OMNIBRAIN: AN EXPERIENCE-DRIVEN AGENT FOR EVIDENCE-GROUNDED CLINICAL EEG DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) diagnosis requires selecting measurements and interpreting ambiguous findings in clinical context. Foundation models provide diagnostic predictions, but these outputs do not specify which findings warrant further checks or how to assess them in context. Learning reusable diagnostic guidance is challenging because case labels evaluate final diagnoses, leaving intermediate analysis decisions unlabeled. We introduce OMNIBRAIN, an EEG agent that learns interpretation and verification guidance from diagnostic feedback without updating model weights. Within each case, a Clinical Evidence Verification Loop (CEVL) organizes hypothesis-driven reasoning: stating expectations before analysis, interpreting measurements in context, and pursuing follow-up checks. The resulting diagnostic records are represented as Diagnostic Inference Graphs (DIGs) that preserve measurement context and evidence dependencies. Comparing successful and failed DIGs from the same recording holds the input fixed, exposing interpretation differences and unmatched follow-up checks. An LLM converts these contrasts into reusable guidance for subsequent cases. At inference, the learned experience guides signal analysis, after which the agent consults foundation-model predictions to form its final diagnosis. Against five frozen-backbone foundation models and their majority vote, OMNIBRAIN achieves the highest task-level balanced accuracy on five of six EEG tasks. DIG-based extraction outperforms direct extraction on both evaluated tasks, and experience learned from one Parkinson's disease dataset transfers across five held-out datasets.

1 INTRODUCTION

Clinical EEG diagnosis is a sequence of evidence-seeking and evidence-interpretation decisions. Reading a recording requires choosing which segments to examine, which spectral or spatial analyses to run, and how to interpret ambiguous patterns in the patient's clinical context (Gélisse et al., 2024; Benbadis & Lin, 2008; Tatum, 2021). Checking a pattern's distribution across channels or persistence across segments helps determine whether to retain an interpretation or investigate another explanation.

Foundation-model predictions and signal measurements offer distinct sources of evidence for EEG diagnosis. EEG foundation models learn representations from unlabeled corpora and are fine-tuned on labeled recordings to predict diagnostic labels (Jiang et al., 2024; Wang et al., 2025; Yang et al., 2023; Zhou et al., 2025; Kostas et al., 2021). Tool-using agents obtain signal findings through executable analyses (Kim et al., 2025; Zhao et al., 2026; Hou et al., 2026; Wang et al., 2026a). However, access to predictions and measurements does not turn past diagnostic outcomes into reusable guidance for choosing checks and interpreting these sources of evidence.

Clinicians build diagnostic experience through clinical practice and feedback on their judgments (Schmidt & Rikers, 2007). Revisiting how findings were interpreted and which checks informed a diagnosis helps refine decisions on subsequent cases. This suggests a route for EEG agents to learn from past diagnoses: examine the reasoning behind each outcome and extract guidance for interpreting findings and selecting further checks.

We study experience learning with case labels as feedback on final diagnoses. This setting poses two difficulties. (i) The case label evaluates the final diagnosis, but does not identify which intermediate interpretations or verification decisions contributed to that outcome (Lightman et al., 2023).

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.

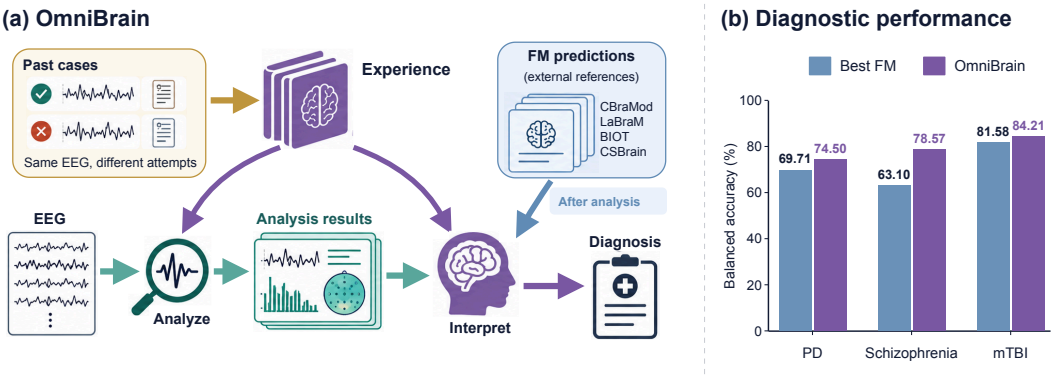


Figure 1: (a) Experience from past diagnostic feedback guides EEG analysis and interpretation. After analyzing the signal, OMNIBRAIN consults individual FM predictions as references for its final diagnosis. (b) Balanced accuracy (%) on three selected tasks, comparing OMNIBRAIN with the best individual frozen-backbone FM: CBraMod for PD, LaBraM for schizophrenia on Albrecht2019, and CSBrain for mTBI.

Learning useful guidance therefore requires relating these decisions to the measurements and analysis conditions on which they were based. (ii) EEG signals differ across subjects (He & Wu, 2020), complicating comparisons of diagnostic judgments across recordings. When both the signal and the analysis change, different diagnostic outcomes may reflect recording difficulty rather than the quality of the analysis.

We introduce OMNIBRAIN, a framework for EEG diagnosis with foundation models and signal-analysis tools that learns reusable experience from diagnostic feedback. Within each case, the Clinical Evidence Verification Loop (CEVL) organizes hypothesis-driven analysis, contextual interpretation, and follow-up verification, linking diagnostic judgments to expectations and measured evidence. Completed training records are represented as typed Diagnostic Inference Graphs (DIGs) that preserve these relationships and their measurement context for comparison across attempts. Comparing successful and failed DIGs from the same recording holds the input fixed, so that contrasts reflect analysis differences rather than signal variation. An LLM uses this evidence to formulate skill items that guide how observations are evaluated and which analyses to perform. The experience library accumulates these items for subsequent cases while model weights remain fixed.

We evaluate OMNIBRAIN on six diagnosis-related EEG classification tasks covering Parkinson's disease, schizophrenia, AD/FTD, OCD, mTBI, and depression. Compared with five frozen-backbone EEG foundation models and their majority vote, OMNIBRAIN achieves the highest task-level balanced accuracy on five of six tasks. Controlled ablations show that DIG-based extraction outperforms direct extraction from the same diagnostic records with absolute BA gains of 12.42% and 23.81% on the two evaluated tasks (Figure 7). Experience learned from one Parkinson's disease dataset improves BA across five held-out Parkinson's disease datasets by an absolute 12.11% on average.

Our contributions are as follows:

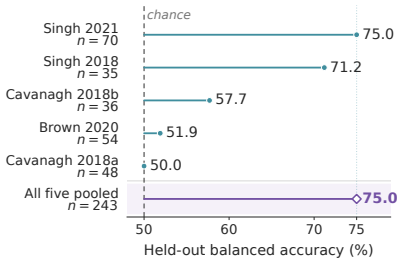
- We study learning diagnostic guidance from case-level EEG feedback, where labels evaluate final diagnoses but leave individual analysis decisions unlabeled.
- We propose OMNIBRAIN, in which CEVL structures diagnostic reasoning around hypotheses, expected observations, and evidence verification, while an LLM learns interpretation and verification guidance by comparing same-recording DIGs without weight updates (§4).
- We evaluate OMNIBRAIN on six EEG tasks, demonstrating the effectiveness of DIG-based experience extraction and cross-dataset transfer (§5).



(a) Coding-agent predictions on schizophrenia and OCD.

	CBraMod	LaBraM	BIOT	CSBrain	BENDR	Majority vote	Best FM
Cavanagh 2018a	**59.4**	56.3	34.4	50.0	46.9	50.0	59.4
Cavanagh 2018b	56.3	56.3	43.8	**75.0**	43.8	56.3	75.0
Singh 2018	68.8	**87.5**	62.5	81.3	68.8	81.3	87.5
Brown 2020	71.9	**75.0**	65.6	**75.0**	68.8	78.1	75.0
Singh 2021	70.4	**78.3**	69.6	64.2	56.7	74.6	78.3
Mean	**65.3**	**70.7**	**55.2**	**69.1**	**57.0**	**68.0**	**75.0**

(b) End-to-end fine-tuned FM performance across Parkinson's disease datasets.



(c) Transfer from different training sources.

Figure 2: Prediction bias and variation across datasets and training sources. (a) OCD labels denote symptom groups. (b) Blue cells mark the best individual models; Best FM is the highest model BA within each dataset. (c) Frozen-backbone CBraMod transfer to Singh2020.

2 Related Work

Learning EEG representations. EEG foundation models learn signal representations that transfer across clinical tasks (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024; 2025; Zhou et al., 2025). EEG-to-language models generate reports from signal representations, sometimes using retrieved examples as context (Pradeepkumar et al., 2026; Wang et al., 2026b; Zhu et al., 2026). We study a complementary learning target, the diagnostic judgments used to interpret EEG measurements and select further checks.

Executing grounded EEG analyses. Tool-using EEG agents organize diagnosis around executable analyses (Kim et al., 2025; Zhao et al., 2026). CogEEGAgent registers analyses before observing outcomes (Hou et al., 2026). Feedback can also revise the analysis procedure itself, as in NeuroWeaver's pipeline search and EEG-SpikeAgent's feature refinement (Wang et al., 2026a; Santhosh et al., 2026). Building on explicit tool execution, CEVL guides the agent to compare observations with prior expectations, interpret them in context, and pursue follow-up checks.

Learning from agent experience. Agent-memory methods extract reusable insights, reasoning strategies, or skills from successful and failed trajectories (Zhao et al., 2024; Ouyang et al., 2026; Ni et al., 2026). GSEM compares successful and failed trajectories from the same training instance (Han et al., 2026). Its dual-layer graph organizes the extracted experiences and their relationships. In OmniBrain, DIGs instead represent the diagnostic decisions being compared, linking measurements to interpretation and verification decisions before reusable guidance is formulated from their contrasts.

3 Motivating Observations

Signal findings need targeted verification. Given raw EEG and signal-processing tools without worked examples or diagnostic experience, Claude Code assigns positive labels to 9 of 11 negative-class recordings across schizophrenia (Albrecht et al., 2019) and OCD (Gründler et al., 2009) tasks (Figure 2a). The negative class comprises healthy controls in schizophrenia and participants with low or moderate symptoms in OCD. In a representative negative-class case, it uses a low posterior alpha

peak and an elevated frontal theta/beta ratio to support a positive reading. Although it acknowledges possible artifact contamination, it makes no further check to distinguish that explanation from the proposed diagnostic interpretation before concluding.

Foundation models provide diagnostic references. Foundation models achieve strong performance on some Parkinson's disease datasets: CBraMod reaches 71.9% balanced accuracy (BA) on Brown2020 (Figure 2b). On the fixed Singh2020 test set, CBraMod's BA ranges from 50.0% to 75.0% across classification-head training sources with the backbone frozen (Figure 2c). The strongest individual model outperforms majority voting in mean BA on these datasets (Figure 2b).

These observations motivate a diagnostic reasoning process that uses foundation-model predictions as references and examines the EEG evidence needed to assess them.

4 OMNIBRAIN

Problem setup. We aim to extract useful information from past diagnostic analyses to improve accuracy on subsequent cases. Let $\mathcal{D}_{\text{past}}$ contain diagnostic records and their case-level outcome feedback, with potentially multiple attempts per case. This feedback evaluates final diagnoses rather than individual judgments. An extraction procedure $\mathcal{F}$ produces an experience library $\mathcal{S}$ containing reusable guidance and its supporting analyses. For task k, the learning objective is

$$\max_{\mathcal{F}} \; J_k(\mathcal{F}), \qquad J_k(\mathcal{F}) = \mathbb{E}_{(x,y)\sim P_k}\mathbb{E}_{\hat{y}\sim\mathcal{A}(\cdot|x,k,\mathcal{S})}\big[\mathbf{1}\{\hat{y}=y\}\big], \quad \mathcal{S} = \mathcal{F}(\mathcal{D}_{\text{past}}). \tag{1}$$

Here, x comprises an EEG recording $\mathbf{X}$ and recording metadata, y is its target label, and P_k is the distribution of subsequent cases. $\mathcal{A}$ denotes the agent, and $\mathcal{A}(\cdot \mid x, k, \mathcal{S})$ is its conditional predictive distribution. The guidance specifies when it applies and remains linked to supporting measurements and judgments for inspection. The objective expresses predictive accuracy, while evaluation uses balanced accuracy for the class-imbalanced tasks.

4.1 WITHIN-CASE DIAGNOSTIC REASONING

Clinical Evidence Verification Loop. CEVL is a diagnostic reasoning pattern that connects hypothesis formation, expectation-guided analysis, contextual interpretation, and follow-up verification (Algorithm 2). The EEG Diagnostic Environment provides signal analysis, waveform inspection, preprocessing, and knowledge retrieval. The agent initializes hypotheses from clinical guidelines, literature, and the incoming experience library. Verification items guide check selection, while interpretation items guide the assessment of observations. For a hypothesis h_t, the agent first states an investigation $q_t = (\hat{o}_t, a_t)$ comprising the expected observation $\hat{o}_t$ and analysis a_t. The EEG Diagnostic Environment executes the analysis on selected channels, intervals, or frequency bands and returns observation o_t with its provenance. The agent's interpretation r_t evaluates the measurement against the expectation in clinical context and records its relation to the hypothesis.

An interpretation that raises an unresolved question can introduce a follow-up hypothesis h' and prompt another check. The loop ends when no hypotheses remain or the cycle budget is reached. During experience learning, the agent synthesizes the diagnosis from signal interpretations alone. At test inference, after the signal-analysis loop, it retrieves the available foundation-model predictions once and interprets each alongside the signal findings before synthesizing the diagnosis. The completed diagnostic record R retains these hypotheses, observations, interpretations, and their relationships.

Diagnostic Inference Graphs. A Diagnostic Inference Graph (DIG) $G = (V, E)$ organizes the diagnostic record for comparison by linking each judgment to the evidence it assesses and the checks it prompts. Its four node types are hypothesis, observation, interpretation, and diagnosis, with each hypothesis node specifying its investigation. Structural edges preserve measurement dependencies, follow-up branches, and contributions to the final diagnosis. Evaluative edges connect interpretations to the hypotheses they assess. Graph construction follows references in the diagnostic record without access to the ground-truth label, retaining source spans and observation provenance.

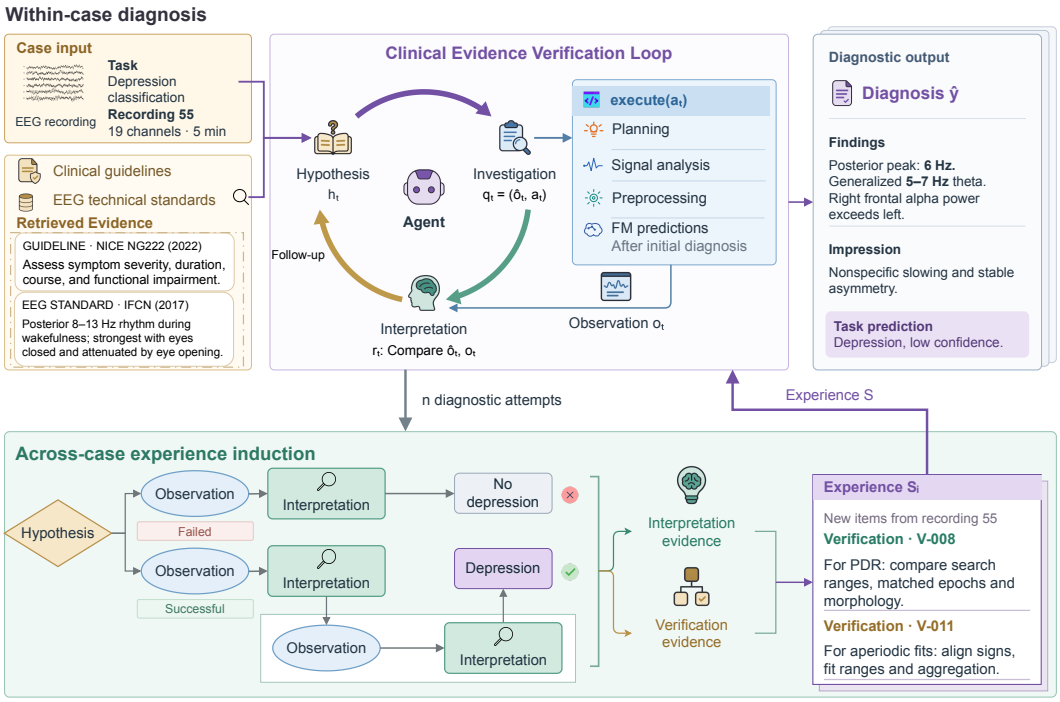


Figure 3: Overview of OMNIBRAIN. Top: CEVL guides within-case reasoning from hypotheses and expected observations through signal analysis, contextual interpretation, and follow-up verification to diagnosis. Bottom: successful and failed attempts on the same recording are represented as DIGs and compared to extract interpretation and verification guidance. The resulting experience library guides analysis and interpretation in subsequent cases.

4.2 ACROSS-CASE EXPERIENCE INDUCTION

Graph-Based Skill Evidence. Experience is extracted from contrasts within each training case and accumulated across cases. Each training case receives n diagnostic attempts using the same case input and incoming library $\mathcal{S}_{i-1}$, producing graphs $\{G_{ij}\}_{j=1}^{n}$. Only after all graphs are constructed is y_i revealed to separate successful graphs $\mathcal{G}_i^+$, whose predictions match y_i, from failed graphs $\mathcal{G}_i^-$. Every pair in $\mathcal{G}_i^+ \times \mathcal{G}_i^-$ is compared. An LLM aligns nodes of the same type using their content, neighboring edges, and measurement provenance, leaving nodes without a counterpart unmatched.

These comparisons yield evidence for skill induction, consisting of contrasts in how diagnostic observations are obtained or interpreted. **Interpretation evidence** captures different assessments of comparable observations against the same hypothesis. **Verification evidence** captures differences in measurement definitions or investigation procedures, including a follow-up check present in the successful graph but absent from the failed one. Retaining unmatched branches preserves these differences instead of restricting comparison to shared steps. Semantically equivalent evidence is deduplicated within the recording into $\Delta_i = \Delta_i^{\text{interpret}} \cup \Delta_i^{\text{verify}}$. Each evidence item retains the contrasting decisions and their source context for skill formation.

Evidence examples.

Interpretation evidence. The same frontal low-frequency increase is interpreted as pathological slowing in one analysis and a possible ocular artifact in another.

Verification evidence. Both analyses find left–right power asymmetry, but only one checks channel quality before using it in diagnosis.

Feedback-Guided Experience Accumulation. The extracted evidence Δ_i is converted into reusable instructions in a once-per-case library update that also records feedback on existing items. Diagnostic outcomes determine which attempts are contrasted, while the source measurements and

Algorithm 1 Across-Case Experience Induction

Require: Agent $\mathcal{A}$, training cases $\{(x_i, y_i)\}_{i=1}^{N}$, task k, attempts n, budget $T_{\max}$
Ensure: Experience library $\mathcal{S}_N$

```
Initialize S_0 with empty Interpretation and Verification sections
for i = 1, ..., N do
    for j = 1, ..., n do
        (ŷ_ij, R_ij, U_ij) ← CEVL(A, x_i, k, S_{i−1}, T_max, learning)
        Build G_ij from the references and provenance in R_ij
    end for
    Reveal y_i and partition {G_ij}_{j=1}^n into G_i^+ and G_i^−
    Δ_i ← ∅
    for all (G^+, G^−) ∈ G_i^+ × G_i^− do
        Align same-type nodes by content, neighboring edges, and measurement provenance
        Add interpretation contrasts and differences in checks, including unmatched branches, to Δ_i
    end for
    Merge equivalent evidence in Δ_i and use an LLM to formulate new skill items
    Form S_i by appending new items to S_{i−1} and updating reported-item counts from all n diagnostic outcomes
end for
return S_N for inference with the library frozen
```

reasoning determine the guidance to extract. For each evidence item, an LLM formulates an applicability condition and an instruction for interpreting a finding or performing a check. New items enter the Interpretation or Verification section of a single skill document, leaving existing items unchanged. Each attempt reports used-item identifiers U_{ij}, with usage counts recording whether its final diagnosis was correct or incorrect. The resulting library $\mathcal{S}_i$, including these counts, guides check selection and interpretation on the next case. After the final training case, $\mathcal{S}_N$ is frozen for test inference. Experience updates change the library while keeping model weights fixed. Algorithm 1 summarizes this process.

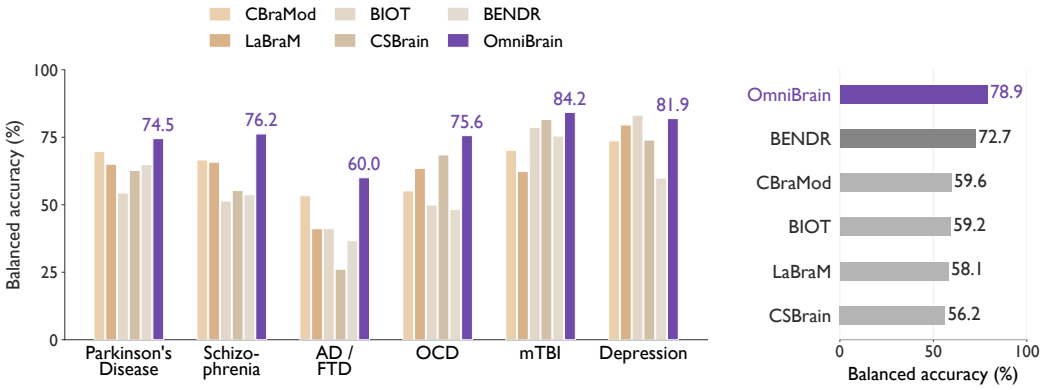
5 Experiments

5.1 Experimental Settings

Datasets. We evaluate six diagnosis-related EEG classification tasks with subjects separated between the training and test sets, covering Parkinson's disease, schizophrenia, AD/FTD, OCD, mTBI, and depression. AD/FTD distinguishes Alzheimer's disease, frontotemporal dementia, and healthy controls; the remaining tasks use binary labels. OCD and depression each include clinical classification and symptom grouping (Sections B.4 and B.5). Schizophrenia includes Albrecht2019 and ASEEG (Section B.3). The external mTBI evaluation applies the frozen Cavanagh2019 checkpoints to 50 DPX recordings from previously unused participants (Table 12). Recordings are resampled to 200 Hz, and Singh2020 is excluded from training and experience construction for cross-dataset evaluation.

