theta} b_i)/\overline{b}$. This gives sensitivity, specificity, precision, and F1 by substituting their confusion-count numerators and denominators. Accuracy uses the centered indicator of a correct classification; balanced accuracy uses the mean of the sensitivity and specificity influences.

For average precision, let s_i be a predicted probability, y_i its binary label, p the positive-class proportion, $a_t = P_n(Y = 1, S = t)$, $b_t = P_n(S \geq t)$, and $r_t = P_n(Y = 1, S \geq t)/b_t$, with P_n denoting the empirical distribution within a test fold. Then $\widehat{\mathrm{AP}} = \sum_t a_t r_t / p$ and its empirical influence is

$$\phi_i = \frac{y_i}{p}\left(r_{s_i} - \widehat{\mathrm{AP}}\right) + \frac{1}{p}\sum_{t \leq s_i} \frac{a_t}{b_t}(y_i - r_t).$$

Tied probabilities share a threshold. The equally sized test folds allow the fold-specific influences to be combined for mean average precision. This delta-method calculation is separate from the AUROC method of LeDell et al. (2015).

For each metric, observation-level influences are summed within patients and divided by the mean number of observations per patient. The standard error is the root mean square patient influence divided by the square root of the number of patients. Normal intervals are truncated to $[0, 1]$. These asymptotic intervals condition on the fitted predictions and threshold; small samples and boundary estimates can limit their coverage.

Table C1: IESS diagnosis performance against patient diagnosis labels.

Method	AUROC	Balanced accuracy	Sensitivity	Specificity
Traditional qEEG				
qEEG triplet	91.4 [87.3–95.5]	85.5 [80.1–90.9]	85.0 [77.5–92.5]	86.0 [78.1–93.9]
Expanded qEEG	93.1 [88.8–97.4]	85.5 [79.7–91.3]	76.0 [65.2–86.8]	95.0 [90.8–99.2]
Frozen encoder				
BIOT	87.6 [81.5–93.7]	81.5 [75.5–87.5]	80.0 [71.2–88.8]	83.0 [74.9–91.1]
LaBraM	94.8 [91.3–98.3]	86.5 [80.6–92.4]	77.0 [66.2–87.8]	96.0 [91.3–100.0]
CBraMod	90.7 [85.3–96.1]	84.5 [78.7–90.3]	74.0 [63.5–84.5]	95.0 [90.0–100.0]
EEGPT	94.7 [90.9–98.4]	87.5 [81.7–93.3]	90.0 [82.7–97.3]	85.0 [76.1–93.9]
LUNA	87.5 [81.0–93.9]	84.0 [78.1–89.9]	74.0 [63.2–84.8]	94.0 [89.5–98.5]
REVE	92.8 [88.2–97.3]	87.5 [82.1–92.9]	86.0 [77.7–94.3]	89.0 [82.0–96.0]
CodeBrain	94.6 [91.0–98.1]	88.0 [82.2–93.8]	84.0 [74.6–93.4]	92.0 [85.0–99.0]