Baselines. The main system comparison uses five EEG foundation models loaded from public pretrained weights with frozen backbones: CBraMod (Wang et al., 2025), LaBraM (Jiang et al., 2024), BIOT (Yang et al., 2023), CSBrain (Zhou et al., 2025), and BENDR (Kostas et al., 2021). Only the classification head is trained, using one 50-epoch run with seed 42, matching the weight-frozen setting of OmniBrain. Dataset splits and baseline configurations are provided in Appendices B and C; an end-to-end fine-tuning comparison appears in Section C.9. Majority vote combines one recording-level prediction from each of these five models, using the same checkpoints and test recordings as the individual baselines, and resolves ties by summed class probabilities.

Agent configuration. OmniBrain uses GPT-5.6-Sol at temperature 0.3 to construct and use a separate experience library for each task. Within each task and phase, the prompt template and Diagnostic Environment remain fixed across cases. The limits per diagnostic attempt are 100 tool calls and 20 CEVL cycles. Each training recording receives four diagnostic attempts with foundation-model predictions disabled before a single library update. Test inference uses the frozen final library and one diagnostic attempt per recording, with available foundation-model predictions consulted after signal analysis. Agent scores are averaged over three evaluation runs per condition.



(a) Within-source evaluation.

(b) Cross-dataset evaluation.

Figure 4: Balanced accuracy (%) of OMNIBRAIN and five EEG foundation models. (a) Within-source test performance across six diagnosis-related classification tasks, with frozen backbones and the task-level aggregation of Table 1. (b) Cross-dataset performance on Singh2020, excluded from training and experience construction; each foundation-model score is the mean BA of five end-to-end fine-tuned source-dataset checkpoints. Purple bars denote OMNIBRAIN.

Table 1: Within-source performance on diagnosis-related EEG classification tasks (%). The five foundation models use frozen backbones. MV (majority vote) combines the five models' recording-level predictions. **Bold**: best; underline: second best. Parkinson averages five datasets; schizophrenia, OCD, and depression each average two datasets or settings.

Task	Metric	CBraMod	LaBraM	BIOT	CSBrain	BENDR	MV	OMNIBRAIN
Parkinson's Disease	BA	69.71	65.00	54.38	62.71	64.92	61.33	**74.50**
	κ	38.53	26.37	7.91	22.91	29.42	20.94	**45.22**
	F1	71.33	61.78	53.46	62.76	66.27	62.39	**73.04**
Schizophrenia	BA	66.59	65.75	51.40	55.27	53.71	70.75	**76.22**
	κ	33.12	31.26	2.92	11.13	7.98	41.12	**51.40**
	F1	65.92	64.28	51.04	47.32	47.52	69.34	**74.95**
AD / FTD	BA	53.33	41.11	41.11	26.11	36.67	23.33	**60.00**
	κ	36.84	13.40	12.50	−5.88	8.70	−6.93	**47.83**
	F1	55.49	42.79	42.33	27.56	38.33	28.70	**60.00**
OCD	BA	55.12	63.45	49.88	68.45	48.21	63.45	**75.60**
	κ	10.71	26.91	−0.59	39.00	−0.51	26.91	**54.01**
	F1	57.15	64.66	48.08	69.85	37.57	64.66	**77.45**
mTBI	BA	70.18	62.28	78.65	81.58	75.44	83.92	**84.21**
	κ	38.12	21.43	50.00	52.43	50.88	**62.37**	58.21
	F1	72.06	62.01	75.82	75.68	78.57	**82.65**	79.24
Depression	BA	73.65	79.55	**83.19**	73.97	59.86	80.75	81.91
	κ	48.13	59.61	**67.10**	50.27	20.46	62.70	63.27
	F1	75.63	81.17	**84.66**	76.10	61.61	82.45	82.12

Evaluation. Within each dataset, all methods predict one label per recording on the same test set. For foundation models, we average class probabilities across non-overlapping analysis windows and select the class with the highest mean probability. Frozen-backbone and end-to-end checkpoints are selected using recording-level and window-level validation, respectively. For foundation models on Singh2020, we apply the same mean-pooling aggregation and report each model's mean BA across five source-dataset checkpoints. For OMNIBRAIN, the label is the final diagnosis from its test-time attempt. Balanced accuracy (BA) is the primary comparison metric, supplemented by Cohen's κ and support-weighted F1. For tasks with multiple datasets or classification settings, each metric is averaged equally across them; details are given in Tables 9 to 11.

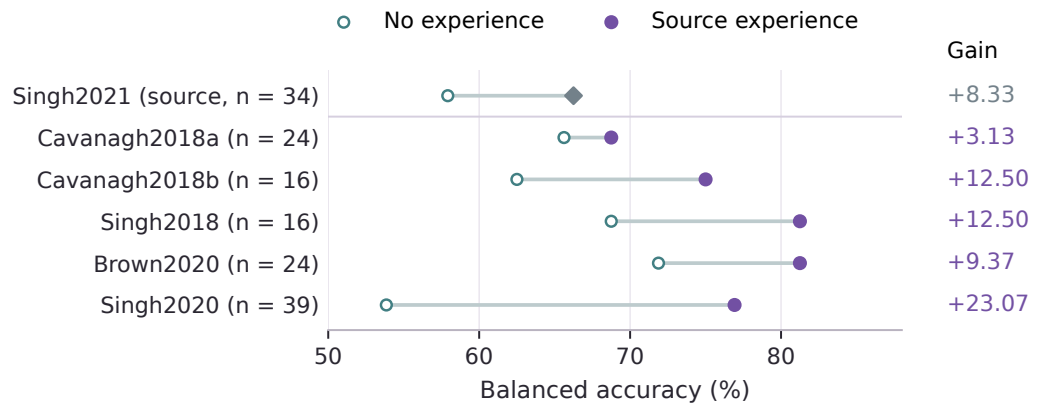


Figure 5: BA (%) with and without frozen Singh2021 experience under CEVL. Singh2021 is the held-out source reference; the remaining five datasets are transfer targets. Right-hand values show absolute BA gains.

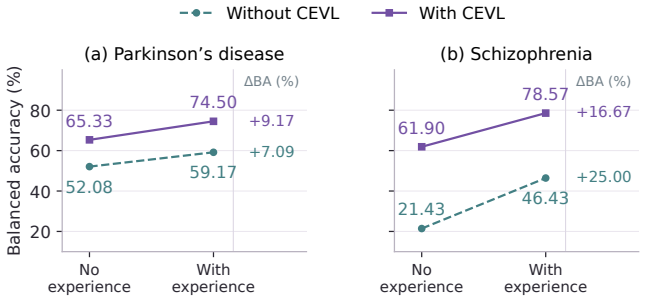


Figure 6: CEVL and experience ablation on (a) Parkinson's disease and (b) schizophrenia (Albrecht2019). Lines compare BA (%) with and without CEVL; ΔBA shows the gain from adding experience.

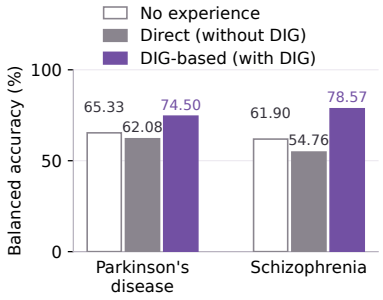


Figure 7: Experience extraction under CEVL on Parkinson's disease and schizophrenia (Albrecht2019): no experience, direct extraction, and DIG-based extraction. Bars show BA (%).

5.2 Overall Diagnostic Performance

Without updating any model parameters, OmniBrain achieves the highest task-level BA on five of six tasks against five frozen-backbone EEG foundation models and their majority vote (Table 1). The largest gains are on OCD (75.60% vs. 68.45%) and AD/FTD (60.00% vs. 53.33%).

Cross-dataset system performance. Using experience from the five Parkinson's disease source datasets, OmniBrain achieves 78.85% BA on Singh2020, compared with 72.7% for the best FM, BENDR, averaged across five source-dataset checkpoints (Figure 4b). Each FM checkpoint in this comparison is trained on one source dataset. Pooled-source results are reported separately in Table 13.

Transfer of diagnostic experience. With CEVL fixed, experience learned from the Singh2021 training set transfers across five held-out Parkinson's disease datasets: adding the frozen library improves BA on every target, by an absolute 12.11% on average (Figure 5).

5.3 Ablation Study

CEVL and experience. Adding experience to CEVL improves BA by an absolute 9.17% and 16.67% on Parkinson's disease and schizophrenia, yielding the best ablation configuration on both tasks (Figure 6). All four configurations use the same model, tools, and budget limits, with a shared frozen library for the two experience-enabled configurations.

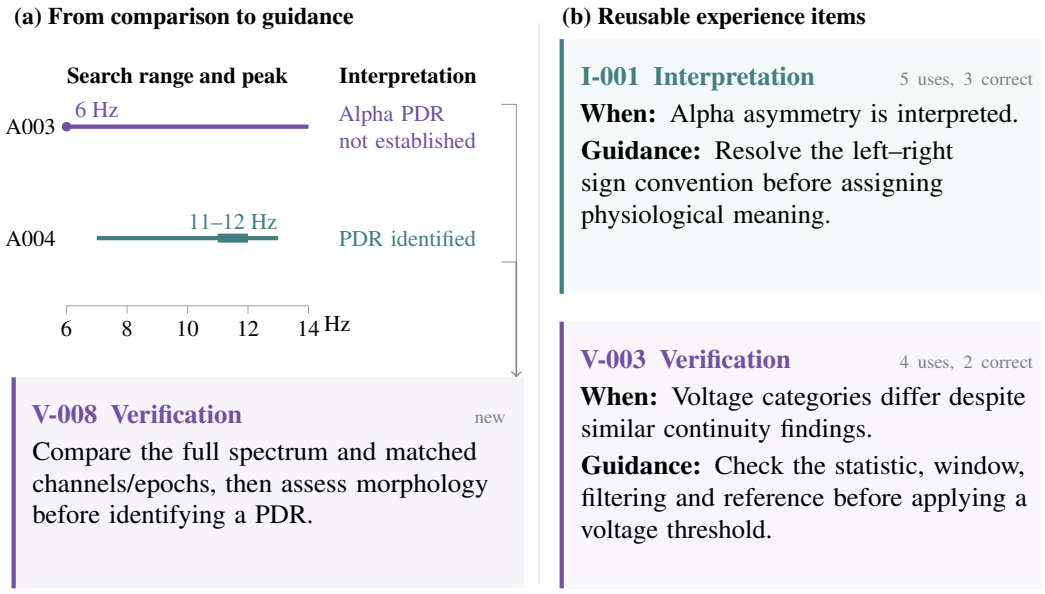


Figure 8: (a) Measurement contrasts for a verification rule: lines show search ranges and markers show reported peaks. (b) Applicability conditions and actions from the incoming experience library.

DIG-based experience extraction. Direct extraction (without DIG) and DIG-based extraction use the same diagnostic records and successful–failed attempt pairs, with matched extraction settings and library budgets under CEVL evaluation (Figure 7). DIG-based extraction improves BA over direct extraction by an absolute 12.42% and 23.81% on Parkinson's disease and schizophrenia, respectively. Only DIG-based extraction improves on the no-experience reference on both tasks.

6 Analysis

Graph comparison exposes measurement differences behind judgments. On Mumtaz2016 recording 55, two attempts disagree about the posterior dominant rhythm (PDR) (Figure 8a). A003 finds a 6 Hz maximum within 6–14 Hz and does not establish an alpha PDR, whereas A004 finds an 11–12 Hz maximum within 7–13 Hz and identifies a PDR. Tracing each interpretation through its observation to the investigation reveals that the narrower search excludes the slower peak. Verification item `V-008` accordingly calls for checking search ranges, matching channels and epochs, and assessing waveform morphology before identifying a PDR.

Experience retains decision rules beyond individual cases. The learned rules pair an applicability condition with an analysis action (Figure 8b). `I-001` calls for resolving the sign convention of alpha asymmetry before assigning physiological meaning. `V-003` requires the amplitude statistic and its measurement conditions to be identified before a voltage threshold is applied. These conditions can recur on later recordings with different measurements.

7 Conclusion

We presented OmniBrain, a framework that uses foundation models and signal-analysis tools for EEG diagnosis and accumulates experience from diagnostic feedback without updating model weights. CEVL structures diagnostic reasoning around hypotheses, expected observations, and evidence verification, while same-recording DIG comparisons yield interpretation and verification guidance for later cases. Against five frozen-backbone baselines and their majority vote, the system achieves the highest task-level balanced accuracy on five of six tasks. DIG-based extraction outperforms direct extraction from the same diagnostic records, and experience learned from one Parkinson's disease dataset transfers across five held-out Parkinson's disease datasets.

AI USE STATEMENT

We used generative AI tools for manuscript polishing, literature searches and citation checks, conceptual and methodological discussions, experimental feedback, result interpretation, and figure and code preparation. The authors reviewed the AI-assisted manuscript edits for accuracy and consistency with the intended meaning. The authors retained responsibility for research decisions and conclusions and take full responsibility for the final text, claims, results, and artifacts.

ETHICS STATEMENT

This study analyzes existing publicly available EEG datasets, with sources and data processing described in Appendix B. No new participants were recruited or clinical interventions performed. OMNIBRAIN is evaluated for research on EEG classification, not for use in patient care.

REFERENCES

Matthew A. Albrecht, James A. Waltz, James F. Cavanagh, Michael J. Frank, and James M. Gold. Increased conflict-induced slowing, but no differences in conflict-induced positive or negative prediction error learning in patients with schizophrenia. *Neuropsychologia*, 123:131–140, 2019. doi: 10.1016/j.neuropsychologia.2018.04.031.

American Clinical Neurophysiology Society. Guideline 7: Guidelines for writing EEG reports. *Journal of Clinical Neurophysiology*, 23(2):118–121, 2006. doi: 10.1097/00004691-200604000-00008.

Claudio Babiloni, Robert J. Barry, Erol Başar, Katarzyna J. Blinowska, Andrzej Cichocki, Wilhelmus H. I. M. Drinkenburg, Wolfgang Klimesch, Robert T. Knight, Fernando Lopes da Silva, Paul Nunez, Robert Oostenveld, Jaeseung Jeong, Roberto Pascual-Marqui, Pedro Valdes-Sosa, and Mark Hallett. International federation of clinical neurophysiology (IFCN)–EEG research workgroup: Recommendations on frequency and topographic analysis of resting state EEG rhythms. part 1: Applications in clinical research studies. *Clinical Neurophysiology*, 131(1):285–307, 2020. doi: 10.1016/j.clinph.2019.06.234.

Sara Bagherzadeh, Mohammadreza Norouzi, Fatemeh Farokhshad, Pouya Tolou Kouroshi, Amirhesam Ghasri, Reza Kazemi, Reza Rostami, and Ahmad Shalbaf. Atieh schizophrenia EEG, a novel high-quality dataset designed to advance biomarker and machine learning research. *Scientific Data*, 13:1211, 2026. doi: 10.1038/s41597-026-07968-9.

Selim R. Benbadis and Kaiwen Lin. Errors in EEG interpretation and misdiagnosis of epilepsy. *European Neurology*, 59(5):267–271, 2008. doi: 10.1159/000115641.

Sándor Beniczky, Harald Aurlien, Jan C. Brøgger, et al. Standardized computer-based organized reporting of EEG: SCORE. *Epilepsia*, 54(6):1112–1124, 2013. doi: 10.1111/epi.12135.

Darin R. Brown, Sarah Pirio Richardson, and James F. Cavanagh. An EEG marker of reward processing is diminished in Parkinson's disease. *Brain Research*, 1727:146541, 2020. doi: 10.1016/j.brainres.2019.146541.

James F. Cavanagh. EEG: DPX cog ctl task in acute mild TBI. OpenNeuro, 2024. URL https://openneuro.org/datasets/ds005114/versions/1.0.0. Dataset ds005114, version 1.0.0.

James F. Cavanagh, Arthur Napolitano, Christopher Wu, and Abdullah Mueen. The patient repository for EEG data + computational tools (PRED+CT). *Frontiers in Neuroinformatics*, 11:67, 2017. doi: 10.3389/fninf.2017.00067.

James F. Cavanagh, Andrew W. Bismark, Michael J. Frank, and John J. B. Allen. Multiple dissociations between comorbid depression and anxiety on reward and punishment processing: Evidence from computationally informed EEG. *Computational Psychiatry*, 3:1–17, 2019a. doi: 10.1162/cpsy_a_00024.

James F. Cavanagh, J. Kevin Wilson, Rebecca E. Rieger, Darbi Gill, James M. Broadway, Jacqueline Hope Story Remer, Violet Fratzke, Andrew R. Mayer, and Davin K. Quinn. ERPs predict symptomatic distress and recovery in sub-acute mild traumatic brain injury. *Neuropsychologia*, 132:107125, 2019b. doi: 10.1016/j.neuropsychologia.2019.107125.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

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, Main Conference, pp. 22541–22577. Curran Associates, Inc., 2025. doi: 10.52202/085713-0760.

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. AutoGuide: Automated generation and selection of context-aware guidelines for large language model agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119919–119948, 2024. doi: 10.52202/079017-3811.

Philippe Gélisse, Selim R. Benbadis, Arielle Crespel, and William O. Tatum. Overcoming traps and pitfalls leading to misinterpretation of normal EEG variants and variation of the background activity. *Journal of Neurology*, 271(7):3869–3878, 2024. doi: 10.1007/s00415-024-12440-y.

Théo Gnassounou, Rémi Flamary, and Alexandre Gramfort. Convolution Monge mapping normalization for learning on sleep data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10457–10476, 2023. doi: 10.52202/075280-0459.

Jean Gotman. Automatic recognition of epileptic seizures in the EEG. *Electroencephalography and Clinical Neurophysiology*, 54(5):530–540, 1982. doi: 10.1016/0013-4694(82)90038-4.

Jean Gotman and Pierre Gloor. Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG. *Electroencephalography and Clinical Neurophysiology*, 41(5):513–529, 1976. doi: 10.1016/0013-4694(76)90063-8.

Theo O. J. Gründler, James F. Cavanagh, Christina M. Figueroa, Michael J. Frank, and John J. B. Allen. Task-related dissociation in ERN amplitude as a function of obsessive-compulsive symptoms. *Neuropsychologia*, 47(8–9):1978–1987, 2009. doi: 10.1016/j.neuropsychologia.2009.03.010.

Xiao Han, Yuzheng Fan, Sendong Zhao, Haochun Wang, and Bing Qin. GSEM: Graph-based self-evolving memory for experience augmented clinical reasoning. *arXiv preprint arXiv:2603.22096*, 2026.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

Lawrence J. Hirsch, Michael W. K. Fong, Markus Leitinger, et al. American Clinical Neurophysiology Society's standardized critical care EEG terminology: 2021 version. *Journal of Clinical Neurophysiology*, 38(1):1–29, 2021. doi: 10.1097/WNP.0000000000000806.

Dengzhe Hou, Lingyu Jiang, Fangzhou Lin, and Kazunori D. Yamada. CogEEGAgent: Toward autonomous cognitive EEG analysis with grounded execution and selection-aware verification. *arXiv preprint arXiv:2607.25045*, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024.

Ayush Rajesh Jhaveri, Anthony GX-Chen, Ilia Sucholutsky, and Eunsol Choi. Failing to falsify: Evaluating and mitigating confirmation bias in language models. *arXiv preprint arXiv:2604.02485*, 2026.

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.

N. Kane, J. Acharya, S. Beniczky, L. Caboclo, S. Finnigan, P. W. Kaplan, H. Shibasaki, R. Pressler, and M. J. A. M. van Putten. A revised glossary of terms most commonly used by clinical electroencephalographers and updated proposal for the report format of the EEG findings. Revision 2017. *Clinical Neurophysiology Practice*, 2:170–185, 2017. doi: 10.1016/j.cnp.2017.07.002.

Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. EEG-Bench: A benchmark for EEG foundation models in clinical applications. *arXiv preprint arXiv:2512.08959*, 2025.

Guzal Khayrullina, Krystsina Liaukovich, Galina Portnova, and Olga Martynova. Raw data of the antisaccade task in healthy volunteers (EEG, eyetracking, vegetatics), 2024a. URL https://doi.org/10.17632/h8xrwrmttj.1.

Guzal Khayrullina, Krystsina Liaukovich, Galina Portnova, and Olga Martynova. Raw data of the antisaccade task in patients with obsessive-compulsive disorder (EEG, eyetracking, vegetatics), 2024b. URL https://doi.org/10.17632/mvyd3bgjv7.1.

Haechan Kim, Yoonho Lee, Gisang Lee, Chelsea Finn, and Kangwook Lee. WHALE: A simple recipe for joint harness-weight optimization. *arXiv preprint arXiv:2609.00196*, 2026. URL https://arxiv.org/abs/2609.00196.

Jonathan W. Kim, Ahmed Alaa, and Danilo Bernardo. Eeg-gpt: Exploring capabilities of large language models for eeg classification and interpretation. In *2025 12th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 951–956, 2025. doi: 10.1109/NER61569.2025.11589079.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-Harness: End-to-end optimization of model harnesses. *arXiv preprint arXiv:2603.28052*, 2026. URL https://arxiv.org/abs/2603.28052.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG-FM-Compass: Progress, benchmarking, and future directions for EEG foundation models. *National Science Review*, pp. nwag466, 2026. doi: 10.1093/nsr/nwag466.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6):95, 2023. doi: 10.3390/data8060095.

Wajid Mumtaz. MDD patients and healthy controls EEG data (new). Figshare dataset, 2016. URL https://figshare.com/articles/dataset/EEG_Data_New/4244171. Version 2.

National Institute for Health and Care Excellence. Psychosis and schizophrenia in adults: prevention and management (CG178), 2014. URL https://www.nice.org.uk/guidance/cg178.

National Institute for Health and Care Excellence. Dementia: assessment, management and support for people living with dementia and their carers (NG97), 2018. URL https://www.nice.org.uk/guidance/ng97.

National Institute for Health and Care Excellence. Depression in adults: treatment and management (NG222), 2022. URL https://www.nice.org.uk/guidance/ng222.

National Institute for Health and Clinical Excellence. Obsessive-compulsive disorder and body dysmorphic disorder: treatment (CG31), 2005. URL https://www.nice.org.uk/guidance/cg31.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*, 2026.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jL7fwchScm.

Ronald B. Postuma, Daniela Berg, Matthew Stern, Werner Poewe, C. Warren Olanow, Wolfgang Oertel, José Obeso, Kenneth Marek, Irene Litvan, Anthony E. Lang, Glenda Halliday, Christopher G. Goetz, Thomas Gasser, Bruno Dubois, Piu Chan, Bastiaan R. Bloem, Charles H. Adler, and Günther Deuschl. MDS clinical diagnostic criteria for parkinson's disease. *Movement Disorders*, 30(12):1591–1601, 2015. doi: 10.1002/mds.26424.

Jathurshan Pradeepkumar, Zheng Chen, and Jimeng Sun. Neural signals generate clinical notes in the wild. *arXiv preprint arXiv:2601.22197*, 2026.

Katya Rascovsky, John R. Hodges, David Knopman, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, 134(9):2456–2477, 2011. doi: 10.1093/brain/awr179.

Sonali Santhosh, Kelly Shuhong Yu, Eugene Chang, Jonathan Kim, Kie Shidara, and Danilo Bernardo. EEG-SpikeAgent: Agentic closed-loop program synthesis for automated EEG spike detection. *arXiv preprint arXiv:2607.04558*, 2026.

Henk G. Schmidt and Remy M. J. P. Rikers. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, 41(12):1133–1139, 2007. doi: 10.1111/j.1365-2923.2007.02915.x.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377.

Lívia Shirahige, Marina Berenguer-Rocha, Sarah Mendonça, Sérgio Rocha, Marcelo Cairrão Rodrigues, and Kátia Monte-Silva. Quantitative electroencephalography characteristics for parkinson's disease: A systematic review. *Journal of Parkinson's Disease*, 10(2):455–470, 2020. doi: 10.3233/JPD-191840.

Noah D. Silverberg, Grant L. Iverson, et al. The American Congress of Rehabilitation Medicine diagnostic criteria for mild traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, 104(8):1343–1355, 2023. doi: 10.1016/j.apmr.2023.03.036.

Arun Singh, Sarah Pirio Richardson, Nandakumar Narayanan, and James F. Cavanagh. Mid-frontal theta activity is diminished during cognitive control in Parkinson's disease. *Neuropsychologia*, 117:113–122, 2018. doi: 10.1016/j.neuropsychologia.2018.05.020.

Arun Singh, Rachel C. Cole, Arturo I. Espinoza, Darin Brown, James F. Cavanagh, and Nandakumar S. Narayanan. Frontal theta and beta oscillations during lower-limb movement in Parkinson's disease. *Clinical Neurophysiology*, 131(3):694–702, 2020. doi: 10.1016/j.clinph.2019.12.399.

Arun Singh, Rachel C. Cole, Arturo I. Espinoza, Aron Evans, Scarlett Cao, James F. Cavanagh, and Nandakumar S. Narayanan. Timing variability and midfrontal $\sim$4 Hz rhythms correlate with cognition in Parkinson's disease. *npj Parkinson's Disease*, 7(1):14, 2021. doi: 10.1038/s41531-021-00158-x.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

William O. Tatum. *Handbook of EEG Interpretation*. Demos Medical Publishing, 3rd edition, 2021. ISBN 9780826147080.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. Unveiling confirmation bias in chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3788–3804. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.195.

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, pp. 39249–39280, 2024. doi: 10.52202/079017-1239.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Guoan Wang, Shihao Yang, Jun-En Ding, and Feng Liu. Neuroweaver: An autonomous evolutionary agent for exploring the programmatic space of eeg analysis pipelines. *arXiv preprint arXiv:2602.13473v2*, 2026a.

Guoan Wang, Shihao Yang, Jun-en Ding, Hao Zhu, and Feng Liu. Neuronarrator: A generalist eeg-to-text foundation model for clinical interpretation via spectro-spatial grounding and temporal state-space reasoning. *arXiv preprint arXiv:2603.16880*, 2026b.

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 (ICLR)*, 2025.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. SkillRL: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593.

Zhiyuan Xu, Yueqing Dai, Junling Li, and Junwen Luo. What the LLM should not say: Boundary-aware context grounding for a seven-channel EEG agent. *arXiv preprint arXiv:2606.26519*, 2026.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (17):19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936.

Sha Zhao, Mingyi Peng, Haiteng Jiang, Tao Li, and Shijian Li. EEG Agent: A unified framework for automated EEG analysis using large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):18063–18071, 2026. doi: 10.1609/aaai.v40i21.38867.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Shijian Li, and Gang Pan. BrainAgent: A large language model-driven multi-agent framework for autonomous brain signal understanding. *arXiv preprint arXiv:2606.25400*, 2026.

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, Main Conference, pp. 87150–87195. Curran Associates, Inc., 2025. doi: 10.52202/085713-2915.

Yu Zhu, Jiayang Guo, Jun Jiang, Peipei Gu, Xin Shu, and Duo Chen. Explainable interictal epileptiform discharge detection method based on scalp eeg and retrieval-augmented generation. *arXiv preprint arXiv:2602.14170*, 2026.

APPENDIX CONTENTS

A Extended Related Work **17**

B Datasets and Splits **19**

B.1 Datasets, tasks, and acquisition . . . 19
B.2 Subjects and split construction . . . 19
B.3 Additional schizophrenia data . . . 20
B.4 Clinical OCD classification . . . 20
B.5 Depressive-symptom classification . . . 21
B.6 External mTBI data . . . 21
B.7 Data use across experiments . . . 21
B.8 Preprocessing and data representations . . . 21

C Experimental Protocols and Supplementary Results **22**

C.1 Foundation-model inputs and architectures . . . 22
C.2 Foundation-model training and evaluation . . . 23
C.3 Agent learning and inference . . . 24
C.4 Dataset-level classification results . . . 24
C.5 External mTBI transfer results . . . 25
C.6 Foundation-model transfer to Singh2020 . . . 25
C.7 Ablation protocols . . . 25
C.8 Experience transfer and ablation results . . . 26
C.9 Full Fine-Tuning Comparison . . . 26
C.10 Released run evidence . . . 27
C.11 Pairwise Disagreement on Parkinson Evaluation Sets . . . 27
C.12 Supplementary Results for Motivating Observations . . . 27

D EEG Diagnostic Environment: Tool Inventory **28**

D.1 Tool execution and data state . . . 29
D.2 Knowledge sources and literature coverage . . . 30
D.3 Collection, indexing, and retrieval . . . 30
D.4 Clinical Evidence Verification Loop . . . 31

E Diagnostic Inference Graph Extraction **31**

E.1 A003 node contents . . . 35
E.2 A003 final diagnosis and report . . . 49

F Experience Library and Update Details **51**

G Case Study **52**

H Representative Coding-Agent Rollouts **53**

H.1 Correctly classified negative case: OCD task, subject 01 . . . 53
H.2 Correctly classified control: schizophrenia task, subject 05 . . . 54
H.3 Misclassified negative case: OCD task, subject 02 . . . 55

A Extended Related Work

Clinical EEG interpretation and automated analysis. Clinical EEG interpretation combines signal characterization with judgments about diagnostic significance. The IFCN glossary defines terminology for EEG findings, while SCORE organizes reporting around recording conditions, background activity, artifacts, and diagnostic significance (Kane et al., 2017; Beniczky et al., 2013). These standards distinguish an observed pattern from its clinical interpretation, a distinction that matters when normal variants or artifacts resemble pathology (Gélisse et al., 2024; Benbadis & Lin, 2008). Automated analysis also predates language-model agents, including methods for recognizing interictal epileptic activity and seizures (Gotman & Gloor, 1976; Gotman, 1982).

Learning transferable EEG representations. EEG representation learning has progressed from task-specific architectures to pretrained backbones shared across datasets and tasks. EEGNet uses compact temporal and spatial convolutions, and EEG Conformer combines convolutional feature extraction with attention (Lawhern et al., 2018; Song et al., 2023). BENDR learns from unlabeled EEG through a contrastive objective, while LaBraM and EEGPT use masked pretraining to learn reusable signal representations (Kostas et al., 2021; Jiang et al., 2024; Wang et al., 2024). BIOT accommodates heterogeneous biosignal inputs through channel-wise tokenization, CBraMod models spatial and temporal dependencies with criss-cross attention, and CSBrain incorporates cross-scale spatiotemporal structure (Yang et al., 2023; Wang et al., 2025; Zhou et al., 2025). REVE addresses variation in electrode arrangements and recording length through position-aware pretraining across diverse datasets (El Ouahidi et al., 2025). These approaches learn representations that are usually adapted through a downstream classifier or fine-tuning.

Distribution alignment and clinical evaluation. Transfer across EEG recordings depends on both the learned representation and the evaluation conditions. Euclidean alignment transforms EEG trials to reduce inter-subject distribution differences, while convolutional Monge mapping normalizes spectral distributions for learning across sleep datasets (He & Wu, 2020; Gnassounou et al., 2023). EEG-Bench evaluates foundation models on clinical tasks under standardized protocols, and EEG-FM-Compass examines model design, adaptation strategies, and generalization across EEG datasets (Kastrati et al., 2025; Liu et al., 2026). Their comparisons show that the relative performance of pretrained models and specialist baselines varies across settings.

EEG-to-language models and tool-using agents. Language-based EEG systems connect signal analysis to textual reports and executable workflows. CELM generates clinical notes from EEG recordings, NeuroNarrator produces clinical narratives from grounded signal representations, and IED-RAG conditions report generation on retrieved EEG–text examples (Pradeepkumar et al., 2026; Wang et al., 2026b; Zhu et al., 2026). These systems make text generation part of the EEG prediction task, with retrieval providing case-specific context.

Tool-using agents expose the intermediate computations used to answer an EEG question. ReAct provides a general pattern for interleaving reasoning with actions and observations (Yao et al., 2023). EEG-GPT and EEG Agent apply language models to EEG analysis and interpretation, while BrainAgent delegates brain-signal workflows to specialized agents (Kim et al., 2025; Zhao et al., 2026; Zhou et al., 2026). NeuraDock separates deterministic numerical processing from a hardware-aware language interface, and CogEEGAgent uses registered analyses and controlled confirmation access to govern adaptive analysis (Xu et al., 2026; Hou et al., 2026).

Learning reusable guidance from experience. Experience-learning agents retain information that changes how they approach later interactions. Reflexion stores verbal feedback in episodic memory for subsequent trials, and Voyager accumulates executable skills for reuse in an embodied environment (Shinn et al., 2023; Wang et al., 2023). ExpeL extracts insights from successful and failed attempts and retrieves similar successful trajectories, while AutoGuide produces conditional guidelines from offline experience and selects them according to the current context (Zhao et al., 2024; Fu et al., 2024). These methods differ in whether they retain reflective text, executable programs, examples, or context-dependent instructions.

More recent methods emphasize extracting and transferring reusable guidance. ReasoningBank extracts structured strategies from successes and failures, with memory-aware test-time scaling

supplying contrastive trajectories for the same query (Ouyang et al., 2026). Trace2Skill consolidates trajectory-local lessons into portable standard operating procedures (Ni et al., 2026).

Structured and graph-based memory. Graph-based memory links experiences to support retrieval and revision. A-MEM links structured notes and updates their contextual representations as new memories arrive (Xu et al., 2025). Mem0 extracts and consolidates conversational information, with a graph variant representing relations among its elements (Chhikara et al., 2025).

GSEM combines clinical reasoning, outcome contrast, and structured experience (Han et al., 2026). It samples multiple trajectories for a training instance, pairs each failed trajectory with the most similar successful one, and extracts decision patterns. Its dual-layer graph represents decision structure within extracted experiences and relations between experiences, supporting applicability-aware retrieval and feedback-based calibration. In OMNIBRAIN, the graphs instead represent diagnostic attempts, linking hypotheses, measurements, and interpretations so that their differences provide evidence for formulating reusable guidance.

Optimizing policies and agent harnesses. Feedback can also update the policy or executable workflow that produces a trajectory. NeuroWeaver searches over EEG analysis pipelines, and EEG-SpikeAgent iteratively proposes signal features and refines them using detection metrics (Wang et al., 2026a; Santhosh et al., 2026). Meta-Harness searches executable harness code using earlier candidates, scores, and execution traces (Lee et al., 2026). WHALE alternates harness search with model-weight updates (Kim et al., 2026). SkillRL couples experience-based skill distillation with reinforcement learning, allowing its skill library and policy to evolve together (Xia et al., 2026). OMNIBRAIN updates an experience library whose interpretation and verification items guide subsequent analyses, with model weights fixed.

Table 2: Recording counts and numbers of classes for each task and dataset. $|y|$ denotes the number of decision classes.

Task	Dataset	Train	Test	$\|y\|$
Parkinson	Cavanagh2018a (Cavanagh et al., 2017)	53	24	2
	Cavanagh2018b (Cavanagh et al., 2017)	40	16	2
	Singh2018 (Singh et al., 2018)	39	16	2
	Brown2020 (Brown et al., 2020)	60	24	2
	Singh2021 (Singh et al., 2021)	78	34	2
	Singh2020 (transfer) (Singh et al., 2020)	0	39	2
Schizophrenia	Albrecht2019 (Albrecht et al., 2019)	63	13	2
	ASEEG (Bagherzadeh et al., 2026)	126	42	2
AD/FTD	ds004504 (Miltiadous et al., 2023)	64	12	3
OCD	Gruendler2009 (Gründler et al., 2009)	25	8	2
	Clinical antisaccade (Khayrullina et al., 2024a;b)	38	13	2
mTBI	Cavanagh2019 (Cavanagh et al., 2019b)	149	28	2
	ds005114 (external) (Cavanagh, 2024)	0	50	2
Depression	Mumtaz2016 (Mumtaz, 2016)	85	18	2
	ds003478 (Cavanagh et al., 2019a)	152	47	2

B Datasets and Splits

B.1 Datasets, tasks, and acquisition

Table 2 summarizes the recording counts and numbers of classes. Where indicated, the original benchmark protocol of Wang et al. (2025) is followed. The five Parkinson training sources contribute 270 training recordings and 114 test recordings from those sources. Singh2020 contributes a separate set of 39 recordings for cross-source evaluation.

OCD task labels. Gruendler2009 recruited unselected participants, with labels based on the Obsessive-Compulsive Inventory (OCI). In the exported split the positive class is the high-symptom OCI tertile ($n = 12$, OCI 22–42) and the negative class comprises the low and middle tertiles ($n = 21$, OCI 0–17). This setting distinguishes high from low or moderate obsessive-compulsive symptom load. The clinical antisaccade setting instead distinguishes diagnosed OCD from healthy controls. Labels, splits, and evaluations are kept separate; the OCD task summary averages the two settings equally (Section B.4).

Acquisition paradigms. The sources include resting-state and task-based recordings. Cavanagh2018b, ds004504, Mumtaz2016, ASEEG, and ds003478 are resting-state; ds004504 is eyes-closed, Mumtaz2016 and ASEEG provide separate eyes-closed and eyes-open recordings, and ds003478 alternates these conditions within each resting run. Behavioral tasks in Table 2 include an oddball task for Cavanagh2018a and Cavanagh2019, cognitive control for Singh2018 and Albrecht2019, a DPX cognitive-control task for ds005114, a flanker task for Gruendler2009, an antisaccade task for the clinical OCD dataset, and a cued-movement task for Brown2020. Every method receives the recording as a continuous signal with no event markers, stimulus codes, or trial segmentation.

B.2 Subjects and split construction

Training and test sets are separated by subject. Table 3 distinguishes subject counts from recording counts for sources with repeated recordings. The remaining datasets in Table 2 contribute one recording per subject. Cavanagh2018a and Cavanagh2018b share 53 participants, with all shared participants assigned consistently to training (37) or test (16) across the two datasets. Their recordings represent different acquisition paradigms.

Table 3: Subject counts for datasets with repeated recordings. Validation denotes a separate explicit split.

Dataset	Recordings per subject	Train	Validation	Test
Cavanagh2018a	Up to two sessions	37	—	16
Brown2020	Up to two sessions	40	—	16
Cavanagh2019	Up to three sessions	58	—	13
Mumtaz2016	Eyes-closed / eyes-open	45	9	9
ASEEG	Eyes-closed / eyes-open	64	16	21
ds003478	Up to two resting runs	76	20	24
ds005114 (external)	Up to three sessions	0	0	20

Table 4: Recording counts used for baseline fitting and validation. "Reserved" validation is drawn from the training pool in Table 2; "explicit" validation is separate from its Train column.

Dataset	Validation split	Fitting	Validation
Cavanagh2018a	Reserved	42	11
Cavanagh2018b	Reserved	32	8
Singh2018	Reserved	31	8
Brown2020	Reserved	48	12
Singh2021	Reserved	62	16
Albrecht2019	Reserved	50	13
Gruendler2009	Reserved	20	5
Cavanagh2019	Reserved	119	30
ds004504	Explicit	64	12
Mumtaz2016	Explicit	85	16
ASEEG	Explicit	126	30
Clinical antisaccade	Explicit	38	10
ds003478	Explicit	152	40
ds005114 (external)	None (test only)	0	0

Internal validation. For the main comparison and the end-to-end comparison in Table 17, datasets without an explicit validation split reserve 20% of the training recordings using class-stratified sampling with seed 42, with the count rounded to the nearest integer. The ds004504 and Mumtaz2016 datasets instead use explicit validation splits; ASEEG, clinical OCD, and ds003478 use the subject-level splits in Sections B.3 to B.5. The reserved validation split is at the recording level, so fitting and validation may include different recordings from the same training subjects. Table 4 gives the resulting recording counts. The pooled transfer experiment uses the separate split described in Section C.6; ds005114 is used only for external testing.

B.3 Additional schizophrenia data

ASEEG (Bagherzadeh et al., 2026) adds 101 participants with 198 eyes-open and eyes-closed resting recordings. Class-stratified subject-level sampling reserves 20% for testing and 20% of the remaining subjects for validation, using seed 42 in both steps. All recordings from a participant remain in one split. The five foundation models use the same 50-epoch frozen-backbone protocol and recording-level validation balanced accuracy as the other binary datasets.

The public ASEEG release contains 128 Hz signals already filtered to 1–45 Hz. We resample these signals to 200 Hz, retain the 19 scalp electrodes, and use the same non-overlapping 10 s windows as the other datasets.

B.4 Clinical OCD classification

The clinical antisaccade dataset comprises 31 participants diagnosed with OCD and 30 healthy controls (Khayrullina et al., 2024a;b). Each participant contributes one continuous task recording. Class-stratified subject-level sampling reserves 20% for testing and 20% of the remaining participants for validation, using seed 42 in both steps. This yields 38/10/13 participants for training, validation,

and testing; the test set contains six controls and seven participants with OCD. The five foundation models use the same 50-epoch frozen-backbone protocol and recording-level validation BA as the other binary datasets.

Signals were acquired at 500 Hz with an FCz reference and a 0.05–70 Hz acquisition bandpass. We retain the 19 standard scalp electrodes and the continuous interval from the first complete antisaccade trial to the end of the final trial's one-second blank screen. This removes pre- and post-task activity while preserving internal task breaks. Event markers define the interval but are not supplied to the models or agent. After resampling to 200 Hz, inputs use the common filtering and non-overlapping 10 s windows described in Appendix C.

B.5 Depressive-symptom classification

The public resting-state release ds003478 provides EEG from participants recruited for consistently high or low Beck Depression Inventory (BDI) scores (Cavanagh et al., 2019a). We classify BDI scores ≥ 13 versus < 7, retaining 120 participants (46 high and 74 low) with 239 recordings. One participant with BDI 7 and the participant marked invalid by the release are excluded. Each resting run remains one recording, including its alternating eyes-open and eyes-closed intervals. These symptom labels form a separate classification setting from the clinical labels in Mumtaz2016.

Class-stratified subject-level sampling reserves 20% for testing and 20% of the remaining participants for validation, with seed 42 for both steps. All runs from a participant remain in one split, yielding 76/20/24 participants and 152/40/47 recordings for fitting, validation, and testing. The test split contains 29 low-symptom and 18 high-symptom recordings. Together with the 18 Mumtaz2016 test recordings, the two settings cover 65 test recordings. We retain 19 standard scalp electrodes from the 500 Hz release and apply the common 200 Hz, 10-second-window preprocessing with a 60 Hz notch. The public release includes source-side channel interpolation; its released signals are used as provided before this preprocessing.

B.6 External mTBI data

The DPX cognitive-control dataset ds005114 (Cavanagh, 2024) supplies an external mTBI test set. We exclude every original participant identifier used in Cavanagh2019 training, validation, or testing. The remaining 20 participants (13 mTBI and 7 controls) contribute 50 recordings across up to three sessions: 33 mTBI and 17 control recordings. These are additional participants from the same research programme, evaluated under a different task. Together with the 28 Cavanagh2019 test recordings, mTBI evaluation covers 78 recordings across the within-source and external settings.

The DPX recordings are used only for external testing, with no target fitting or validation. We retain the 19 standard scalp electrodes, resample from 500 to 200 Hz, apply a 60 Hz notch with the common bandpass, and form 10-second non-overlapping windows. The export contains 10,703 windows. External inference reuses the five Cavanagh2019 frozen-backbone checkpoints selected on their source validation data; classifier parameters also remain fixed. Results are reported separately from the within-source mTBI row of Table 1.