Table C1 continued

CSBrain	87.9 [82.1–93.7]	80.5 [74.4–86.6]	69.0 [58.3–79.7]	92.0 [86.2–97.8]
Full fine-tuning				
BIOT	97.2 [94.9–99.4]	90.5 [86.0–95.0]	92.0 [86.2–97.8]	89.0 [82.0–96.0]
LaBraM	97.6 [95.0–100.0]	93.5 [89.2–97.8]	91.0 [83.8–98.2]	96.0 [91.3–100.0]
CBraMod	96.3 [92.7–99.9]	91.0 [86.1–95.9]	89.0 [81.5–96.5]	93.0 [86.8–99.2]
EEGPT	95.4 [91.8–99.0]	91.0 [86.3–95.7]	90.0 [83.2–96.8]	92.0 [85.6–98.4]
LUNA	93.9 [89.0–98.8]	87.0 [81.5–92.5]	81.0 [71.4–90.6]	93.0 [87.4–98.6]
REVE	94.8 [90.8–98.7]	88.5 [83.0–94.0]	92.0 [85.6–98.4]	85.0 [76.1–93.9]
CodeBrain	96.1 [92.9–99.3]	90.5 [85.6–95.4]	90.0 [82.7–97.3]	91.0 [84.4–97.6]
CSBrain	96.5 [93.6–99.4]	91.0 [86.4–95.6]	83.0 [74.0–92.0]	99.0 [97.1–100.0]
Method	Accuracy	Precision	F1	Average precision
Traditional qEEG				
qEEG triplet	85.5 [80.1–90.9]	85.9 [77.5–94.2]	85.4 [79.5–91.4]	93.4 [89.5–97.3]
Expanded qEEG	85.5 [79.4–91.6]	93.8 [88.4–99.2]	84.0 [76.7–91.2]	94.6 [90.9–98.3]
Frozen encoder				
BIOT	81.5 [75.5–87.5]	82.5 [73.4–91.6]	81.2 [74.4–88.1]	89.9 [84.1–95.7]
LaBraM	86.5 [80.3–92.7]	95.1 [89.2–100.0]	85.1 [77.8–92.3]	96.2 [93.4–98.9]
CBraMod	84.5 [78.3–90.7]	93.7 [87.3–100.0]	82.7 [75.4–90.0]	93.1 [88.4–97.8]
EEGPT	87.5 [81.7–93.3]	85.7 [77.0–94.5]	87.8 [81.7–93.9]	95.6 [92.2–99.0]
LUNA	84.0 [77.8–90.2]	92.5 [86.5–98.5]	82.2 [74.8–89.7]	91.5 [86.5–96.5]
REVE	87.5 [82.1–92.9]	88.7 [81.1–96.2]	87.3 [81.3–93.3]	93.9 [89.6–98.2]
CodeBrain	88.0 [82.1–93.9]	91.3 [83.6–99.0]	87.5 [81.0–94.0]	95.6 [92.5–98.7]
CSBrain	80.5 [74.0–87.0]	89.6 [81.8–97.4]	78.0 [70.0–85.9]	90.3 [84.8–95.9]
Full fine-tuning				
BIOT	90.5 [86.0–95.0]	89.3 [82.2–96.4]	90.6 [85.8–95.4]	97.3 [95.1–99.6]
LaBraM	93.5 [89.2–97.8]	95.8 [90.8–100.0]	93.3 [88.8–97.9]	97.8 [95.3–100.0]
CBraMod	91.0 [86.1–95.9]	92.7 [86.1–99.3]	90.8 [85.6–96.0]	96.3 [92.2–100.0]
EEGPT	91.0 [86.3–95.7]	91.8 [85.1–98.6]	90.9 [85.9–95.9]	96.6 [93.9–99.3]
LUNA	87.0 [81.3–92.7]	92.0 [85.5–98.6]	86.2 [79.7–92.6]	96.2 [93.1–99.3]
REVE	88.5 [83.0–94.0]	86.0 [77.4–94.6]	88.9 [83.2–94.6]	95.0 [90.8–99.2]
CodeBrain	90.5 [85.6–95.4]	90.9 [84.0–97.8]	90.5 [85.2–95.7]	96.5 [93.2–99.8]
CSBrain	91.0 [86.1–95.9]	98.8 [96.5–100.0]	90.2 [84.7–95.7]	97.0 [94.2–99.7]

Table C2: IESS diagnosis performance against clinician diagnosis labels.

Method	AUROC	Balanced accuracy	Sensitivity	Specificity
Traditional qEEG				
qEEG triplet	90.7 [85.9–95.4]	83.5 [77.9–89.2]	84.4 [76.9–91.9]	82.7 [74.2–91.2]
Expanded qEEG	94.7 [91.3–98.1]	86.2 [80.7–91.7]	78.1 [67.9–88.4]	94.2 [89.9–98.6]
Frozen encoder				
BIOT	88.4 [82.9–94.0]	81.5 [75.7–87.3]	81.3 [72.5–90.0]	81.7 [73.7–89.8]
LaBraM	97.3 [95.3–99.4]	89.2 [83.9–94.4]	81.3 [71.6–90.9]	97.1 [93.0–100.0]
CBraMod	92.2 [87.8–96.6]	85.1 [79.6–90.6]	76.0 [66.1–85.9]	94.2 [89.1–99.4]
EEGPT	93.2 [88.9–97.5]	84.7 [78.5–90.8]	88.5 [81.0–96.1]	80.8 [71.1–90.5]
LUNA	88.1 [81.6–94.6]	84.7 [78.8–90.5]	76.0 [65.3–86.8]	93.3 [88.6–97.9]
REVE	91.1 [86.3–96.0]	85.5 [79.9–91.1]	85.4 [77.4–93.4]	85.6 [77.5–93.7]
CodeBrain	95.1 [91.9–98.4]	89.9 [84.7–95.1]	87.5 [79.4–95.6]	92.3 [85.6–99.0]
CSBrain	90.2 [84.8–95.6]	81.1 [75.2–86.9]	70.8 [60.5–81.1]	91.3 [85.5–97.2]
Full fine-tuning				
BIOT	95.0 [91.7–98.2]	86.6 [81.2–92.0]	89.6 [82.9–96.3]	83.7 [75.2–92.1]
LaBraM	97.4 [94.9–99.9]	92.5 [88.2–96.7]	91.7 [85.4–97.9]	93.3 [87.3–99.3]
CBraMod	95.2 [91.4–99.0]	90.0 [85.4–94.6]	89.6 [83.5–95.7]	90.4 [83.3–97.5]
EEGPT	92.6 [88.1–97.1]	87.0 [81.4–92.6]	87.5 [79.8–95.2]	86.5 [78.5–94.6]
LUNA	92.4 [87.2–97.7]	85.8 [80.0–91.6]	81.3 [71.9–90.6]	90.4 [83.8–97.0]
REVE	91.4 [86.6–96.3]	84.7 [78.6–90.8]	89.6 [82.8–96.3]	79.8 [69.7–89.9]
CodeBrain	96.2 [93.0–99.5]	90.5 [85.9–95.2]	91.7 [85.4–97.9]	89.4 [82.2–96.6]

Table C2 continued

CSBrain	96.4 [93.2–99.6]	92.7 [88.8–96.7]	86.5 [78.7–94.2]	99.0 [97.2–100.0]
Method	Accuracy	Precision	F1	Average precision
Traditional qEEG				
qEEG triplet	83.5 [77.8–89.2]	81.8 [72.5–91.1]	83.1 [76.5–89.6]	91.6 [86.3–96.8]
Expanded qEEG	86.5 [81.0–92.0]	92.6 [86.8–98.4]	84.7 [77.9–91.6]	95.7 [92.4–98.9]
Frozen encoder				
BIOT	81.5 [75.7–87.3]	80.4 [71.1–89.8]	80.8 [73.9–87.8]	89.9 [84.3–95.6]
LaBraM	89.5 [84.3–94.7]	96.3 [91.0–100.0]	88.1 [81.8–94.4]	97.7 [95.7–99.6]
CBraMod	85.5 [79.9–91.1]	92.4 [85.6–99.2]	83.4 [76.5–90.3]	93.6 [89.0–98.1]
EEGPT	84.5 [78.2–90.8]	81.0 [71.0–90.9]	84.6 [77.6–91.6]	94.2 [90.2–98.3]
LUNA	85.0 [79.2–90.8]	91.3 [84.8–97.7]	83.0 [75.6–90.3]	91.4 [86.3–96.5]
REVE	85.5 [79.9–91.1]	84.5 [75.5–93.6]	85.0 [78.4–91.5]	92.5 [87.8–97.3]
CodeBrain	90.0 [84.8–95.2]	91.3 [83.6–99.0]	89.4 [83.4–95.3]	95.9 [92.9–99.0]
CSBrain	81.5 [75.5–87.5]	88.3 [80.2–96.4]	78.6 [71.0–86.2]	92.1 [87.1–97.1]
Full fine-tuning				
BIOT	86.5 [81.0–92.0]	83.5 [74.6–92.4]	86.4 [80.3–92.6]	95.4 [92.1–98.7]
LaBraM	92.5 [88.3–96.7]	92.6 [86.0–99.2]	92.1 [87.4–96.9]	97.4 [94.8–100.0]
CBraMod	90.0 [85.4–94.6]	89.6 [81.8–97.3]	89.6 [84.4–94.8]	95.2 [90.6–99.8]
EEGPT	87.0 [81.3–92.7]	85.7 [76.9–94.5]	86.6 [80.2–93.0]	94.1 [90.2–97.9]
LUNA	86.0 [80.3–91.7]	88.6 [80.6–96.6]	84.8 [77.9–91.6]	94.8 [91.1–98.6]
REVE	84.5 [78.2–90.8]	80.4 [70.3–90.5]	84.7 [77.9–91.6]	92.0 [86.9–97.2]
CodeBrain	90.5 [85.8–95.2]	88.9 [81.1–96.7]	90.3 [85.0–95.6]	96.7 [93.4–99.9]
CSBrain	93.0 [89.1–96.9]	98.8 [96.5–100.0]	92.2 [87.6–96.8]	97.6 [95.4–99.8]