B.7 Data use across experiments

Table 5 identifies the learning and evaluation data for the main comparison and transfer experiments. Experience libraries are frozen before evaluation. Singh2020 is excluded from foundation-model training, validation, and experience construction.

B.8 Preprocessing and data representations

We resample recordings to 200 Hz, apply a 0.3–75 Hz bandpass filter, and standardize the channel set per dataset. Source acquisition bandwidths for ASEEG and clinical OCD are specified in Sections B.3 and B.4. The notch frequency is 60 Hz for Gruendler2009, ds003478, and ds005114, and 50 Hz for the remaining sources. All five foundation models use the raw-microvolt export with the model-specific normalization and channel mapping in Table 6. BENDR's input adapter subsequently interpolates each window to 256 Hz.

Table 5: Data scope for the main and transfer experiments.

Experiment	Learning data	Evaluation data
Main comparison	Task training recordings for experience learning; source-specific fitting/validation splits for FMs	Corresponding task test sets, including 114 Parkinson recordings
Transfer to Singh2020	Parkinson training pool for agent experience; each source separately for FMs	39 Singh2020 recordings
Pooled FM transfer	Union of the five source training pools (243 fitting / 27 validation recordings) for FMs	The same 39 Singh2020 recordings
Experience transfer	Singh2021 training recordings	34 Singh2021 test recordings and the five other Parkinson test sets
External mTBI transfer	Cavanagh2019 source checkpoints, fixed before transfer	50 DPX recordings from 20 previously unused participants
CEVL / extraction ablations	Parkinson and schizophrenia training recordings	The same 114 Parkinson and 13 schizophrenia test recordings

Table 6: Default per-model input shaping and normalization. "Channels" is the effective channel count after padding or cropping. Task-specific inputs are described below.

Model	Input source	Channels	Normalization	Pretrained weights
CBraMod	raw μV	19 (real)	μV $\div 100$	Official 12-layer checkpoint
CSBrain	raw μV	19 (real)	μV $\div 100$	Official checkpoint, region-aware
LaBraM	raw μV	19 (real)	μV $\div 100$	`labram-base.pth` (student)
BIOT	raw μV	19 (real)	95th-percentile per channel	18-channel six-dataset checkpoint
BENDR	raw μV	19 $\rightarrow$ 20	joint EEG min–max + zero scale	Encoder + conditioning weights

C Experimental Protocols and Supplementary Results

C.1 Foundation-model inputs and architectures

We compare against five representative EEG foundation models. Each model uses its own input pipeline. Table 6 summarizes the default per-model input shaping.

CBraMod (Wang et al., 2025) is a 12-layer criss-cross transformer (d=200, d_{ff}=800, 8 heads) with the official all-patch-reps classification head. Input is raw microvolts divided by 100, following the original pretraining pipeline's physical-amplitude scaling. The 19 standard 10–20 electrodes are used directly.

CSBrain (Zhou et al., 2025) shares CBraMod's patch embedding but replaces its criss-cross attention with cross-scale spatiotemporal tokenization (CST), per-region spatial embedding across five brain regions (frontal, parietal, temporal, occipital, central), and structured sparse attention (SSA) that alternates intra-window temporal attention with cross-region spatial attention under a learned sparse mask. Each of the 19 standard 10–20 electrodes is mapped to its anatomical region via a fixed lookup table. Input and normalization match CBraMod (raw μV $\div 100$).

LaBraM-Base (Jiang et al., 2024) is a ViT-style masked-prediction transformer with relative position bias. It consumes raw microvolts divided by 100, excludes PAD channels, and maps each real electrode to its standard spatial embedding index. Its native downstream path averages patch tokens, applies `fc_norm`, and uses a linear classifier. The normalization layer belongs to the backbone and remains fixed in the frozen-backbone condition.

BIOT (Yang et al., 2023) tokenizes each channel independently via STFT (n_{fft}=200, hop 100) and processes the resulting spectrogram tokens with a transformer (d=256, 8 heads, depth 4). It is loaded from the 18-channel six-dataset checkpoint pretrained on MGH, SHHS, TUAB, TUEV, CHB-MIT, and IIIC Seizure. Only the 19 real EEG channels are passed to BIOT (padding channels excluded);

Table 7: Default foundation-model training configuration. The LaBraM and BIOT settings described above override the corresponding defaults.

Setting	Value
Optimizer	AdamW, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$
Learning rate / decay	10^{-4} / 0.05
Gradient clipping	1.0
Schedule	Minibatch-level cosine annealing, $\eta_{\min} = 10^{-6}$, $T_{\max}$ = epochs × |train_loader|
Loss	Cross-entropy, inverse-frequency class weights, label smoothing 0.1
Epochs	50
Batch size / seed	64 / 42
Checkpoint selection	Validation Cohen's κ for AD/FTD, validation BA for binary tasks
Validation unit	Recording for frozen-backbone training; window for end-to-end fine-tuning
Main-table updates	Classification head only (backbone frozen)

the first 18 channel-token embeddings come from the pretrained checkpoint and the 19th is randomly initialised. The 19th embedding is updated only in the end-to-end condition (Section C.9); under the frozen-backbone condition of Table 1, the entire backbone including all channel-token embeddings stays fixed. Per-channel normalization divides by the 95th-percentile absolute amplitude.

BENDR (Kostas et al., 2021) follows the `LinearHeadBENDR` downstream architecture. Its input adapter maps real electrodes by name to the 19-channel scalp layout, applies joint EEG min–max normalization within each window, and appends a fixed zero scale channel because no dataset amplitude range is supplied. Normalization includes zero-valued missing channel slots, and PAD channels are excluded. Nearest-neighbor interpolation converts the 200 Hz windows to 256 Hz before encoding. The convolutional encoder loads `encoder.pt`, and `EncodingAugment` loads the corresponding conditioning weights from `contextualizer.pt`, without using the full transformer context network. Conditioning parameters remain frozen in both protocols, while gradients pass through them to the encoder during end-to-end training. The conditioned sequence is pooled into four temporal bins before classification.

C.2 Foundation-model training and evaluation

Table 7 summarizes the training configuration. LaBraM uses a base learning rate of 5×10^{-4} in both training conditions; end-to-end fine-tuning additionally applies layer-wise learning-rate decay of 0.65. LaBraM and BIOT use a 10% linear warmup followed by cosine decay with a minimum learning-rate multiplier of 10^{-3}, and gradient clipping 5.0. BIOT retains a base learning rate of 10^{-4} without layer-wise decay. In the frozen-backbone condition, all backbone parameters are fixed and the backbone is evaluated in inference mode while the classification head is trained.

Classification heads. CBraMod and CSBrain use a three-layer MLP over flattened per-window patch features (input dimension $C\times10\times200$). LaBraM uses its native linear classifier over the normalized 200-dimensional pooled feature. BIOT uses an ELU followed by a linear classifier. BENDR flattens its four pooled bins to 6,144 features, projects them to 2,048 dimensions with dropout 0.4, ReLU, and batch normalization, and applies a final linear classifier.

Per-recording inference. Models use non-overlapping 10 s windows. The model produces per-window class probabilities, which are mean-pooled across all windows of a recording. The argmax of the pooled distribution is the recording prediction. MV combines one recording-level prediction from each of the five main-table models using the same checkpoints and test recordings. Vote ties are resolved by the summed recording-level probabilities among the tied classes, with the lowest class index resolving an exact probability tie. Four AD/FTD recordings require vote tie-breaking in the main comparison.

Hardware and reproducibility. Foundation-model training uses a single NVIDIA RTX 5090 D (32 GB) with PyTorch 2.11.0+cu128 and Python 3.10. The frozen-backbone, end-to-end, and pooled training runs in Tables 1, 13 and 17 use seed 42 for weight initialisation, data shuffling, and

Table 8: Agent configuration during experience learning and test inference. LLM weights remain fixed in both phases.

Setting	Experience learning	Test inference
Case data	Training recordings	Test recordings
Attempts per recording	Four	One per evaluation run
Tool-call limit	100 per attempt	100 per attempt
CEVL cycle limit	20 per attempt	20 per attempt
Experience library	One update after the four attempts	Frozen final library
FM predictions	Disabled	Consulted after signal analysis

Table 9: Dataset-level schizophrenia results (%). F1 is support-weighted. The main table averages the two datasets equally.

Dataset	Metric	CBraMod	LaBraM	BIOT	CSBrain	BENDR	MV	OMNIBRAIN
Albrecht2019	BA	54.76	63.10	46.43	41.67	58.33	63.10	**78.57**
	κ	9.30	25.29	−7.06	−15.56	17.72	25.29	**55.17**
	F1	53.30	60.14	46.15	25.64	52.86	60.14	**76.08**
ASEEG	BA	**78.41**	68.41	56.36	68.86	49.09	**78.41**	73.86
	κ	**56.95**	37.24	12.90	37.81	−1.76	**56.95**	47.62
	F1	**78.53**	68.42	55.93	68.99	42.19	**78.53**	73.82

Table 10: OCD classification by type (%). Symptom grouping uses Gruendler2009; clinical classification uses the antisaccade dataset. F1 is support-weighted. The main table averages the two settings equally.

Type	Metric	CBraMod	LaBraM	BIOT	CSBrain	BENDR	MV	OMNIBRAIN
Symptom grouping	BA	56.67	56.67	53.33	66.67	33.33	56.67	**83.33**
	κ	14.29	14.29	5.88	38.46	−26.32	14.29	**71.43**
	F1	60.45	60.45	50.00	70.83	15.00	60.45	**86.82**
Clinical classification	BA	53.57	**70.24**	46.43	**70.24**	63.10	**70.24**	67.86
	κ	7.14	**39.53**	−7.06	**39.53**	25.29	**39.53**	36.59
	F1	53.85	**68.86**	46.15	**68.86**	60.14	**68.86**	68.08

validation sampling. Each model–dataset–protocol configuration has one training run. Parkinson scores are unweighted means across the five source datasets. The scope and per-run evidence of the accompanying release are detailed in Section C.10.

Metrics. All F1 scores use support-weighted averaging.

C.3 AGENT LEARNING AND INFERENCE

OMNIBRAIN uses GPT-5.6-Sol at temperature 0.3, with one experience library per task. The prompt template and Diagnostic Environment remain fixed within each task and phase. Table 8 distinguishes experience learning from test inference. Agent scores are averaged over three evaluation runs per condition.

C.4 DATASET-LEVEL CLASSIFICATION RESULTS

The ASEEG scores in Table 9 use 42 test recordings from 21 participants; window probabilities are averaged within each recording. The schizophrenia task score in Table 1 is the unweighted mean of Albrecht2019 and ASEEG.

Depression settings. Table 11 reports clinical and symptom-grouping results separately. The depression row of Table 1 averages the two settings equally for each metric.

Table 11: Depression classification by setting (%). Mumtaz2016 uses clinical labels; ds003478 uses high versus low BDI symptoms. F1 is support-weighted.

Dataset	Metric	CBraMod	LaBraM	BIOT	CSBrain	BENDR	MV	OMNIBRAIN
Mumtaz2016	BA	93.75	**100.00**	**100.00**	87.50	70.00	**100.00**	82.50
	κ	88.61	**100.00**	**100.00**	76.92	41.56	**100.00**	65.82
	F1	94.39	**100.00**	**100.00**	88.60	70.83	**100.00**	83.17
ds003478 (BDI)	BA	53.54	59.10	66.38	60.44	49.71	61.49	**81.32**
	κ	7.65	19.21	34.20	23.62	−0.64	25.40	**60.72**
	F1	56.87	62.35	69.31	63.61	52.39	64.91	**81.06**

Table 12: External mTBI performance on ds005114 (%) using fixed Cavanagh2019 checkpoints. F1 is support-weighted.

Metric	CBraMod	LaBraM	BIOT	CSBrain	BENDR	MV	OMNIBRAIN
BA	49.02	50.09	42.87	59.36	43.23	48.84	**64.17**
κ	−1.60	0.12	−10.74	14.89	−11.71	−1.76	**26.11**
F1	44.00	24.33	29.66	51.31	43.29	36.79	**64.90**

C.5 EXTERNAL MTBI TRANSFER RESULTS

Table 12 evaluates the source-trained frozen-backbone models on the 50 DPX recordings in Section B.6. Each model uses its Cavanagh2019 checkpoint without updating the backbone or classification head. Window probabilities are averaged within each recording, and MV combines the five resulting binary predictions.

C.6 FOUNDATION-MODEL TRANSFER TO SINGH2020

The cross-dataset transfer experiment (Figure 4b) evaluates the 25 Parkinson source-dataset checkpoints from the end-to-end condition on Singh2020, using five checkpoints per model. Each checkpoint is selected on its source validation set before transfer, and each model's score is the mean across its five source checkpoints. Singh2020's 39 recordings yield 1,179 windows and are used only for testing. Inputs follow the model-specific pipelines above, and recording predictions use mean window probabilities as in Appendix C. The in-distribution column in Table 13 reports the unweighted mean BA across the five source-dataset test sets.

The pooled column in Table 13 instead fine-tunes one checkpoint per model on the union of the five source training pools (270 recordings). Within each source, sorted recordings are shuffled with `random.Random(42)` and the first 10% are reserved for validation, yielding 243 fitting, 27 validation, and 39 test recordings; validation recordings may share participants with fitting. Checkpoints are selected by window-level validation BA, and test inference is unchanged. CBraMod and CSBrain consume the raw microvolt signal divided once by 100 with padding channels removed, and CSBrain uses its historical 19-channel region layout; LaBraM, BIOT, and BENDR keep the input policies of Appendix C. All pooled runs use seed 42 and 50 epochs. Frozen-backbone pooling and pooled end-to-end fine-tuning are evaluated separately in Tables 13 and 21.

C.7 ABLATION PROTOCOLS

The joint ablation compares CEVL and experience use on the Parkinson's disease and schizophrenia (Albrecht2019) test splits in Table 5. PD balanced accuracy is the unweighted mean of five per-dataset balanced accuracies, matching the metric in Table 1. The two conditions with experience share the same frozen library constructed with CEVL on training recordings. The CEVL switch changes the inference procedure, while the experience switch controls access to this library.

The control without CEVL allows freely organized analysis while retaining the same signal-analysis capabilities. It omits compulsory expectations before analysis, structured interpretations after each analysis, and the pending-hypothesis queue. The LLM, temperature, inputs, knowledge resources,

Table 13: Parkinson's disease classification performance on source datasets and Singh2020 (BA, %). The first Singh2020 column averages five source-dataset checkpoints; the pooled column fine-tunes one checkpoint per model on the union of the five source training sets.

Method	In-distribution	Singh2020	Singh2020 (pooled)
CBraMod	65.3	59.6	78.8
LaBraM	70.7	58.1	69.2
BIOT	55.2	59.2	82.7
CSBrain	69.1	56.2	75.0
BENDR	57.0	72.7	80.8
Best FM	70.7	72.7	82.7
OMNIBRAIN	**74.5**	**78.85**	78.85

Table 14: Cross-dataset transfer of diagnostic experience from Singh2021 (BA, %).

Dataset (test recordings)	No experience (%)	Source experience (%)	ΔBA (%)
Singh2021 (held-out source, n = 34)	57.92	66.25	+8.33
Cavanagh2018a (n = 24)	65.62	68.75	+3.13
Cavanagh2018b (n = 16)	62.50	75.00	+12.50
Singh2018 (n = 16)	68.75	81.25	+12.50
Brown2020 (n = 24)	71.88	81.25	+9.37
Singh2020 (n = 39)	53.85	76.92	+23.07
Mean over five target datasets	64.52	76.63	+12.11

Table 15: Joint CEVL and experience ablation. ✓ and ✗ indicate enabled and disabled components at inference.

Setting	CEVL	Experience	PD BA (%)	Schizophrenia BA (%)
Tool-using agent	✗	✗	52.08	21.43
CEVL only	✓	✗	65.33	61.90
Experience only	✗	✓	59.17	46.43
OMNIBRAIN	✓	✓	74.50	78.57

Table 16: Experience-extraction ablation with CEVL fixed.

Extraction	PD BA (%)	Schizophrenia BA (%)
None (CEVL only)	65.33	61.90
Direct	62.08	54.76
DIG-based	74.50	78.57

tool access, and token and tool-call limits remain fixed, with prompt and planning constraints changed only to enable or remove CEVL. Evaluation uses recording-level BA.

The extraction ablation holds source records fixed to isolate extraction from changes in subsequent training rollouts. Direct and DIG-based extraction share the same training recordings, pairs of successful and failed attempts, LLM, consolidation procedure, and skill schema. The resulting libraries are matched in entry count and token budget and evaluated with CEVL on the same held-out recordings. The CEVL-only condition from the joint ablation supplies the no-experience reference.

C.8 Experience transfer and ablation results

C.9 Full Fine-Tuning Comparison

Table 17 reports end-to-end fine-tuning with seed 42 for 50 epochs using the corresponding dataset splits and model-specific input pipelines from the frozen-backbone comparison. Backbone and classification-head parameters are updated, except for BENDR's fixed conditioning parameters described above. Checkpoint selection uses window-level validation, while test metrics use recording-

Table 17: End-to-end fine-tuning classification performance (BA, %). Bold marks the best model per row.

Target (test split)	CBraMod	LaBraM	BIOT	CSBrain	BENDR
Parkinson (five-dataset mean)	65.3	**70.7**	55.2	69.1	57.0
Schizophrenia (Albrecht2019)	32.1	52.4	38.1	**57.1**	54.8
AD/FTD (ds004504, 3-class)	28.3	30.0	21.7	**50.6**	38.9
OCD (Gruendler2009)	66.7	**83.3**	56.7	66.7	66.7
mTBI (Cavanagh2019)	62.3	54.1	75.4	67.8	**81.6**
Depression (Mumtaz2016)	93.8	76.3	**100.0**	87.5	87.5
Average	58.1	61.1	57.8	**66.5**	64.4

Table 18: Pairwise agreement across five Parkinson evaluation datasets (N=114 recordings). Prediction κ compares predicted labels; correctness κ compares binary correctness indicators.

Model pair	Agreement (%)	Prediction κ	Correctness κ
BENDR vs. LaBraM	73.7	0.391	0.390
BENDR vs. CBraMod	68.4	0.323	0.299
LaBraM vs. CBraMod	73.7	0.439	0.364
BENDR vs. BIOT	64.0	0.211	0.240
BENDR vs. CSBrain	69.3	0.390	0.323
LaBraM vs. BIOT	72.8	0.409	0.399
LaBraM vs. CSBrain	65.8	0.320	0.182
CBraMod vs. BIOT	74.6	0.470	0.454
CBraMod vs. CSBrain	69.3	0.388	0.305
BIOT vs. CSBrain	64.9	0.301	0.250
Mean	69.6	0.364	0.320

level predictions. The average is the unweighted mean BA across the six rows shown, with Albrecht2019 representing schizophrenia and Gruendler2009 representing OCD.

C.10 Released run evidence

The accompanying results release covering the five Parkinson sources and the Albrecht2019, ds004504, Gruendler2009, Cavanagh2019, and Mumtaz2016 datasets, under `result/unified_seed42/`, includes metrics and recording-level predictions for 50 frozen-backbone runs, 50 end-to-end runs, five pooled runs, and 25 single-source transfer evaluations. `CURRENT.json` identifies the result version, and the release ledger links each reported baseline to its constituent runs and implementation. Each training configuration has one seed-42 execution, and the released records support recomputation of BA, Cohen's κ, and support-weighted F1.

C.11 Pairwise Disagreement on Parkinson Evaluation Sets

Table 18 reports pairwise prediction agreement and Cohen's κ for five end-to-end fine-tuned EEG foundation models over 114 test recordings from Brown2020, Cavanagh2018a, Cavanagh2018b, Singh2018, and Singh2021. Predictions are aligned by dataset and recording identifier. Correctness κ measures agreement on which recordings each model classifies correctly, complementing the performance comparison in Section 3.

C.12 Supplementary Results for Motivating Observations

Tables 19 to 21 give the numerical results shown in Figure 2. The dataset comparison uses end-to-end fine-tuned models. Majority vote combines their recording-level predictions; Best FM is the highest model BA within each dataset, and Mean averages each column across datasets. The separate training-source comparison freezes the CBraMod backbone and trains heads on individual sources or

Table 19: Classification counts underlying Figure 2a. TN and TP are correct negative and positive predictions; FP and FN are false positive and false negative predictions.

Task	TN	FP	FN	TP	Total
OCD	1	4	0	3	8
Schizophrenia	1	5	1	6	13
Total	2	9	1	9	21

Table 20: Dataset-level balanced accuracy (%) underlying Figure 2b.

Dataset	CBraMod	LaBraM	BIOT	CSBrain	BENDR	Majority vote	Best FM
Cavanagh2018a	59.4	56.3	34.4	50.0	46.9	50.0	59.4
Cavanagh2018b	56.3	56.3	43.8	75.0	43.8	56.3	75.0
Singh2018	68.8	87.5	62.5	81.3	68.8	81.3	87.5
Brown2020	71.9	75.0	65.6	75.0	68.8	78.1	75.0
Singh2021	70.4	78.3	69.6	64.2	56.7	74.6	78.3
Mean	65.3	70.7	55.2	69.1	57.0	68.0	75.0

Table 21: CBraMod training-source comparison on Singh2020 (BA, %).

Training source	Training recordings	Singh2020 BA (%)
Singh2021	70	75.0
Singh2018	35	71.2
Cavanagh2018b	36	57.7
Brown2020	54	51.9
Cavanagh2018a	48	50.0
All five pooled	243	75.0

their pooled data, evaluating each on the same 39 Singh2020 recordings. Its fitting-set sizes are given in Table 21.

D EEG Diagnostic Environment: Tool Inventory

The EEG Diagnostic Environment exposes five tool groups to the agent. Table 22 lists the concrete tools in each group with their functions. The only component updated across cases is the experience library $\mathcal{S}$, which shapes *when* and *how* the agent uses these tools. The case input x in Algorithm 2 includes the EEG recording and recording metadata describing acquisition and preprocessing. The task prompt specifies the candidate classes, while the prompt's "Clinical context" explains how those classes are interpreted. The recording's target label is used as outcome feedback after the training attempts and graphs are complete, as specified in Algorithm 1.