Table C3: Pre-treatment response prediction performance on original long EEG recordings.

Method	AUROC	Balanced accuracy	Sensitivity	Specificity
Traditional qEEG				
Response qEEG	62.3 [46.7–77.9]	61.2 [47.7–74.7]	67.9 [50.6–85.2]	54.5 [33.7–75.4]
Frozen encoder				
BIOT	73.0 [59.7–86.3]	55.2 [42.8–67.5]	78.6 [63.4–93.8]	31.8 [12.4–51.3]
LaBraM	70.6 [55.4–85.7]	69.8 [57.0–82.6]	71.4 [54.7–88.2]	68.2 [48.7–87.6]
CBraMod	54.0 [38.7–69.3]	51.8 [37.8–65.7]	53.6 [35.1–72.0]	50.0 [29.1–70.9]
EEGPT	58.6 [42.3–74.9]	56.7 [43.1–70.2]	67.9 [50.6–85.2]	45.5 [24.6–66.3]
LUNA	60.5 [48.8–72.3]	57.6 [43.8–71.4]	60.7 [42.6–78.8]	54.5 [33.7–75.4]
REVE	59.2 [42.7–75.6]	50.0 [36.0–64.0]	50.0 [31.5–68.5]	50.0 [29.1–70.9]
CodeBrain	51.5 [36.2–66.8]	57.8 [44.7–70.9]	42.9 [24.5–61.2]	72.7 [54.1–91.3]
CSBrain	55.0 [40.1–69.8]	51.0 [37.2–64.7]	42.9 [24.5–61.2]	59.1 [38.5–79.6]
Full fine-tuning				
BIOT	67.9 [53.0–82.9]	54.1 [40.1–68.0]	53.6 [35.1–72.0]	54.5 [33.7–75.4]
LaBraM	68.1 [52.9–83.3]	63.0 [49.6–76.3]	71.4 [54.7–88.2]	54.5 [33.7–75.4]
CBraMod	55.7 [40.1–71.2]	50.3 [36.9–63.7]	64.3 [46.5–82.0]	36.4 [16.3–56.5]
EEGPT	67.3 [51.2–83.4]	62.5 [49.3–75.7]	75.0 [59.0–91.0]	50.0 [29.1–70.9]
LUNA	59.9 [44.1–75.6]	54.4 [41.0–67.8]	67.9 [50.6–85.2]	40.9 [20.4–61.5]
REVE	57.5 [40.6–74.3]	47.1 [35.0–59.2]	71.4 [54.7–88.2]	22.7 [5.2–40.2]
CodeBrain	59.1 [43.2–75.1]	63.5 [50.0–76.9]	67.9 [50.6–85.2]	59.1 [38.5–79.6]
CSBrain	61.4 [46.2–76.6]	54.1 [40.1–68.0]	53.6 [35.1–72.0]	54.5 [33.7–75.4]
Method	Accuracy	Precision	F1	Average precision
Traditional qEEG				
Response qEEG	62.0 [48.5–75.5]	65.5 [48.2–82.8]	66.7 [52.5–80.8]	71.3 [57.2–85.4]
Frozen encoder				
BIOT	58.0 [44.3–71.7]	59.5 [43.6–75.3]	67.7 [54.6–80.8]	77.4 [64.4–90.4]
LaBraM	70.0 [57.3–82.7]	74.1 [57.5–90.6]	72.7 [59.4–86.0]	78.6 [65.4–91.8]
CBraMod	52.0 [38.2–65.8]	57.7 [38.7–76.7]	55.6 [39.6–71.5]	70.6 [56.4–84.8]