Targeted tool access supports analyses of selected channels, intervals, and frequency bands without placing the full recording in the language model's context. Retrieval of distributed evidence becomes less reliable as language-model context grows (Liu et al., 2024; Hsieh et al., 2024). Planning tools record hypotheses and interpretations, signal-analysis tools execute generated Python and render waveforms, and preprocessing tools provide filtering and artifact handling. Knowledge retrieval covers research literature, clinical references, disease-specific guidelines, and EEG technical standards. For Parkinson's disease, these resources include the MDS diagnostic criteria (Postuma et al., 2015) and a review of quantitative EEG findings (Shirahige et al., 2020). At test inference, foundation-model predictions from CBraMod, LaBraM, BIOT, and CSBrain become available after the agent completes its own analysis.

For every tool execution, the Environment records the code or routine, software versions, parameters, selected channels and time intervals, input dependencies, and produced outputs. Generated Python pro-

Table 22: Tool inventory of the EEG Diagnostic Environment.

Group	Tool	Function
Planning	`create_plan`	Create a hypothesis-driven analysis plan
	`reason_step`	Declare the hypothesis under test before executing code
	`interpret`	Compare measurement result against the stated expectation
	`revise_plan`	Modify the hypothesis plan (add, skip, or reorder steps)
Signal analysis	`run_python`	Execute Python code in a persistent analysis namespace (NumPy, SciPy, MNE, Matplotlib)
	`show_eeg`	Render EEG waveforms for visual morphology inspection (spike shape, burst patterns, artifacts)
Preprocessing	`bandpass_filter`	Apply bandpass filter
	`run_ica`	ICA artifact removal
	`detrend`	Baseline drift removal (linear detrend + zero-phase highpass)
	`detect_bad_channels`	Detect channels with abnormal amplitude
	`remove_channels`	Remove specified channels from the recording
Foundation model	`get_foundation_model_predictions`	Independent predictions from pretrained EEG models (CBraMod, LaBraM, BIOT, CSBrain); enabled at test inference after signal analysis, disabled during experience learning
Knowledge retrieval	`search_papers`	Primary research literature (qEEG biomarkers, cutoff values, normative ranges)
	`search_references`	Clinical reference material (differential diagnosis, contextual interpretation)
	`search_disease_guidelines`	Disease-specific diagnostic criteria and assessment or management guidance (PD, schizophrenia, OCD, mTBI, depression, AD)
	`search_eeg_standards`	ACNS/IFCN technical standards (EEG terminology, measurement definitions, reporting conventions)

grams create derived artifacts without overwriting the source recording. Verification items in $\mathcal{S}$ guide the choice of follow-up analyses, whose observations are stored with the same provenance metadata.

D.1 Tool execution and data state

The clinical environment registers the 16 tools in Table 22. The Python interface provides the current EEG array in microvolts, sampling rate, channel names, recording condition, and analysis interval together with NumPy, SciPy, MNE, and Matplotlib. Variables persist between calls, allowing later analyses to reuse derived measurements. Waveform inspection through `show_eeg` accepts a time interval and optional channel selection and returns a rendered image. Spectral estimates, band-power calculations, and connectivity measures are implemented through Python library calls.

Preprocessing acts on the current analysis state. Bandpass filtering uses MNE's FIR implementation. Detrending applies linear detrending in 10-second segments followed by a zero-phase FIR highpass, with a default cutoff of 1 Hz. The ICA interface requests 20 components by default, capped at the number of channels minus one, and uses FastICA with seed 42. Component exclusion uses EOG and muscle checks together with kurtosis and variance heuristics. Bad-channel detection flags channel standard deviations above five times or below one fifth of the across-channel median; channel removal is a separate operation. The agent selects these operations for the current analysis.

Foundation-model predictions are read from each recording's precomputed results and returned separately for each available model, preserving its label and probabilities. When planning is enabled, access requires an idle analysis cycle and a completed plan or exhausted analysis budget. The experience interface can additionally expose `list_skills`, `load_skill`, and `read_skill_resource`; these load experience content and are separate from the 16 environment tools.

Table 23: Representative clinical sources in the knowledge resources.

Task	Representative source	Reference role
Parkinson	MDS clinical diagnostic criteria (Postuma et al., 2015)	Diagnostic criteria
Schizophrenia	NICE CG178 (2014)	Recognition, assessment, and management
AD/FTD	NICE NG97 (2018); revised bvFTD criteria (Rascovsky et al., 2011)	Dementia assessment and management; bvFTD diagnostic criteria
OCD	NICE CG31 (2005)	Assessment and treatment context
mTBI	ACRM diagnostic criteria (Silverberg et al., 2023)	Diagnostic criteria
Depression	NICE NG222 (2022)	Assessment, treatment, and management

Table 24: Knowledge-resource inventory for the six clinical task domains on 25 September 2026. Counts distinguish source entries from indexed passages.

Domain	Research	Review	Guideline	Textbook	Entries	Passages
Parkinson	33	4	6	3	46	3,361
Schizophrenia	41	4	7	2	54	4,736
AD/dementia	48	3	7	9	67	6,680
OCD	13	2	6	1	22	1,438
mTBI	22	1	6	2	31	1,812
Depression	48	6	5	10	69	5,562
Total	205	20	37	27	289	23,589

D.2 Knowledge sources and literature coverage

The knowledge resources provide three kinds of context for EEG reasoning. Research articles and reviews describe signal findings and their measurement conditions, including spectral power, slowing, posterior alpha peaks, asymmetry, connectivity, aperiodic activity, microstates, and event-related potentials. Clinical references provide disease definitions, differential-diagnostic context, and assessment or management guidance. EEG standards supply terminology and reporting conventions. Table 23 gives representative clinical sources for the six tasks and the roles for which they are indexed.

The resource inventory dated 25 September 2026 contains 289 domain-specific source entries and 23,589 indexed passages for these six clinical domains (Table 24). Counts refer to entries within a domain, so a publication indexed in multiple domains contributes more than once. Source types follow the index metadata; the guideline category also includes technical-standard summaries.

Three local summaries separately support EEG-standard lookup: the ACNS critical-care terminology (Hirsch et al., 2021), ACNS reporting guidance (American Clinical Neurophysiology Society, 2006), and the IFCN glossary and report-format proposal (Kane et al., 2017). Their 49 lookup sections cover background frequency, PDR, continuity, reactivity, voltage, waveform morphology, and report organization. Disease resources include accessible article text and reference cards combining bibliographic records with locally authored paraphrases, as used for the MDS, ACRM, and NICE sources.

D.3 Collection, indexing, and retrieval

The collection pipeline uses task-specific queries to Europe PMC with open-access and full-text availability filters, supplemented by guideline searches, NCBI Bookshelf material, and manually imported official references. Configured title and abstract filters screen for the relevant modality and topic, and PMC identifiers support deduplication during discovery. JATS processing extracts titles, abstracts, and body text while removing reference lists and back matter. PDF text is extracted page by page with PyMuPDF. The current text-processing configuration uses 800-character windows

with 200-character overlap, retaining trimmed passages longer than 50 characters and filtering reference-list-like passages.

The domain collections use Chroma with cosine similarity; the persisted embeddings have 3,072 dimensions. Passage metadata carry the source, locator, and available bibliographic and source-type fields. PDF locators refer to extracted pages, while TXT locators identify a logical document. The inventory counts in Table 24 are measured from the persisted indexes.

The three vector-retrieval tools use distinct source filters. `search_papers` retrieves research articles and reviews using the target, population, comparator, metric, and question. `search_references` retrieves textbook material and professional or supplemental disease guidance. `search_disease_guidelines` selects disease-scoped entries assigned the core-guidance tier. Each tool returns at most five passages, with at most one passage per source, from an expanded candidate pool. Source filters remain in force when no matching passage is available. In contrast, `search_eeg_standards` searches the local Markdown summaries by keyword matching and returns at most two positively scored sections.

Retrieved passages carry retrieval, source, and hit identifiers that can be attached to a CEVL target as reference context. This context supplies measurement definitions and applicability conditions; observations about the current case come from analysis of its EEG. Retrieval similarity ranks passages.

D.4 Clinical Evidence Verification Loop

CEVL organizes diagnostic reasoning around hypotheses, expectations stated before analysis, contextual interpretation, and follow-up verification. EEG interpretation depends on clinical context and the diagnostic task (Kane et al., 2017; Gélisse et al., 2024; Babiloni et al., 2020), while language models can favor evidence that confirms an initial hypothesis (Jhaveri et al., 2026; Wan et al., 2025). For hypothesis h_t, investigation $q_t = (\hat{o}_t, a_t)$ pairs an expected observation with an analysis. The Environment executes a_t to obtain o_t, and interpretation r_t assesses it in context, assigning Supports, Opposes, Conflicts, or Inconclusive. The completed diagnostic record R links the final diagnosis to its contributing interpretations. The agent also reports the distinct identifiers U of the skill items used.

Algorithm 2 Clinical Evidence Verification Loop (CEVL)

Require: Agent $\mathcal{A}$, case x, task k, experience $\mathcal{S}$, budget $T_{\max}$, phase $m \in \{\text{learning}, \text{inference}\}$
Ensure: Diagnosis $\hat{y}$, diagnostic record R, used-item identifiers U

$\mathcal{E} \leftarrow \textsc{Environment}(x)$
Initialize $\mathcal{H}$ from clinical guidelines, literature, and $\mathcal{S}$
for $t = 1, 2, \ldots, T_{\max}$ **while** $\mathcal{H} \neq \emptyset$ **do**
  Select $h_t \in \mathcal{H}$ *// Hypothesis*
  $q_t = (\hat{o}_t, a_t) \leftarrow \mathcal{A}(\text{investigate} \mid h_t, \mathcal{E}, \mathcal{S})$ *// Investigation*
  $o_t \leftarrow \mathcal{E}.\text{execute}(a_t)$ *// Observation*
  $(r_t, \mathcal{H}_{\text{followup},t}) \leftarrow \mathcal{A}(\text{interpret} \mid h_t, \hat{o}_t, o_t, \mathcal{S})$ *// Interpretation*
  $\mathcal{H} \leftarrow (\mathcal{H} \setminus \{h_t\}) \cup \mathcal{H}_{\text{followup},t}$
end for
if $m =$ inference and foundation-model predictions are available **then**
  Retrieve their predictions once, after the signal-analysis loop
  Interpret each prediction alongside the accumulated signal findings
end if
Synthesize $\hat{y}$ from the signal interpretations and available model evidence *// Diagnosis*
Complete diagnostic record R and report used-item identifiers U
return $\hat{y}$, R, U

E Diagnostic Inference Graph Extraction

Table 25 gives the node schema and complete edge vocabulary used to map completed CEVL records to DIGs. The four node types preserve the hypothesis, observation, interpretation, and diagnosis roles introduced in Section 4.1. Structural edges encode workflow dependencies, while evaluative edges link interpretations to the hypotheses they assess. Observation nodes retain their Environment identifiers and execution metadata alongside the source spans preserved for all nodes.

Table 25: DIG node schema and edge vocabulary.

Element	Type & Signature	Meaning
Node	HYPOTHESIS	investigated property, expectation, action, and initial/follow-up role
	OBSERVATION	analysis output with source and execution metadata
	INTERPRETATION	contextual assessment of an observation and its assigned relation to the evaluated hypothesis
	DIAGNOSIS	final prediction, explanation, and references to the interpretations used
Structural Edge	EXECUTES: Hypothesis → Observation	the Environment executes the hypothesis's analysis a_t, producing observation o_t
	INFORMS: Observation → Interpretation	feeds measurement o_t into interpretation r_t
	TRIGGERS: Interpretation → Hypothesis	links an interpretation to a follow-up hypothesis it generates
	SYNTHESIZES: Interpretation → Diagnosis	contributes the interpretation to the final diagnostic prediction $\hat{y}$
Evaluative Edge	SUPPORTS: Interpretation → Hypothesis	judges the observations to support the hypothesis in the stated context
	OPPOSES: Interpretation → Hypothesis	judges the observations to oppose the hypothesis in the stated context
	CONFLICTS: Interpretation → Hypothesis	identifies conflicting evidence that prevents a consistent assessment of the hypothesis
	INCONCLUSIVE: Interpretation → Hypothesis	finds insufficient evidence to judge the hypothesis, including when artifacts or noise prevent assessment

The completed diagnostic record supplies the references needed to construct these edges. It contains the final diagnosis and its explanation with references to the contributing interpretations. It also retains observation identifiers, execution metadata, and the parent interpretation that generated each follow-up hypothesis.

The agent separately reports the distinct identifiers of the skill items it used in U for the subsequent usage-count update.

Figure 9 illustrates this representation with the complete A003 graph and an expanded PDR branch. Its final depression diagnosis synthesizes all ten interpretations.

The following templates summarize the fixed extraction instructions. The schema, outcome blinding, and extraction constraints remain unchanged across records.

```
EXTRACT
Input: one completed diagnostic record R. The ground-truth label and
the correctness of R are not provided.

Extract the atomic units needed to represent the diagnostic inference.
For each unit, return:
  id, type in {Hypothesis, Observation, Interpretation, Diagnosis}, text,
  source_span, and the fields required by its type.
For Hypothesis, retain the observable property, expected observation,
analysis action, role in {initial, followup}, and
parent_interpretation_id
for a follow-up hypothesis. For Observation, retain observation_id and
the source and execution metadata.
For Interpretation, retain the contextual assessment, assigned relation,
evaluated hypothesis_id, and upstream observation_ids.
For Diagnosis, retain the final prediction, explanation, and referenced
interpretation_ids from the completed diagnostic record R.
Preserve measured values and
do not add information absent from R. Represent every completed
follow-up check as another Hypothesis--Observation--Interpretation cycle.
```

```
RELATE
Input: R and the typed units returned by EXTRACT. The ground-truth
label and the correctness of R are not provided.

Link only relations stated in R. Direct structural edges according
to signature:
  Executes: Hypothesis to Observation
  Informs: Observation to Interpretation
  Triggers: Interpretation to a follow-up Hypothesis
  Synthesizes: Interpretation to Diagnosis
Create Synthesizes edges from the interpretation references recorded
in the final diagnosis.
Direct evaluative relation edges (Supports, Opposes, Conflicts,
Inconclusive) from an Interpretation to the Hypothesis it evaluates.
For each edge, return:
  source_id, target_id,
  type in {Executes, Informs, Triggers, Synthesizes,
           Supports, Opposes, Conflicts, Inconclusive},
  and source_span.
Use Inconclusive only when R explicitly states that the source does
not resolve the hypothesis.
Use Conflicts when R explicitly identifies conflicting evidence that
prevents a consistent assessment of the hypothesis.
Inconclusive covers insufficient evidence even when signal quality is
adequate. Artifacts and noise are possible reasons for this judgment.
The absence of a relation is not Inconclusive.
Do not infer physiological causality or create an edge solely because
two units occur near one another.
```

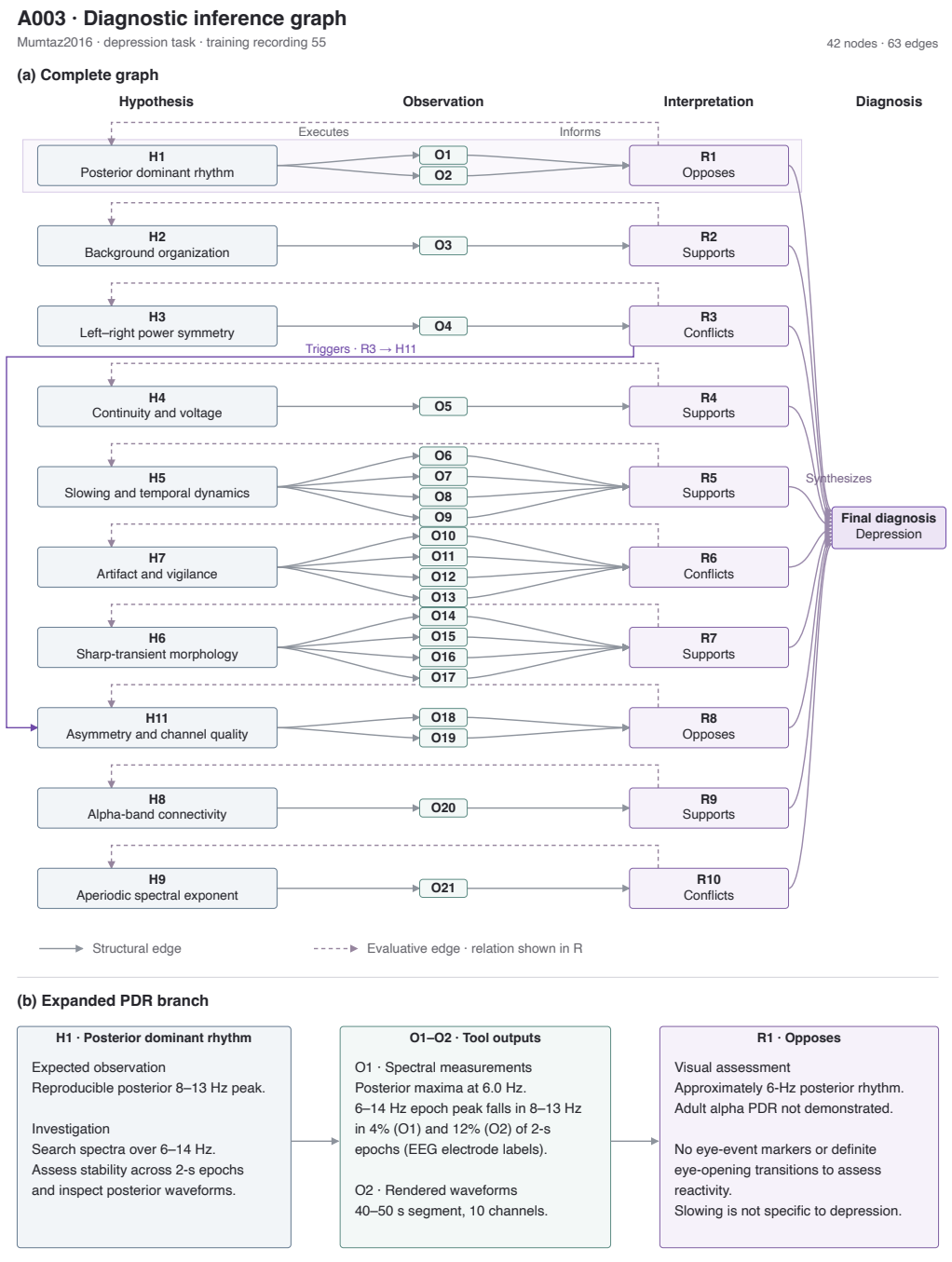


Figure 9: DIG from attempt A003 on Mumtaz2016 training recording 55. (a) Complete graph with 42 nodes and 63 edges. (b) Expanded PDR branch.

E.1 A003 NODE CONTENTS

The entries below expand all 42 nodes in Figure 9, grouped by analysis cycle. Figure labels H and R correspond to the source record's target identifiers T and cycle identifiers C, respectively. The entries reproduce the stored expectations, procedures, observation outputs, and interpretations. Waveform observations are shown at their corresponding nodes below.

H1 / R1: POSTERIOR DOMINANT RHYTHM FREQUENCY, AMPLITUDE, POSTERIOR DISTRIBUTION, INTERHEMISPHERIC SYMMETRY, AND REACTIVITY

H1 | Hypothesis (T1). Posterior dominant rhythm frequency, amplitude, posterior distribution, interhemispheric symmetry, and reactivity

Context. Resting-state adult EEG; reactivity requires identifiable eye-opening/closure transitions or event markers

Expected observation: indeterminate. A posterior alpha peak without eye-event markers leaves reactivity indeterminate; pervasive artifact or mixed vigilance may also prevent characterization.

Expected observation: refutes. No posterior 8-13 Hz peak, a nonposterior maximum, or a marked homologous difference refutes a clearly present symmetric PDR.

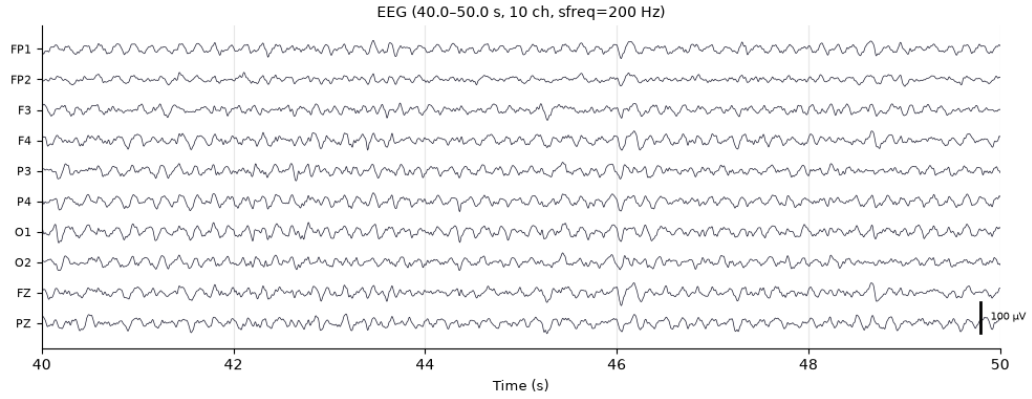
Expected observation: supports. A reproducible 8-13 Hz peak maximal over O1/O2 and P3/P4, with comparable bilateral frequency/amplitude, supports a present symmetric PDR.

Investigation. Compute Welch spectra over 2-s epochs for posterior and frontal channels across the full record, identify 6-14 Hz peaks, alpha amplitudes and posterior/anterior ratios, and inspect representative 10-s posterior-inclusive waveforms for rhythmic morphology and eye-related attenuation if visible.

O1 | Observation **Tool.** run_python

```
posterior peaks {'O1': (6.0, 95.99881744384766), 'O2': (6.0, 43.947105407714844), '
    P3': (6.0, 49.52297592163086), 'P4': (6.0, 83.51531982421875)}
alpha power {'O1': np.float32(26.74), 'O2': np.float32(33.371), 'P3': np.float32
    (27.487), 'P4': np.float32(24.065), 'FP1': np.float32(15.397), 'FP2': np.
    float32(17.237), 'F3': np.float32(23.673), 'F4': np.float32(30.212)}
post/front alpha ratio 1.2906282
O1 epoch peak median/IQR 6.0 [6. 6.] 8-13 fraction 0.04 alpha_rms_uV
    3.717875332643846
O2 epoch peak median/IQR 6.0 [6. 6.] 8-13 fraction 0.12 alpha_rms_uV
    3.684045900917395
```

O2 | Waveform observation. Tool: show_eeg.
Displayed interval. 40.0–50.0 s; channels: FP1, FP2, F3, F4, P3, P4, O1, O2, FZ, PZ



R1 | Interpretation (C1) Relation to H1. Opposes

Observation references. O1, O2

Comparison with expectation. The frozen support criterion required a reproducible posterior 8-13 Hz peak. Quantitatively, all posterior spectra instead peaked at 6.0 Hz, only 4% of O1 and 12% of O2 2-s epochs had an 8-13 Hz maximum, and posterior/frontal alpha power ratio was only 1.29. Morphologically, the rendered posterior traces showed a fairly regular approximately 6-Hz posterior rhythm rather than a sustained alpha-frequency PDR. Thus the observed data refute a normal alpha-frequency PDR; bilateral alpha RMS was similar (~3.7 uV), but eye-opening reactivity could not be established.

Recorded interpretation. A posteriorly emphasized, approximately 6-Hz rhythmic background is present, but an adult alpha-frequency PDR is not demonstrated. Posterior rhythm frequency is bilaterally similar; amplitude symmetry is addressed more fully under homologous-channel analysis. PDR reactivity is not assessable because no eye-event markers or definite eye-opening transitions were available. Does this result raise a new target? Yes: whether the 6-Hz rhythm reflects persistent cerebral slowing versus drowsiness or low-frequency artifact is already captured by T5 and T7.

Implication for depression. Nonspecific posterior rhythm slowing can occur for many state, medication, or encephalopathic reasons and provides no direct depression classification.

Implication for no_depression. The absence of a normal alpha PDR does not establish depression and therefore only weakens any inference based on a conventionally normal awake background.

H2 / R2: PREDOMINANT BACKGROUND FREQUENCY, ANTERIOR-POSTERIOR ORGANIZATION, AND NONDOMINANT BETA, THETA, AND DELTA DISTRIBUTIONS

H2 | Hypothesis (T2). Predominant background frequency, anterior-posterior organization, and nondominant beta, theta, and delta distributions

Context. Most awake or alert portions of the full resting recording

Expected observation: indeterminate. Mixed state or severe artifact prevents assigning predominant frequency or gradient.

Expected observation: refutes. No stable predominant band, reversed distribution, or artifact-dominated power refutes an organized background characterization.

Expected observation: supports. A stable band clearly dominates globally, with posterior enhancement and a coherent anterior-posterior gradient; nondominant bands have quantifiable lesser contributions.

Investigation. Calculate robust 2-s epoch relative delta/theta/alpha/beta power by channel and region, global dominant band, anterior/posterior ratios, and temporal distributions over the full recording.

O3 | Observation Tool. run_python

```
global median fractions {'delta': np.float64(0.113), 'theta': np.float64(0.567), '
    alpha': np.float64(0.118), 'beta': np.float64(0.065)}
epoch global dominant proportions {'delta': np.float64(0.02), 'theta': np.float64
    (0.88), 'alpha': np.float64(0.1), 'beta': np.float64(0.0)}
front {'delta': np.float64(0.145), 'theta': np.float64(0.528), 'alpha': np.float64
    (0.12), 'beta': np.float64(0.062)}
central {'delta': np.float64(0.148), 'theta': np.float64(0.479), 'alpha': np.
    float64(0.139), 'beta': np.float64(0.065)}
posterior {'delta': np.float64(0.085), 'theta': np.float64(0.605), 'alpha': np.
    float64(0.114), 'beta': np.float64(0.068)}
delta post/front median IQR 0.65 [0.469 0.917]
theta post/front median IQR 1.142 [0.993 1.293]
alpha post/front median IQR 0.901 [0.665 1.199]
beta post/front median IQR 1.107 [0.921 1.372]
```

R2 | Interpretation (C2) Relation to H2. Supports

Observation references. O3

Comparison with expectation. The frozen support criterion required a clearly dominant, stable band and quantifiable spatial organization. Theta comprised a median 56.7% of 1-30 Hz power and was globally dominant in 88% of epochs, with posterior theta enrichment (posterior median 60.5%, posterior/frontal ratio 1.14). Alpha was only 11.8% globally and lacked posterior dominance (ratio 0.90). This supports a stable theta-predominant background with a posterior gradient for theta, not a normal posterior alpha gradient.

Recorded interpretation. Predominant background frequency is theta, centered near 6 Hz. Non-dominant activity includes delta 11.3%, alpha 11.8%, and beta 6.5% median relative power; delta is frontally weighted, theta posteriorly weighted, and alpha has no posterior gradient. Does this result raise a new target? No new target beyond T5 (persistence/cerebral versus drift), T7 (vigilance/artifact), and T9 (aperiodic contribution), which are already pending.

Implication for depression. Theta predominance is nonspecific; it may be seen with drowsiness, medication effects, developmental factors, or diffuse dysfunction and is not a depression criterion.

Implication for no_depression. The atypical awake-background organization prevents describing the EEG as conventionally normal, but abnormality itself does not discriminate cohort membership.

H3 / R3: Homologous left-right spectral power symmetry, including frontal alpha asymmetry and whether any difference is stable over time

H3 | Hypothesis (T3). Homologous left-right spectral power symmetry, including frontal alpha asymmetry and whether any difference is stable over time

Context. Resting EEG; interpretation depends on asymmetry convention, reference, vigilance, and artifact

Expected observation: indeterminate. Artifact, reference dependence, or broad distributions make laterality unreliable.

Expected observation: refutes. Marked (>=50%) voltage/power-equivalent asymmetry, >1 Hz frequency difference, or unstable sign across epochs refutes stable symmetry or stable frontal asymmetry.

Expected observation: supports. Homologous channel alpha powers and peak frequencies show no marked differences, or any frontal alpha difference has a consistent left-versus-right direction across epochs.

Investigation. Compute homologous-pair broadband RMS and band powers, percent side difference and peak-frequency differences, plus log alpha right-minus-left for F3/F4, F7/F8, and FP1/FP2 in artifact-robust 2-s epochs with bootstrap confidence intervals and sign stability.

O4 | Observation Tool. run_python

```
('FP1', 'FP2') alpha L R pctdiff 6.05 6.87 12.6 RMS_pctdiff 29.5 peakdiffHz 0.0
('F3', 'F4') alpha L R pctdiff 9.67 14.1 37.3 RMS_pctdiff 16.7 peakdiffHz 0.0
('C3', 'C4') alpha L R pctdiff 12.78 9.71 27.3 RMS_pctdiff 14.5 peakdiffHz 0.0
('P3', 'P4') alpha L R pctdiff 12.86 9.4 31.1 RMS_pctdiff 9.6 peakdiffHz 0.0
('O1', 'O2') alpha L R pctdiff 12.1 13.11 8.0 RMS_pctdiff 19.4 peakdiffHz 0.0
('F7', 'F8') alpha L R pctdiff 5.8 10.25 55.5 RMS_pctdiff 30.3 peakdiffHz 0.0
('T7', 'T8') alpha L R pctdiff 7.78 5.11 41.4 RMS_pctdiff 20.2 peakdiffHz 0.0
('P7', 'P8') alpha L R pctdiff 12.05 5.56 73.6 RMS_pctdiff 62.4 peakdiffHz 0.0
('FP1', 'FP2') logalpha R-L median CI signpos 0.08554647322733833 [0.01812829
    0.15372741] 0.6071428571428571
('F3', 'F4') logalpha R-L median CI signpos 0.26980730123273655 [0.2216183
    0.35695199] 0.7785714285714286
('F7', 'F8') logalpha R-L median CI signpos 0.5263905879621921 [0.45016893
    0.61624136] 0.8676470588235294
```

R3 | Interpretation (C3) Relation to H3. Conflicts

Observation references. O4

Comparison with expectation. The frozen criterion anticipated either broadly symmetric homologous activity or a stable frontal alpha direction, whereas marked >=50% alpha differences would refute global symmetry. Frontal alpha was consistently greater on the right (right-minus-left log alpha medians 0.086 FP, 0.270 F3/F4, 0.526 F7/F8; positive in 61%, 78%, and 87% of retained epochs), but F7/F8 had a 55.5% alpha difference. P7/P8 also showed marked alpha (73.6%) and RMS (62.4%) differences, while central and occipital pairs were mild and all peak-frequency differences were 0 Hz. Thus stable right-greater frontal alpha coexists with marked lateral-pair amplitude asymmetry, conflicting with global symmetry.

Recorded interpretation. There is reproducible right-greater frontal alpha power, strongest at F8 relative to F7, plus marked P7/P8 amplitude asymmetry; central/occipital frequency is symmetric. Translating the convention explicitly, positive log(right alpha)-log(left alpha) means greater right alpha power, not greater right cortical activation. This is reference-, state-, and artifact-sensitive and does not establish depression. Does this result raise a new target? Yes: determine whether lateral temporal/posterior asymmetry is driven by channel quality or focal artifact; this is incorporated as a finding-linked addition.

Implication for depression. Right-greater frontal alpha is a weak, inconsistent group-level research association whose direction depends on power-versus-activation convention; it provides at most limited support and cannot classify individually.

Implication for no_depression. Stable asymmetry is not typical evidence of perfect symmetry, but marked lateral differences may be technical; it provides only limited counterevidence pending artifact assessment.

Resolution limit. No reference montage was supplied, and alpha asymmetry lacks validated individual-level diagnostic specificity.

H4 / R4: Background continuity, attenuation or suppression burden, and typical voltage

H4 | Hypothesis (T4). Background continuity, attenuation or suppression burden, and typical voltage

Context. Full recording with separation of temporal continuity from amplitude statistic

Expected observation: indeterminate. Estimator or montage mismatch prevents applying voltage categories; continuity can still be reported separately.

Expected observation: refutes. Typical amplitude <20 uV, suppression <10 uV, or >=10% attenuation/suppression refutes continuous normal-voltage background.

Expected observation: supports. Typical peak-to-trough or robust window amplitude is >=20 uV with <1% attenuated/suppressed windows, supporting continuous normal-voltage background.

Investigation. Compute per-channel and global 1-s peak-to-peak amplitudes, robust percentiles, and percentages of windows below 20 and 10 uV; separately estimate attenuation as windows below 50% of each channel's median background amplitude.

O5 | Observation Tool. run_python

```
channel median 1s ptp uV {'FP1': np.float32(41.3), 'FP2': np.float32(33.2), 'F3':
    np.float32(45.1), 'F4': np.float32(52.8), 'C3': np.float32(51.0), 'C4': np.
    float32(45.7), 'P3': np.float32(46.4), 'P4': np.float32(48.0), 'O1': np.
    float32(52.4), 'O2': np.float32(45.1), 'F7': np.float32(37.2), 'F8': np.
    float32(50.3), 'T7': np.float32(41.4), 'T8': np.float32(34.9), 'P7': np.
    float32(65.8), 'P8': np.float32(38.6), 'FZ': np.float32(51.7), 'CZ': np.
    float32(46.5), 'PZ': np.float32(52.3)}
global window-channel ptp percentiles [25.7 30.6 39.1 46.3 55.7 75.  95.1]
percent <20uV 0.03 percent <10uV 0.0
attenuation below half channel median percent 0.0
global median-window ptp 45.716087 pct global windows <half 0.0 <20 0.0 <10 0.0
```

R4 | Interpretation (C4) Relation to H4. Supports

Observation references. O5

Comparison with expectation. The frozen support criterion required typical amplitude >=20 uV and <1% attenuation/suppression. Median 1-s peak-to-peak amplitudes were 33.2-65.8 uV by channel, global median 46.3 uV, only 0.03% of channel-windows were <20 uV, none were <10 uV, and no global windows fell below half the median background. This matches continuous normal-voltage background.

Recorded interpretation. Background is continuous with normal typical voltage by the measured referential window statistic: 0% global attenuation/suppression and no suppressed windows. Does this result raise a new target? No; amplitude estimator and montage limitations will be retained in the report.

Implication for depression. Continuous normal-voltage background is a technical negative finding and does not exclude depression.

Implication for no_depression. This conventional background property is compatible with no_depression but is nondiscriminative because depression often has normal routine EEG voltage and continuity.

H5 / R5: Presence, frequency band, topography, continuity, and temporal dynamics of focal or generalized slowing

H5 | Hypothesis (T5). Presence, frequency band, topography, continuity, and temporal dynamics of focal or generalized slowing

Context. Full recording; low-frequency activity must be distinguished from drift, ocular artifact, and drowsiness

Expected observation: indeterminate. Mixed vigilance or unavailable state markers prevents separating drowsiness from pathological slowing despite drift control.

Expected observation: refutes. Low-frequency excess substantially disappears after detrending or is confined to ocular/frontal channels and transient periods, refuting persistent cerebral slowing.

Expected observation: supports. Theta/delta excess remains after drift control, is spatially coherent rather than confined to frontal poles, and persists across most epochs, supporting true generalized slowing.

Investigation. Apply justified 1-Hz drift control because the posterior peak is 6 Hz and no alpha PDR is clear; then recompute regional band fractions and epoch persistence with comparable settings, and visually inspect early/middle/late representative windows for generalized versus frontal low-frequency morphology.

O6 | Observation Tool. detrend

```
Detrend + 1.0 Hz zero-phase high-pass applied. Sub-2Hz relative power: 7.6% ->
    7.0%.
```

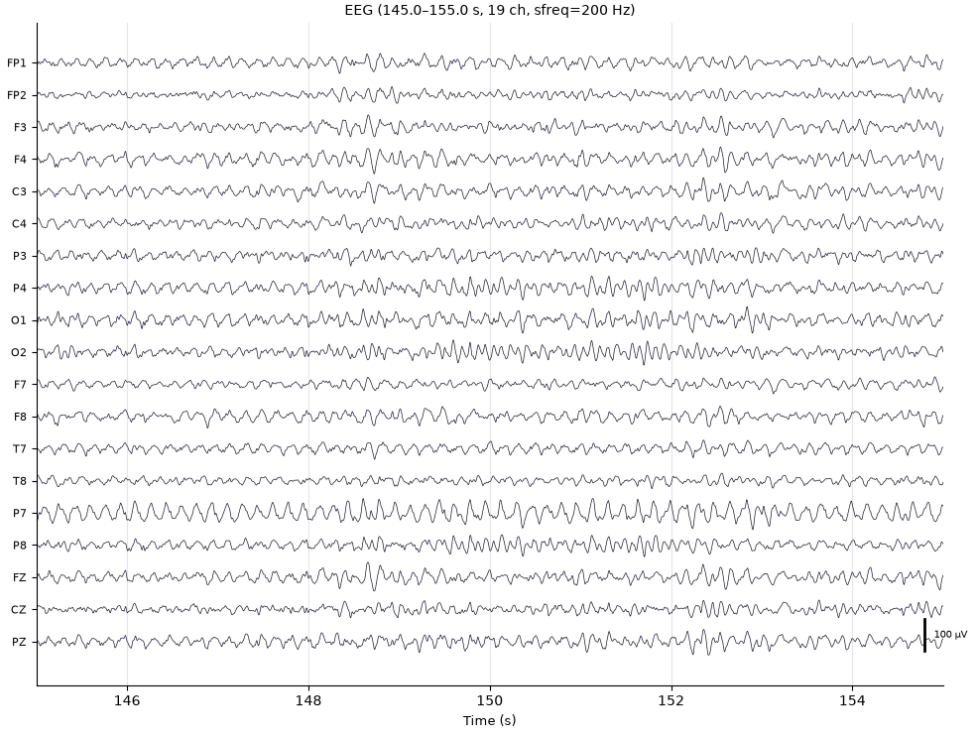
O7 | Observation Tool. run_python

```
post-detrend global median {'delta': np.float64(0.113), 'theta': np.float64(0.567),
     'alpha': np.float64(0.118), 'beta': np.float64(0.065)}
epoch dominant {'delta': np.float64(0.02), 'theta': np.float64(0.88), 'alpha': np.
    float64(0.1), 'beta': np.float64(0.0)}
early {'delta': np.float64(0.092), 'theta': np.float64(0.556), 'alpha': np.float64
    (0.144), 'beta': np.float64(0.066)}
middle {'delta': np.float64(0.129), 'theta': np.float64(0.564), 'alpha': np.
    float64(0.103), 'beta': np.float64(0.064)}
late {'delta': np.float64(0.122), 'theta': np.float64(0.576), 'alpha': np.float64
    (0.112), 'beta': np.float64(0.064)}
delta {'FP1': np.float64(0.09), 'FP2': np.float64(0.24), 'F3': np.float64(0.13), '
    F4': np.float64(0.13), 'C3': np.float64(0.11), 'C4': np.float64(0.15), 'P3':
```

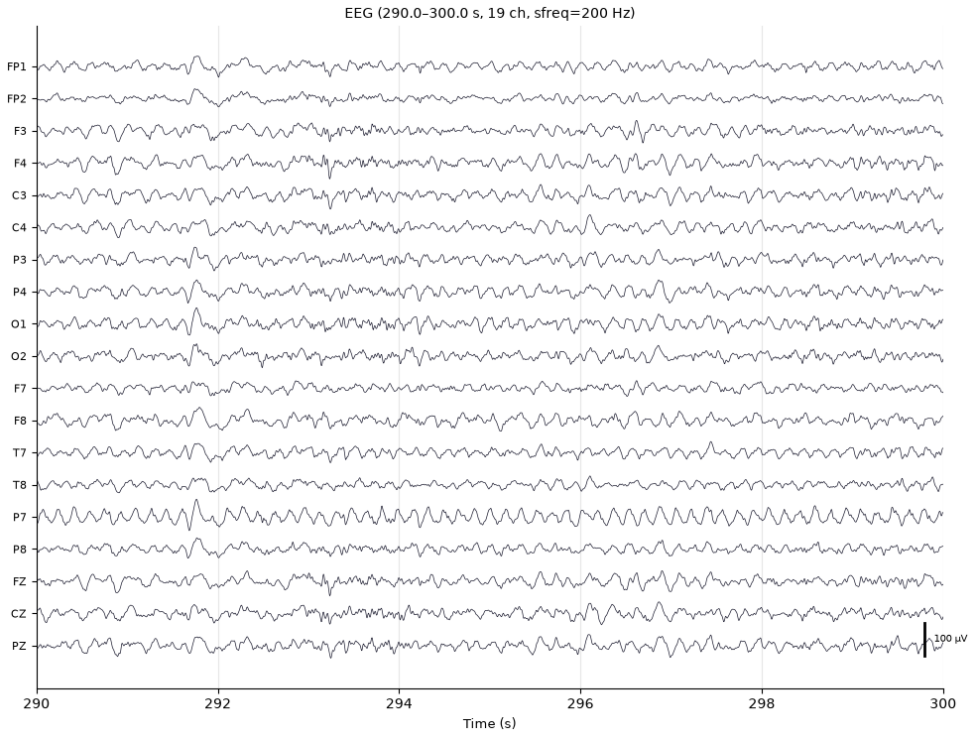
```
    np.float64(0.13), 'P4': np.float64(0.1), 'O1': np.float64(0.08), 'O2': np.
    float64(0.11), 'F7': np.float64(0.13), 'F8': np.float64(0.18), 'T7': np.
    float64(0.06), 'T8': np.float64(0.1), 'P7': np.float64(0.03), 'P8': np.float64
    (0.09), 'FZ': np.float64(0.14), 'CZ': np.float64(0.19), 'PZ': np.float64(0.11)}

theta {'FP1': np.float64(0.66), 'FP2': np.float64(0.28), 'F3': np.float64(0.5), '
    F4': np.float64(0.53), 'C3': np.float64(0.6), 'C4': np.float64(0.5), 'P3': np.
    float64(0.5), 'P4': np.float64(0.63), 'O1': np.float64(0.64), 'O2': np.float64
    (0.51), 'F7': np.float64(0.59), 'F8': np.float64(0.56), 'T7': np.float64(0.74),
     'T8': np.float64(0.62), 'P7': np.float64(0.86), 'P8': np.float64(0.64), 'FZ':
     np.float64(0.51), 'CZ': np.float64(0.32), 'PZ': np.float64(0.57)}
```

O8 | Waveform observation. Tool: show_eeg.
Displayed interval. 145.0–155.0 s; all recorded channels



O9 | Waveform observation. Tool: show_eeg.
Displayed interval. 290.0–300.0 s; all recorded channels



R5 | Interpretation (C5) **Relation to H5.** Supports

Observation references. O6, O7, O8, O9

Comparison with expectation. The frozen support criterion required low-frequency excess to persist after drift control, remain spatially coherent, and occupy most epochs. Sub-2-Hz power changed only from 7.6% to 7.0% after detrending, while theta remained 56.7% globally and dominant in 88% of epochs. Early, middle, and late theta fractions were similar (55.6%, 56.4%, 57.6%). Visual samples showed widespread rhythmic 5-7 Hz activity across both hemispheres, especially posteriorly, not merely frontal drift. This matches persistent generalized rhythmic theta slowing, although vigilance remains unresolved.

Recorded interpretation. There is continuous/persistent generalized theta slowing at approximately 5-7 Hz, posteriorly accentuated, with only modest frontal delta. It is robust to 1-Hz drift control and not explained by baseline drift alone. The morphology is rhythmic and broadly distributed; no consistent focal polymorphic delta focus is evident in sampled windows. Does this result raise a new target? No new target: distinguishing drowsiness from other causes is already T7.

Implication for depression. Persistent generalized theta slowing is not a recognized individual diagnostic marker of depression and may reflect vigilance, medication, age, or diffuse cerebral dysfunction; no direct label assignment follows.

Implication for no_depression. The finding argues against a conventionally normal alert adult EEG, but it does not specifically argue against no_depression cohort membership because nonspecific slowing can occur independently.