Table C3 continued

EEGPT	58.0 [44.3–71.7]	61.3 [44.1–78.4]	64.4 [50.2–78.6]	71.8 [58.7–84.8]
LUNA	58.0 [44.3–71.7]	63.0 [44.7–81.2]	61.8 [46.7–76.9]	74.7 [63.3–86.1]
REVE	50.0 [36.1–63.9]	56.0 [36.5–75.5]	52.8 [36.5–69.1]	73.5 [59.9–87.2]
CodeBrain	56.0 [42.2–69.8]	66.7 [44.9–88.4]	52.2 [34.6–69.7]	67.6 [52.7–82.5]
CSBrain	50.0 [36.1–63.9]	57.1 [36.0–78.3]	49.0 [31.8–66.2]	71.6 [58.0–85.3]
Full fine-tuning				
BIOT	54.0 [40.2–67.8]	60.0 [40.8–79.2]	56.6 [40.6–72.6]	76.5 [62.4–90.5]
LaBraM	64.0 [50.7–77.3]	66.7 [49.8–83.5]	69.0 [55.3–82.6]	77.9 [64.5–91.2]
CBraMod	52.0 [38.2–65.8]	56.3 [39.1–73.4]	60.0 [45.3–74.7]	70.8 [56.3–85.4]
EEGPT	64.0 [50.7–77.3]	65.6 [49.2–82.1]	70.0 [56.8–83.2]	75.8 [61.7–89.8]
LUNA	56.0 [42.2–69.8]	59.4 [42.4–76.4]	63.3 [49.1–77.6]	70.8 [56.7–84.9]
REVE	50.0 [36.1–63.9]	54.1 [38.0–70.1]	61.5 [47.6–75.5]	71.1 [56.4–85.8]
CodeBrain	64.0 [50.7–77.3]	67.9 [50.6–85.2]	67.9 [53.8–81.9]	71.6 [57.0–86.1]
CSBrain	54.0 [40.2–67.8]	60.0 [40.8–79.2]	56.6 [40.6–72.6]	73.7 [60.1–87.4]

Table C4: Post-treatment response prediction performance on original long EEG recordings.

Method	AUROC	Balanced accuracy	Sensitivity	Specificity
Traditional qEEG				
Response qEEG	77.0 [63.1–90.9]	65.3 [52.0–78.5]	71.4 [54.7–88.2]	59.1 [38.5–79.6]
Frozen encoder				
BIOT	57.8 [43.5–72.0]	61.4 [48.2–74.5]	50.0 [31.5–68.5]	72.7 [54.1–91.3]
LaBraM	68.6 [54.1–83.1]	64.8 [51.6–77.9]	75.0 [59.0–91.0]	54.5 [33.7–75.4]
CBraMod	75.2 [62.2–88.3]	65.6 [53.3–77.8]	85.7 [72.8–98.7]	45.5 [24.6–66.3]
EEGPT	58.1 [42.4–73.9]	51.8 [37.8–65.7]	53.6 [35.1–72.0]	50.0 [29.1–70.9]
LUNA	58.8 [42.6–75.0]	55.8 [42.0–69.7]	57.1 [38.8–75.5]	54.5 [33.7–75.4]
REVE	59.5 [44.5–74.4]	61.2 [47.7–74.7]	67.9 [50.6–85.2]	54.5 [33.7–75.4]
CodeBrain	65.4 [49.7–81.1]	58.0 [44.9–71.0]	75.0 [59.0–91.0]	40.9 [20.4–61.5]
CSBrain	63.8 [46.8–80.9]	66.6 [53.7–79.4]	78.6 [63.4–93.8]	54.5 [33.7–75.4]
Full fine-tuning				
BIOT	64.7 [49.7–79.8]	56.2 [42.9–69.4]	71.4 [54.7–88.2]	40.9 [20.4–61.5]
LaBraM	68.7 [54.1–83.2]	60.7 [47.3–74.1]	71.4 [54.7–88.2]	50.0 [29.1–70.9]
CBraMod	79.4 [66.3–92.6]	64.8 [51.6–77.9]	75.0 [59.0–91.0]	54.5 [33.7–75.4]
EEGPT	62.9 [47.8–78.0]	53.7 [43.8–63.6]	89.3 [77.8–100.0]	18.2 [2.1–34.3]
LUNA	44.8 [30.5–59.0]	47.7 [33.8–61.7]	50.0 [31.5–68.5]	45.5 [24.6–66.3]
REVE	59.6 [44.5–74.8]	55.7 [42.8–68.5]	75.0 [59.0–91.0]	36.4 [16.3–56.5]
CodeBrain	73.7 [59.7–87.6]	57.5 [44.9–70.1]	78.6 [63.4–93.8]	36.4 [16.3–56.5]
CSBrain	59.7 [43.8–75.6]	55.8 [42.0–69.7]	57.1 [38.8–75.5]	54.5 [33.7–75.4]