Resolution limit. Without age, medications, behavioral state, eye markers, or clinical examination, physiological drowsiness cannot be fully separated from pathological generalized slowing.

H7 / R6: Artifact types and distributions, vigilance states, state transitions, and their impact on interpretation

H7 | Hypothesis (T7). Artifact types and distributions, vigilance states, state transitions, and their impact on interpretation

Context. Resting recording without supplied event markers

Expected observation: indeterminate. Absence of EOG, EMG, video, and event markers prevents definitive vigilance or artifact attribution.

Expected observation: refutes. Clear vertex waves, spindles/K-complexes, progressive alpha dropout, or pervasive ocular/muscle artifact refutes stable awake-state interpretation.

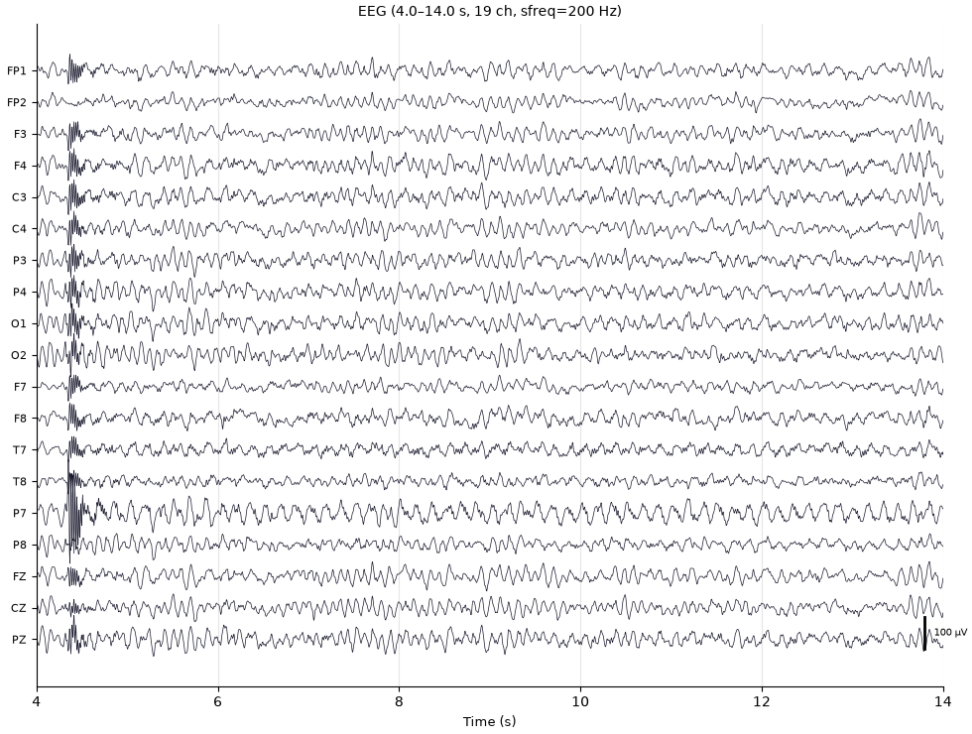
Expected observation: supports. Stable rhythmic background without sleep transients, major eye/muscle contamination, or abrupt state transitions supports a predominantly awake or quietly drowsy interpretable record.

Investigation. Quantify epoch-wise frontal delta, posterior theta/alpha, high-frequency muscle proxy, and amplitude outliers; identify transition and extreme-artifact epochs, then inspect selected windows for eye movements, muscle, sleep transients, and state changes.

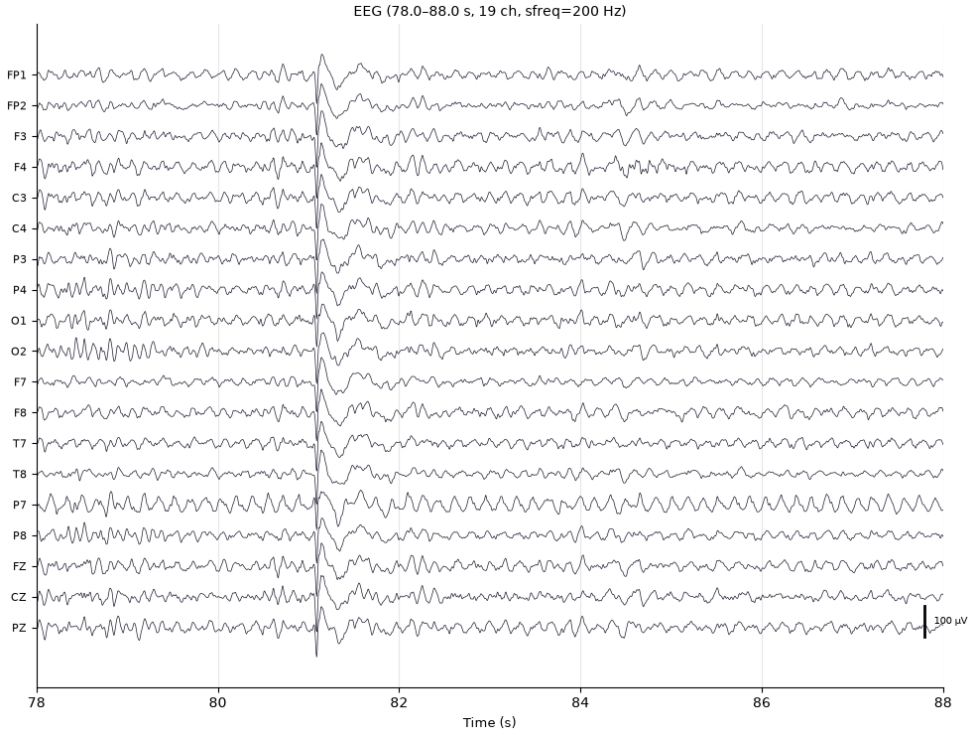
O10 | Observation **Tool.** run_python

```
features median IQR frontalDelta postTheta postAlpha lateral30-45 medianPTP
0 12.659 [ 7.583 19.176]
1 59.686 [52.971 70.862]
2 12.515 [ 8.348 31.136]
3 1.111 [0.981 1.284]
4 50.611 [46.55  58.257]
frontal_delta top epoch starts [np.int64(80), np.int64(110), np.int64(218), np.
    int64(292), np.int64(182)] z [np.float64(27.1), np.float64(19.1), np.float64
    (11.4), np.float64(5.8), np.float64(5.5)]
post_theta top epoch starts [np.int64(80), np.int64(134), np.int64(40), np.int64
    (136), np.int64(296)] z [np.float64(9.9), np.float64(4.5), np.float64(3.7), np.
    float64(3.5), np.float64(3.2)]
post_alpha top epoch starts [np.int64(74), np.int64(4), np.int64(76), np.int64(132)
    , np.int64(134)] z [np.float64(32.5), np.float64(25.7), np.float64(21.2), np.
    float64(21.0), np.float64(20.5)]
muscle30_45 top epoch starts [np.int64(4), np.int64(8), np.int64(16), np.int64(6),
     np.int64(12)] z [np.float64(45.1), np.float64(31.9), np.float64(29.1), np.
    float64(27.6), np.float64(27.1)]
amplitude top epoch starts [np.int64(80), np.int64(4), np.int64(134), np.int64(74),
     np.int64(52)] z [np.float64(15.4), np.float64(5.7), np.float64(3.9), np.
    float64(3.6), np.float64(3.3)]
trend frontal_delta -0.022
trend post_theta 0.046
trend post_alpha -0.236
trend muscle -0.341
trend amp -0.118
```

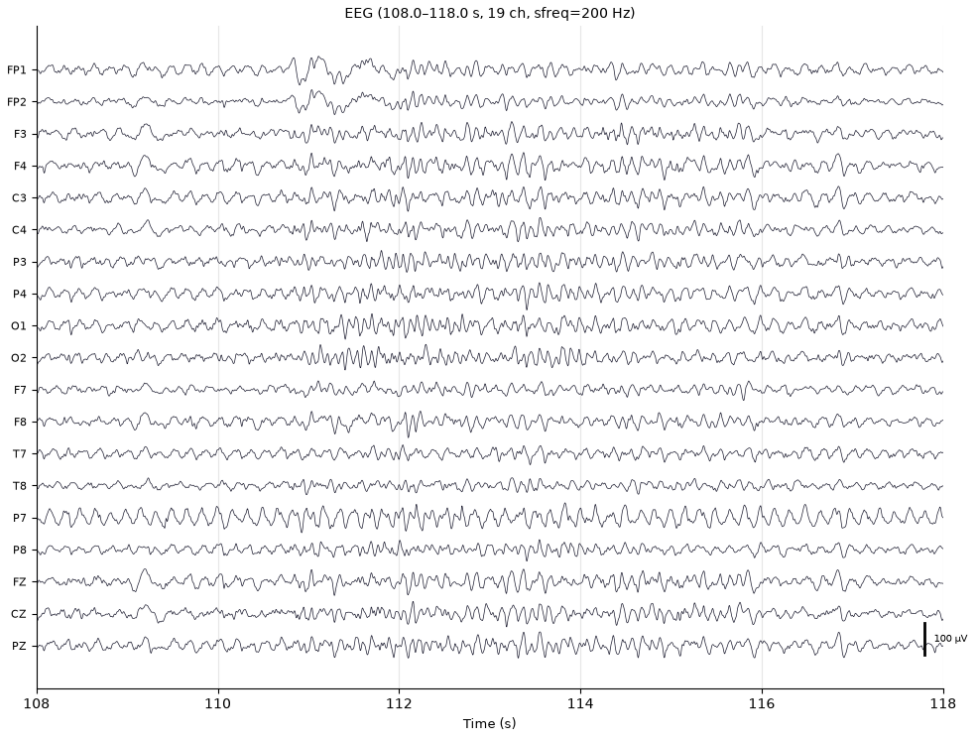
O11 | Waveform observation. Tool: show_eeg.
Displayed interval. 4.0–14.0 s; all recorded channels



O12 | Waveform observation. Tool: show_eeg.
Displayed interval. 78.0–88.0 s; all recorded channels



O13 | Waveform observation. Tool: show_eeg.
Displayed interval. 108.0–118.0 s; all recorded channels



R6 | Interpretation (C6) Relation to H7. Conflicts

Observation references. O10, O11, O12, O13

Comparison with expectation. The frozen support criterion anticipated stable interpretable background without major artifact or sleep transients; the record is largely stable but contains discrete artifacts. Quantitative high-frequency activity was strongly concentrated early (top muscle-proxy epochs 4-16 s; negative time trend -0.341), and frontal-delta outliers occurred at 80, 110, 218, 292, and 182 s. Visual review showed a brief generalized high-frequency movement/electrode artifact near 4.4 s and large synchronous slow transients around 81 and 111 s, likely movement/ocular artifact rather than cerebral discharges. Between these, the 5-7 Hz rhythm remained organized. No definite spindles, K-complexes, or vertex waves were seen in inspected windows, but absent EOG/video makes awake versus drowsy state indeterminate.

Recorded interpretation. The recording is interpretable with intermittent movement/electrode and likely ocular slow artifacts, strongest early and at isolated times; artifact is not pervasive enough to explain the sustained theta rhythm. Vigilance appears awake-to-drowsy or indeterminate: no definite N2 sleep transients were observed, but a normal alpha PDR is absent and no behavioral/eye markers are available. Does this result raise a new target? No; T11 will specifically assess lateral channel quality.

Implication for depression. Drowsiness and intermittent artifact can alter power/asymmetry and reduce confidence in research biomarkers; they do not specifically support depression.

Implication for no_depression. The same state/artifact uncertainty prevents using slowing or asymmetry as strong evidence against no_depression.

Resolution limit. Definitive vigilance, eye-opening reactivity, and ocular-artifact identification are not assessable without EOG, video, or event markers.

H6 / R7: Morphology, location, and prevalence of epileptiform discharges or other sharp transients

H6 | Hypothesis (T6). Morphology, location, and prevalence of epileptiform discharges or other sharp transients

Context. Visual review across representative and quantitatively selected high-transient intervals

Expected observation: indeterminate. Movement/electrode artifacts dominate candidate events or limited sampling prevents reliable morphology assessment.

Expected observation: refutes. Recurrent sharply contoured transients with consistent spatial field, duration, and aftergoing slow wave refute absence and support epileptiform activity.

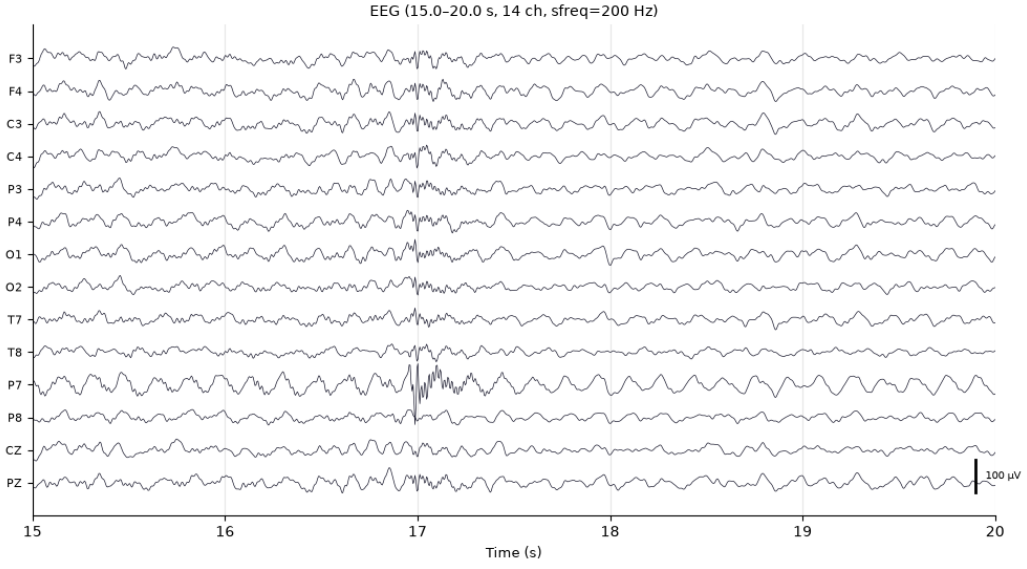
Expected observation: supports. No reproducible spikes or sharp waves with plausible field and aftergoing slow wave across quantitative high-kurtosis intervals and representative samples supports absence of epileptiform discharges.

Investigation. Detect high line-length/kurtosis 1-s windows after excluding global synchronous artifact, list candidate times and channel concentration, then visually inspect the highest remaining candidates for spike/sharp-wave morphology, field, and aftergoing slow wave.

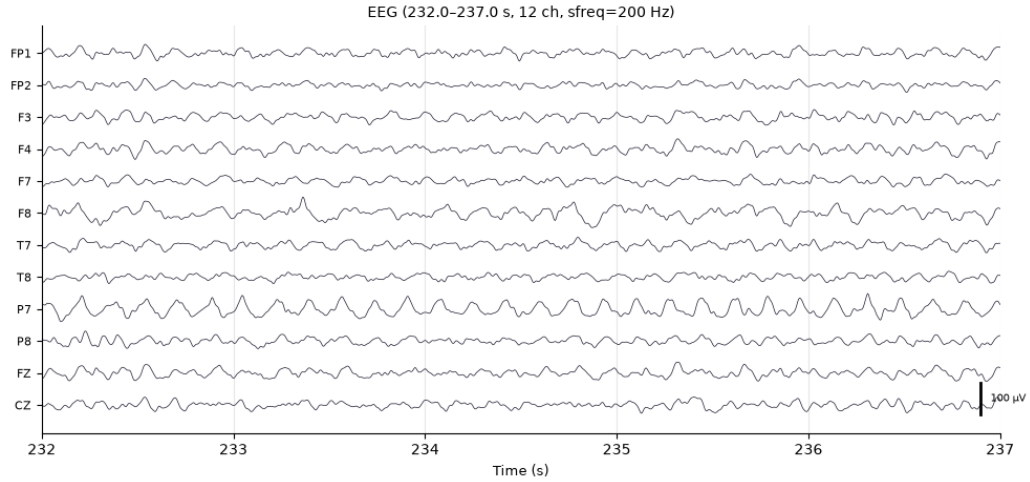
O14 | Observation **Tool.** run_python

```
candidate focal transient windows: score,start_s,channel,global_ptp,line_length,
    kurtosis
13.32 16 P7 58.0 7.28 6.25
9.49 17 P7 54.7 8.21 2.96
8.95 234 F8 41.2 2.92 2.8
8.04 251 F8 42.1 2.83 5.31
7.58 282 P7 44.1 2.81 2.14
7.31 44 P7 35.9 3.07 1.86
6.87 63 P7 41.1 3.14 2.17
6.81 161 CZ 50.2 2.12 6.88
6.71 245 P7 41.4 3.23 2.23
6.29 128 P7 44.0 3.12 1.86
6.22 127 P7 34.9 3.22 2.02
6.05 204 P7 47.1 3.32 2.1
```

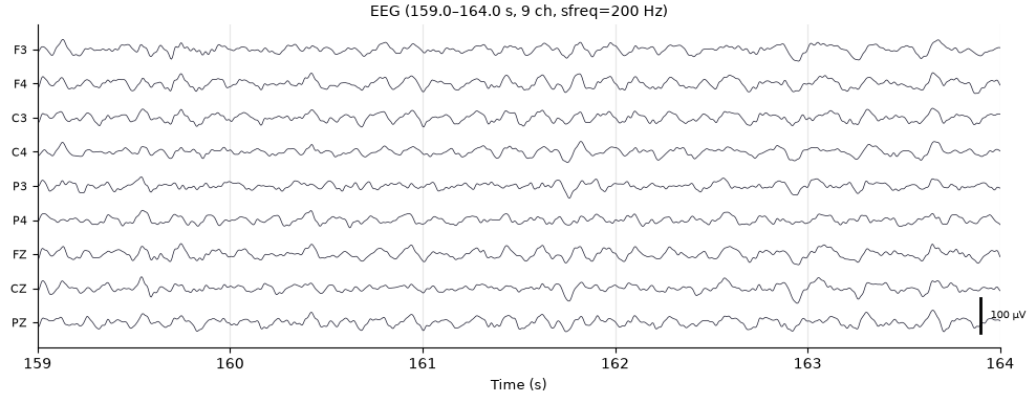
O15 | Waveform observation. Tool: show_eeg.
Displayed interval. 15.0–20.0 s; channels: F3, F4, C3, C4, P3, P4, O1, O2, T7, T8, P7, P8, CZ, PZ



O16 | Waveform observation. Tool: show_eeg.
Displayed interval. 232.0–237.0 s; channels: FP1, FP2, F3, F4, F7, F8, T7, T8, P7, P8, FZ, CZ



O17 | Waveform observation. Tool: show_eeg.
Displayed interval. 159.0–164.0 s; channels: F3, F4, C3, C4, P3, P4, FZ, CZ, PZ



R7 | Interpretation (C7) **Relation to H6.** Supports

Observation references. O14, O15, O16, O17

Comparison with expectation. The frozen support criterion required no reproducible spikes/sharp waves with a plausible field and aftergoing slow wave. Quantitative candidates were dominated by P7 rhythmic/high-line-length activity and isolated F8/CZ kurtosis. Visual inspection showed brief multichannel high-frequency movement/electrode artifact near 17 s, smoothly contoured rhythmic theta at P7/F8, and a broad slow transient near 162.8 s without a focal spike field or aftergoing slow wave. No recurrent epileptiform morphology was demonstrated, matching the support criterion.

Recorded interpretation. No epileptiform discharges are identified in representative and quantitatively enriched candidate windows. Candidate sharp activity is better explained by movement/electrode artifact or sharply contoured rhythmic theta without a consistent field or aftergoing slow wave. Does this result raise a new target? No; persistent P7 amplitude prominence remains addressed by T11.

Implication for depression. Absence of epileptiform activity removes an obvious neurologic comorbidity in sampled data but does not support or exclude depression.

Implication for no_depression. Likewise nondiscriminative for cohort membership.

Resolution limit. A 5-minute routine recording cannot exclude rare interictal discharges.

H11 / R8: Whether marked F7/F8 and P7/P8 asymmetries reflect channel-quality differences, focal artifact, or reproducible cerebral activity

H11 | Hypothesis (T11). Whether marked F7/F8 and P7/P8 asymmetries reflect channel-quality differences, focal artifact, or reproducible cerebral activity

Context. Homologous lateral channels showed >=50% differences while midline-adjacent pairs did not

Expected observation: indeterminate. No impedance/reference metadata and mixed artifact prevent distinguishing technical from cerebral origin.

Expected observation: refutes. Similar spectral shape, physiologic correlation with adjacent channels, and stable rhythmic morphology without excess high-frequency/noise refutes a purely technical explanation and supports reproducible cerebral or reference-related asymmetry.

Expected observation: supports. Persistent variance, spectral-shape, or high-frequency abnormalities confined to F8/P7 across clean epochs, with poor correlation to adjacent channels, supports channel-quality or focal artifact as the source of marked asymmetry.

Investigation. Run channel-quality detection on the current detrended data and compare F7/F8/P7/P8 variance, line-noise/high-frequency fractions, neighbor correlations, and epoch stability against homologous and adjacent channels.