Method	Accuracy	Precision	F1	Average precision
Traditional qEEG				
Response qEEG	66.0 [52.9–79.1]	69.0 [52.1–85.8]	70.2 [56.6–83.7]	82.1 [69.7–94.4]
Frozen encoder				
BIOT	60.0 [46.4–73.6]	70.0 [49.9–90.1]	58.3 [41.7–74.9]	73.0 [60.2–85.9]
LaBraM	66.0 [52.9–79.1]	67.7 [51.3–84.2]	71.2 [58.1–84.3]	76.8 [63.4–90.2]
CBraMod	68.0 [55.1–80.9]	66.7 [51.3–82.1]	75.0 [63.1–86.9]	85.1 [74.8–95.3]
EEGPT	52.0 [38.2–65.8]	57.7 [38.7–76.7]	55.6 [39.6–71.5]	71.9 [57.8–86.0]
LUNA	56.0 [42.2–69.8]	61.5 [42.8–80.2]	59.3 [43.7–74.8]	71.4 [56.6–86.1]
REVE	62.0 [48.5–75.5]	65.5 [48.2–82.8]	66.7 [52.5–80.8]	70.4 [56.3–84.5]
CodeBrain	60.0 [46.4–73.6]	61.8 [45.4–78.1]	67.7 [54.4–81.1]	76.5 [63.1–90.0]
CSBrain	68.0 [55.1–80.9]	68.8 [52.7–84.8]	73.3 [60.7–85.9]	69.0 [53.0–84.9]
Full fine-tuning				
BIOT	58.0 [44.3–71.7]	60.6 [43.9–77.3]	65.6 [51.7–79.4]	78.9 [67.0–90.8]
LaBraM	62.0 [48.5–75.5]	64.5 [47.7–81.4]	67.8 [54.1–81.5]	77.8 [64.6–90.9]
CBraMod	66.0 [52.9–79.1]	67.7 [51.3–84.2]	71.2 [58.1–84.3]	87.5 [78.4–96.6]
EEGPT	58.0 [44.3–71.7]	58.1 [43.4–72.9]	70.4 [58.3–82.5]	71.0 [55.5–86.4]

Table C4 continued

LUNA	48.0 [34.2–61.8]	53.8 [34.7–73.0]	51.9 [35.6–68.1]	61.1 [46.6–75.5]
REVE	58.0 [44.3–71.7]	60.0 [43.8–76.2]	66.7 [53.2–80.1]	71.7 [58.3–85.1]
CodeBrain	60.0 [46.4–73.6]	61.1 [45.2–77.0]	68.8 [55.7–81.8]	84.8 [74.9–94.6]
CSBrain	56.0 [42.2–69.8]	61.5 [42.8–80.2]	59.3 [43.7–74.8]	71.4 [56.8–86.1]

C.2 Segment Embeddings with Whole-Recording qEEG and Clinical Metadata

Inputs. We use ten randomly sampled 30-minute segments from each of the 50 original pre-treatment long EEG recordings and 50 original post-treatment long EEG recordings. The saved starting positions are shared across all comparisons, and segments may overlap. Each segment inherits the source patient's treatment-response label. Frozen BIOT and CBraMod use 30-second windows and frozen LaBraM uses 10-second windows. We average the native window embeddings within each segment. E denotes the five-dimensional representation obtained after the projection described below. Q contains the three response qEEG features defined in Appendix B.2, evaluated on the complete source recording for the main-text comparison. All ten segments from that recording receive the same Q vector. L is $\ln(1 + \mathrm{LeadtimeMo})$, where lead time is measured from disease onset to the first EEG in months. Each task contains 500 segments from 50 patients. Models with Q receive EEG information from outside the 30-minute segment.

Training and validation. We use the fixed response patient folds. Fold f supplies test patients, fold $(f + 1) \bmod 5$ supplies validation patients, and the other three folds supply training patients. There are 300 training, 100 validation, and 100 test segments from 30, 10, and 10 patients, respectively. Embeddings are standardized on training segments and reduced to five components with partial least squares (PLS) regression supervised by training treatment-response labels (Wold et al., 2001). We use `PLSRegression` with internal scaling disabled, a maximum of 500 iterations, and tolerance 10^{-6}. The dimension is fixed across folds. Each model's projection is shared across its feature combinations within that fold. Q and L are appended after projection.

The concatenated inputs are standardized on training segments and passed to L2-regularized logistic regression. Validation segment AUROC selects C from $\{0.001, 0.01, 0.1, 1, 10\}$, retaining the first candidate when scores differ by at most 10^{-12}. We use the `lbfgs` solver, a maximum of 2,000 iterations, tolerance 10^{-9}, and seed $20260925 + f$. The selected classifier is tested without refitting. Each training segment has weight 0.1, so its patient's ten segments have total weight one. This weighting keeps the regularization of models using only Q and L unchanged by replication of their identical input rows. The same weighting applies to every feature combination.

Evaluation and uncertainty. The analysis evaluates the 500 sampled segments separately for each response task. AUROC is the mean across the five test folds, and balanced accuracy uses a threshold of 0.5. Confidence intervals group observations by patient using Appendix B.6. Paired AUROC intervals use the difference between the two methods' patient influence values. These analyses followed earlier exploratory comparisons of feature combinations, projections, and classifiers. Their intervals are conditional on the fitted predictions and do not account for this exploration. Tables C6 and C7 report all feature combinations for pre-treatment and post-treatment prediction, respectively.

For the pre-treatment whole-recording comparison, the paired AUROC gains from adding Q to E are 2.1 percentage points for BIOT (95% CI $[-5.9, 10.2]$), 4.5 for LaBraM ($[-2.9, 11.9]$), and 8.9 for CBraMod ($[-1.8, 19.6]$). Table C5 gives the remaining paired comparisons.

Table C5: Paired differences in segment-level AUROC for whole-recording qEEG combinations, in percentage points with 95% confidence intervals.

Input	Comparison	BIOT	LaBraM	CBraMod
Pre-treatment	E+Q vs. E	2.1 [-5.9, 10.2]	4.5 [-2.9, 11.9]	8.9 [-1.8, 19.6]
Pre-treatment	E+Q vs. Q	7.9 [-2.1, 17.9]	9.9 [-3.0, 22.7]	4.3 [-7.1, 15.7]
Pre-treatment	E+L vs. E	3.9 [-1.0, 8.7]	4.0 [-0.6, 8.6]	10.6 [1.4, 19.7]
Pre-treatment	E+Q+L vs. E+Q	3.1 [-1.7, 7.9]	2.2 [-2.0, 6.4]	4.3 [-2.1, 10.7]
Pre-treatment	E+Q+L vs. E+L	1.4 [-5.3, 8.1]	2.6 [-4.0, 9.2]	2.6 [-5.6, 10.8]

Input	Comparison	BIOT	LaBraM	CBraMod
Pre-treatment	E+Q+L vs. Q+L	3.7 [-4.7, 12.2]	4.7 [-5.9, 15.3]	1.3 [-8.8, 11.3]
Post-treatment	E+Q vs. E	8.8 [-1.5, 19.1]	11.4 [3.9, 19.0]	7.4 [-1.3, 16.1]
Post-treatment	E+Q vs. Q	-7.2 [-15.5, 1.1]	-14.4 [-24.8, -3.9]	-6.0 [-17.2, 5.2]
Post-treatment	E+L vs. E	10.5 [-1.0, 22.0]	9.5 [2.7, 16.4]	5.2 [-1.9, 12.3]
Post-treatment	E+Q+L vs. E+Q	5.0 [-4.0, 14.0]	5.6 [0.2, 11.0]	3.7 [-2.9, 10.3]
Post-treatment	E+Q+L vs. E+L	3.3 [-7.8, 14.3]	7.5 [-0.0, 15.0]	5.9 [-0.3, 12.1]
Post-treatment	E+Q+L vs. Q+L	-4.6 [-14.8, 5.5]	-11.2 [-21.4, -1.0]	-4.7 [-14.1, 4.7]

Table C6: Pre-treatment response with 500 random 30-minute segments. Values are percentages with 95% confidence intervals.

Foundation model	Features	AUROC	Balanced accuracy
Embedding and lead-time references			
Shared reference	L	66.7 [51.3, 82.1]	50.0 [50.0, 50.0]
BIOT	E	68.1 [56.6, 79.5]	53.2 [46.3, 60.2]
BIOT	E+L	71.9 [59.4, 84.4]	52.0 [48.3, 55.6]
LaBraM	E	67.6 [55.9, 79.3]	56.8 [45.9, 67.7]
LaBraM	E+L	71.7 [59.5, 83.8]	56.1 [45.0, 67.3]
CBraMod	E	57.7 [46.0, 69.4]	57.0 [45.9, 68.2]
CBraMod	E+L	68.3 [56.2, 80.3]	57.1 [47.3, 66.9]
qEEG from the complete source recording			
Shared reference	Q	62.3 [46.7, 77.9]	50.0 [50.0, 50.0]
Shared reference	Q+L	69.6 [54.9, 84.3]	50.0 [50.0, 50.0]
BIOT	E+Q	70.2 [57.4, 83.0]	50.0 [50.0, 50.0]
BIOT	E+Q+L	73.3 [60.5, 86.2]	50.0 [50.0, 50.0]
LaBraM	E+Q	72.1 [59.3, 84.9]	50.7 [46.2, 55.2]
LaBraM	E+Q+L	74.3 [61.4, 87.3]	50.0 [50.0, 50.0]
CBraMod	E+Q	66.6 [54.8, 78.3]	50.0 [50.0, 50.0]
CBraMod	E+Q+L	70.9 [59.5, 82.2]	52.0 [46.8, 57.3]

Table C7: Post-treatment response with 500 random 30-minute segments. Values are percentages with 95% confidence intervals.

Foundation model	Features	AUROC	Balanced accuracy
Embedding and lead-time references			
Shared reference	L	66.7 [51.3, 82.1]	50.0 [50.0, 50.0]
BIOT	E	60.2 [48.2, 72.2]	53.7 [45.1, 62.3]
BIOT	E+L	70.7 [59.6, 81.9]	51.7 [44.0, 59.4]
LaBraM	E	50.4 [40.5, 60.3]	53.7 [44.9, 62.4]
LaBraM	E+L	60.0 [48.6, 71.4]	54.2 [46.8, 61.5]
CBraMod	E	62.8 [51.9, 73.7]	55.0 [47.9, 62.2]
CBraMod	E+L	68.0 [56.2, 79.9]	56.0 [47.9, 64.1]
qEEG from the complete source recording			
Shared reference	Q	76.2 [62.2, 90.2]	55.0 [47.1, 63.0]
Shared reference	Q+L	78.6 [66.1, 91.1]	62.3 [51.5, 73.2]
BIOT	E+Q	69.0 [55.9, 82.1]	49.9 [41.7, 58.1]
BIOT	E+Q+L	74.0 [62.6, 85.4]	61.3 [50.2, 72.5]
LaBraM	E+Q	61.8 [49.9, 73.8]	52.6 [43.4, 61.8]
LaBraM	E+Q+L	67.4 [55.5, 79.4]	59.9 [50.5, 69.2]
CBraMod	E+Q	70.2 [59.3, 81.1]	55.3 [47.2, 63.5]
CBraMod	E+Q+L	73.9 [63.1, 84.7]	63.1 [54.4, 71.7]