Follow-up from. C3

O18 | Observation **Tool.** detect_bad_channels

```
No bad channels detected.
```

O19 | Observation **Tool.** run_python

```
F7 var 71.9 hf_frac 0.015 line_frac 0.00066 neighbor_corr {'FP1': np.float64
    (-0.034), 'F3': np.float64(0.85), 'T7': np.float64(0.316)}
F8 var 137.4 hf_frac 0.0125 line_frac 0.00044 neighbor_corr {'FP2': np.float64
    (0.415), 'F4': np.float64(0.753), 'T8': np.float64(0.296)}
P7 var 292.9 hf_frac 0.0131 line_frac 0.00257 neighbor_corr {'T7': np.float64
    (0.331), 'P3': np.float64(0.267), 'O1': np.float64(0.865)}
P8 var 89.3 hf_frac 0.0151 line_frac 4e-05 neighbor_corr {'T8': np.float64(-0.014),
     'P4': np.float64(0.911), 'O2': np.float64(0.365)}
('F7', 'F8') logR/L median IQR 0.292 [0.212 0.427] signpos 0.987
('P7', 'P8') logR/L median IQR -0.627 [-0.683 -0.557] signpos 0.0
```

R8 | Interpretation (C8) **Relation to H11.** Opposes

Observation references. O18, O19

Comparison with expectation. The frozen technical-artifact expectation required isolated noise-like spectral abnormalities or poor physiologic correlation. Automated screening found no bad channels. F8 and P7 had higher variance, but high-frequency fractions were low and similar (1.25-1.51%), and both correlated strongly with an adjacent cerebral channel (F8-F4 r=0.753; P7-O1 r=0.865). The asymmetries were extremely stable across epochs (F8>F7 in 98.7%; P7>P8 in 100%). Thus a purely noisy-channel explanation is refuted; stable reference-related or cerebral amplitude differences remain possible.

Recorded interpretation. Marked lateral asymmetries are reproducible and not explained by obvious high-frequency or line-noise channel failure. F8 has greater amplitude than F7 and P7 greater than P8 across nearly all epochs; however, absent impedance and reference metadata prevents deciding whether this is cerebral, anatomical, or reference-related. Does this result raise a new target? No; the uncertainty is a reporting limitation rather than a distinct unresolved signal property.

Implication for depression. Reproducible right-greater frontal alpha remains weakly compatible with some research reports, but stable P7 asymmetry and reference uncertainty make it nonspecific and insufficient for individual classification.

Implication for no_depression. The asymmetries reduce confidence in a fully symmetric normal technical description but do not specifically oppose no_depression.

Resolution limit. No electrode impedances or original reference description were supplied.

H8 / R9: Interregional alpha-band functional connectivity and its temporal stability

H8 | Hypothesis (T8). Interregional alpha-band functional connectivity and its temporal stability

Context. Resting EEG; coherence is sensitive to reference, segment length, volume conduction, and artifact

Expected observation: indeterminate. Low alpha power, unknown reference, and volume conduction may prevent disease-directed interpretation even if estimates are stable.

Expected observation: refutes. Connectivity is uniformly saturated, dominated by zero-lag/reference effects, or unstable across matched temporal subsets, refuting interpretable topology.

Expected observation: supports. Alpha-band coherence shows reproducible regional topology across matched epochs and homologous pairs, supporting an interpretable higher-order connectivity estimate.

Investigation. Compute magnitude-squared coherence in theta and alpha for predefined homologous frontal, central, posterior, and long-range pairs using identical 4-s segments; compare first versus second half and report pairwise topology and stability.

O20 | Observation Tool. run_python

```
('FP1', 'FP2') half1 theta alpha, half2, full [0.781 0.796 0.786 0.769 0.777
    0.782]
('F3', 'F4') half1 theta alpha, half2, full [0.723 0.725 0.736 0.696 0.735 0.708]
('C3', 'C4') half1 theta alpha, half2, full [0.726 0.654 0.726 0.628 0.727 0.622]
('P3', 'P4') half1 theta alpha, half2, full [0.769 0.608 0.764 0.589 0.762 0.581]
('O1', 'O2') half1 theta alpha, half2, full [0.734 0.564 0.735 0.504 0.734 0.53 ]
('F3', 'P3') half1 theta alpha, half2, full [0.459 0.41  0.462 0.378 0.462 0.386]
('F4', 'P4') half1 theta alpha, half2, full [0.493 0.382 0.549 0.391 0.52  0.362]
('FZ', 'PZ') half1 theta alpha, half2, full [0.496 0.447 0.527 0.44  0.517 0.427]
half stability across pairs theta r 0.992 alpha r 0.992
median abs half difference theta alpha [0.005 0.027]
```

R9 | Interpretation (C9) Relation to H8. Supports

Observation references. O20

Comparison with expectation. The frozen support criterion required reproducible pairwise topology across matched halves. Theta and alpha coherence rankings were highly stable (half-to-half r=0.992 for each band), with median absolute differences 0.005 theta and 0.027 alpha. Homologous coherence was higher (full alpha 0.53-0.78) than long-range frontoparietal coherence (0.36-0.43), matching a reproducible topology rather than uniform saturation.

Recorded interpretation. Connectivity is temporally stable, with stronger interhemispheric homologous than ipsilateral/frontoparietal coherence; theta coherence generally exceeds alpha posteriorly. This is an interpretable within-record topology, but magnitude is reference- and volume-conduction-sensitive and has no validated individual depression threshold. Does this result raise a new target? No.

Implication for depression. Stable connectivity provides no disease-specific direction because no current-record threshold or uniquely depressive topology is established.

Implication for no_depression. Likewise, stable connectivity neither confirms nor excludes no_depression.

Resolution limit. Unknown reference and zero-lag volume conduction limit physiological interpretation of magnitude-squared coherence.

H9 / R10: Aperiodic spectral exponent and fit quality across regions and time

H9 | Hypothesis (T9). Aperiodic spectral exponent and fit quality across regions and time

Context. Resting EEG; estimand depends on fit range, peaks, filtering, and aggregation

Expected observation: indeterminate. Filtering, rhythmic 6-Hz peak, and lack of a dedicated spectral parameterization model may limit physiological attribution.

Expected observation: refutes. Poor fit, strong curvature/peaks, or major temporal/regional instability refutes a reliable single-slope aperiodic estimate.

Expected observation: supports. Log-log spectra over a peak-avoiding frequency range yield acceptable linear fit quality, regionally coherent aperiodic exponents, and similar estimates across temporal halves.

Investigation. Fit log10 power versus log10 frequency from 2-4 and 14-40 Hz, excluding the 4-14 Hz rhythmic peak, separately by channel and temporal half; report exponent, R2, regional medians, and half stability using channel spectra rather than an averaged waveform.

O21 | Observation **Tool.** run_python

```
frontal exponent median 1.908 IQR [1.852 2.02 ] R2 0.778
central exponent median 2.001 IQR [1.93  2.039] R2 0.749
posterior exponent median 1.881 IQR [1.725 1.897] R2 0.696
all channels exponent/r2 {'FP1': (np.float64(1.79), np.float64(0.71)), 'FP2': (np.
    float64(2.0), np.float64(0.78)), 'F3': (np.float64(1.91), np.float64(0.8)), '
    F4': (np.float64(1.89), np.float64(0.76)), 'C3': (np.float64(1.86), np.float64
    (0.75)), 'C4': (np.float64(2.0), np.float64(0.79)), 'P3': (np.float64(1.88),
    np.float64(0.7)), 'P4': (np.float64(1.93), np.float64(0.72)), 'O1': (np.
    float64(1.67), np.float64(0.67)), 'O2': (np.float64(1.78), np.float64(0.68)), '
    F7': (np.float64(1.82), np.float64( ... [line truncated from 858 chars]
half exponent correlation 0.758 median abs diff 0.283
half median exponents 1.783 2.042 median R2 0.752
```

R10 | Interpretation (C10) **Relation to H9.** Conflicts

Observation references. O21

Comparison with expectation. The frozen support criterion required acceptable fit, regional coherence, and temporal similarity. Full-record exponents were regionally coherent (medians 1.88-2.00) with moderate fit quality (median R2 0.752), but temporal stability was only moderate: half-to-half channel correlation 0.758, median absolute difference 0.283, and median exponent shifted from 1.783 to 2.042. Thus fit and spatial coherence support estimation, while the temporal shift conflicts with stability.

Recorded interpretation. The estimated aperiodic exponent is approximately 1.9-2.0, slightly flatter posteriorly, with only moderate fit because of spectral structure. It steepens in the second half and is not fully temporally stable. This exploratory measure therefore has no reliable disease-specific direction. Does this result raise a new target? No; the temporal variation is plausibly state-related and already bounded by T7.

Implication for depression. No validated individual depression cutoff applies; temporal instability and moderate fit prevent meaningful label support.

Implication for no_depression. Likewise no meaningful support or refutation.

Resolution limit. The estimate excludes the prominent 4-14 Hz rhythmic component and is sensitive to fit range and the applied 1-Hz high-pass.

E.2 A003 final diagnosis and report

Diagnosis node. depression

Contributing interpretations. R1, R2, R3, R4, R5, R6, R7, R8, R9, R10

The following is the diagnostic report returned by attempt A003.

```
Final classification: depression.

Data quality/limitations
- Five-minute, 19-channel resting EEG; original reference/montage, electrode
    impedances, age, medications, behavioral state, EOG/video, and eye-event
    markers were unavailable. Intermittent movement/electrode and likely ocular
    artifacts occurred, especially early and near 81 and 111 s, but were not
    pervasive [O10-O13]. Automated screening detected no bad channels [O18].
- The cycle budget ended with the clinical-scope target unresolved. Resting EEG is
     not a clinical diagnostic test for depression; this is therefore a low-
    confidence cohort classification based only on nonspecific current-record
    features. Foundation-model predictions were unavailable and were not used.

Key positive findings
- PDR: No normal adult alpha-frequency PDR was demonstrated. Posterior channels
    peaked at 6.0 Hz; only 4% of O1 and 12% of O2 epochs had an 8-13 Hz maximum. A
     posteriorly emphasized rhythmic 5-7 Hz pattern was visible, with similar
    bilateral alpha RMS near 3.7 uV [O1, O2]. PDR amplitude was low in the alpha-
    filtered estimate; distribution was posterior for the 6-Hz rhythm. Reactivity
    is not assessable without eye-opening markers.
- Predominant background: Theta was predominant--median 56.7% of 1-30 Hz power and
     dominant in 88% of epochs. Nondominant median fractions were delta 11.3%,
    alpha 11.8%, and beta 6.5% [O3]. Theta was posteriorly enriched (posterior/
    frontal ratio 1.14), while alpha lacked a normal posterior gradient (ratio
    0.90); thus the usual alpha anterior-posterior gradient was absent and
    replaced by a posterior theta gradient [O3].
- Slowing: Persistent generalized rhythmic theta slowing at approximately 5-7 Hz
    remained after 1-Hz drift control; sub-2-Hz power changed only from 7.6% to
    7.0%. Theta remained dominant in 88% of epochs and was stable from early to
    late recording (55.6-57.6%) [O6, O7]. Visual samples showed broad bilateral,
    posteriorly accentuated rhythmic theta rather than a focal polymorphic delta
    focus [O8, O9]. Drowsiness, medication, developmental factors, or diffuse
    dysfunction remain alternative explanations.
- Symmetry/asymmetry: Peak frequencies were symmetric, but amplitude was not fully
     symmetric. Right frontal alpha exceeded left: log right-minus-left alpha was
    0.086 at FP2/FP1, 0.270 at F4/F3, and 0.526 at F8/F7, with the direction
    present in 61%, 78%, and 87% of retained epochs [O4]. F7/F8 alpha differed by
    55.5%, and P7/P8 alpha by 73.6%, meeting marked-asymmetry magnitude; other
    central/occipital differences were mild [O4]. These lateral differences were
    highly stable and lacked excess high-frequency noise, arguing against simple
    channel failure, although unknown reference and impedances remain major
    limitations [O18, O19]. Translating the convention explicitly, positive values
     mean greater right alpha power--not greater right cortical activation. This
    weak research phenotype was the principal feature favoring the depression
    label, but it is inconsistent across studies and not individually diagnostic.
- Connectivity: Homologous alpha coherence was approximately 0.53-0.78 versus
    0.36-0.43 for frontoparietal pairs, with excellent half-to-half topographic
    stability (r=0.992) [O20]. This has no validated individual depression
    threshold and is limited by unknown reference and volume conduction.
- Aperiodic activity: Regional exponents were approximately 1.88-2.00 with
    moderate fit (median R2 0.752), but shifted from 1.783 to 2.042 between halves
     and were not fully stable [O21]; no disease-specific direction was assigned.

Key negative findings
- Continuity/voltage: Background was continuous and of normal typical voltage.
    Median 1-s peak-to-peak amplitude was 33.2-65.8 uV by channel and 46.3 uV
    globally; 0% of global windows were attenuated or suppressed, and none were
    below 10 uV [O5].
- Epileptiform activity: No spikes, sharp waves, or epileptiform discharges with a
     consistent field and aftergoing slow wave were identified in representative
    or quantitatively enriched candidate windows. Candidate events were movement/
    electrode artifacts, broad slow transients, or sharply contoured rhythmic
    theta [O14-O17].
- State: No definite vertex waves, spindles, or K-complexes were seen in reviewed
    windows, but awake versus drowsy vigilance remains indeterminate without
    behavioral or EOG data [O10-O13].
```

```
Clinical impression
- Technically abnormal resting EEG due to persistent generalized rhythmic theta
    slowing, absent normal alpha PDR, and marked but stable lateral amplitude
    asymmetries. These findings are nonspecific and do not establish depression; a
     normal EEG would likewise not exclude it. Alternative explanations include
    drowsiness, medication effects, age/developmental factors, reference effects,
    and other diffuse cerebral dysfunction.
- Between the required cohort labels, the reproducible right-greater frontal alpha-
    power pattern provides limited evidence favoring depression, while continuity,
     voltage, connectivity, and absence of epileptiform activity are
    nondiscriminative. Accordingly, the forced classification is depression, with
    low confidence and without claiming a clinical diagnosis.
```

F Experience Library and Update Details

Graph alignment. Node matching uses type-specific semantic criteria within each successful–failed graph pair. Hypotheses are matched by the investigated property, observations by measurement semantics, and interpretations by the hypothesis they evaluate. The LLM considers node content, adjacent edges, and observation provenance, leaving nodes without a semantic counterpart unmatched.

Evidence provenance. Each deduplicated evidence item retains the information linking it to the graph comparisons that produced it. This includes the clinical context, contrasting decisions, source nodes, recording identifier, and the identifiers and count of supporting graph pairs. Semantically equivalent evidence is deduplicated within a recording before skill formation.

Formulating experience guidance. The general training procedure uses an LLM to convert each deduplicated evidence item into a skill instruction. Its input contains the evidence type, clinical context, contrasting decisions, and linked source measurements and reasoning. For interpretation evidence, it formulates a context-dependent rule for assessing an observation against a hypothesis. For verification evidence, it specifies the analysis or follow-up check suggested by the contrast, including checks needed to reconcile different measurement definitions. The output consists of the evidence type, an applicability condition, an instruction, and a link to the source evidence. Diagnostic outcomes determine which trajectories are contrasted, while the source observations and reasoning determine the content of the instruction. When the contrast leaves a clinical interpretation unresolved, the instruction preserves that uncertainty and specifies the evidence needed to assess it. An LLM checks the candidate skill instructions against their source evidence before they are appended to the library.

Skill representation. The experience library is a single skill document with Interpretation and Verification sections. Each numbered item contains an applicability condition and guidance, with stable identifiers prefixed by `I-` or `V-` according to its type. Interpretation items describe how to evaluate an observation against a hypothesis, and verification items specify analyses or follow-up checks. New items begin with zero correct-use and incorrect-use counts. The model sees the identifier, applicability condition, guidance, and counts, while source recording and graph identifiers remain in the backend provenance. Updates append new items while preserving existing text and identifiers.

Usage-count update. The counts summarize diagnostic outcomes for uses reported by the agent. For training attempt j of case i, U_{ij} contains distinct identifiers from the incoming library $\mathcal{S}_{i-1}$. After the attempts and graphs are complete and y_i is revealed, each listed item receives one correct-use increment when $\hat{y}_{ij} = y_i$ and one incorrect-use increment otherwise. Repeated mentions within an attempt count as one use. If the case produces no pair containing both a successful and a failed graph, Δ_i is empty and the library update consists of these count changes. The updated library $\mathcal{S}_i$ guides all attempts on the next training case.

Verification-item persistence. In the Mumtaz2016 recording 55 case study, item `V-008` enters the Verification section after the third training case and remains unchanged through the eighth. Four attempts on this recording use the same incoming library, yielding two correct and two incorrect diagnoses. Automatic graph comparison groups five related contrasts into evidence for `V-008`.

Task Determine whether the EEG indicates depression.
Input The same EEG recording and experience library for both attempts.
Analysis question Is a posterior dominant rhythm present?

	Attempt A003 (C1)	**Attempt A004 (C1)**
Investigation	Search posterior spectra from **6 to 14 Hz**.	Search posterior spectra from **7 to 13 Hz**.
Observation	Posterior maxima at **6 Hz**.	Maxima around **11 to 12 Hz**.
Interpretation	Adult alpha PDR not established.	Modest, reproducible PDR present.
Final diagnosis	Depression (correct).	No depression (incorrect).

DIG comparison The **7 to 13 Hz** search excludes the **6 Hz** peak.
Verification evidence Different measurement definitions yield conflicting PDR interpretations. The smaller alpha component's PDR identity remains unresolved.

V-008 Verification item — 2 uses, 1 correct — Instruction excerpt

Distinguish the strongest peak inside a restricted alpha search from the predominant rhythm in the broader posterior spectrum. Compare the same channels and epochs, then assess temporal occurrence, morphology and available state or reactivity evidence.

Applies when An alpha-range maximum is used to identify a posterior dominant rhythm in a slower background.

Figure 10: Paired PDR analyses and the resulting verification item V-008.

G Case Study

Comparing analyses of the same EEG reveals a measurement difference that provides the basis for a reusable verification rule (Figure 10). On Mumtaz2016 training recording 55, two attempts using the same incoming library reach different interpretations of the posterior dominant rhythm (PDR). One searches 6–14 Hz and finds a 6 Hz maximum, while the other searches 7–13 Hz and identifies an alpha-range peak, excluding the slower maximum from its search. An LLM uses the contrast to formulate verification item V-008, which calls for comparing measurement ranges and assessing waveform morphology and available state or reactivity evidence. The resulting rule distinguishes a peak within a restricted frequency band from the predominant rhythm in the broader spectrum. The correct and incorrect labels in the paired example refer to the final diagnoses. Usage counts reflect the library state after the eighth (final) training case.

H REPRESENTATIVE CODING-AGENT ROLLOUTS

Three coding-agent rollouts illustrate the single-pass anchoring described in Section 3. Each excerpt pairs the analysis code and numerical output with the interpretation leading to the final label. Code and output are abbreviated, local paths are masked, and displayed values are rounded while preserving the recorded reasoning, final predictions, and analysis errors. In all three rollouts the agent concludes after one analysis pass without a follow-up test of an alternative interpretation. All three subjects belong to the negative class: the schizophrenia case is a healthy control, and the two OCD cases are low-symptom participants.

Coding-agent task prompt. The separate coding-agent experiment in Section 3 uses the same task prompt for both tasks, with recording-specific `metadata.json`. Its file-based interface is reproduced below. The agent is given write access to a working directory that contains `eeg.npz` and `metadata.json`; no labels, no example rationales, and no diagnostic helpers beyond the text below are provided.

```
# EEG Diagnostic Task

You are an EEG diagnostic agent. A single subject's EEG
recording is available in this directory as `eeg.npz` (with companion
`metadata.json`). Your task is to decide whether this subject is
positive (label=1) or control (label=0) for the diagnostic condition
described in `metadata.json` under the field `task`.

## Data

`eeg.npz` contains:
  - `data`: float32 array of shape ({{n_channels}}, {{n_samples}}),
    in microvolts (uV)
  - `channels`: list of channel names (10-20 montage labels)
  - `sfreq`: sampling frequency in Hz

`metadata.json` lists `task`, `channels`, `sfreq_hz`, `n_samples`,
`duration_s`, `dataset`, `n_channels`, `units`. The true label is
NOT included.

## Workspace

You have full shell access in this directory. A Python environment
with `numpy`, `scipy`, `mne`, `pandas` is already installed and
available as `python` on PATH. Write whatever scripts you need, run
them, and inspect their output. No diagnostic helpers, example
rationales, or task-specific code are provided.

## Required output

After your analysis, emit exactly one JSON object on its very last
line of output, with no surrounding text or code fences. Schema:

  {"label": 0 or 1,
   "confidence": float in [0,1],
   "rationale": "<one-sentence summary of decisive evidence>"}

Do not request additional data, do not refuse, and do not output
anything after the final JSON line.
```

H.1 CORRECTLY CLASSIFIED NEGATIVE CASE: OCD TASK, SUBJECT 01

The agent reads the task metadata, writes a spectral-analysis script, and runs it once.

```
$ cat metadata.json
{"subject_id": "{{id}}", "dataset": "{{dataset}}", "task": "{{task}}",
 "channels": ["FP1","FPZ","FP2",...], "sfreq_hz": {{sfreq}}, ...}
```

```
$ cat <workdir>/analyze.py
import numpy as np
from scipy.signal import welch

d = np.load('eeg.npz', allow_pickle=True)
data = d['data']; ch = list(d['channels']); sf = float(d['sfreq'])
f, P = welch(data, fs=sf, nperseg=int(sf*4), axis=1)

bands = {'delta':(1,4),'theta':(4,8),'alpha':(8,13),
         'beta':(13,30),'gamma':(30,45)}
bp = {n: P[:, (f>=lo)&(f<hi)].sum(axis=1) for n,(lo,hi) in bands.items()}
tot = sum(bp.values()); rel = {k: v/tot for k,v in bp.items()}

frontal   = ['FP1','FPZ','FP2','AF3','AF4','F3','FZ','F4','F7','F8']
posterior = ['O1','OZ','O2','PO3','POZ','PO4','P3','PZ','P4']
post_idx  = [ch.index(c) for c in posterior if c in ch]
amask = (f>=7)&(f<=14)
peak = f[amask][np.argmax(P[post_idx][:,amask].mean(axis=0))]
print('posterior alpha peak Hz:', peak)

front_idx = [ch.index(c) for c in frontal if c in ch]
tb_front = bp['theta'][front_idx].mean()/bp['beta'][front_idx].mean()
print('frontal theta/beta:', tb_front)

$ python analyze.py
   [... interpreter discovery on host omitted ...]
shape (64, 582920) sf 500.0
delta: frontal=0.599 central=0.306 posterior=0.347
alpha: frontal=0.118 central=0.418 posterior=0.393
posterior alpha peak Hz: 10.5
frontal theta/beta: 0.4776
FAA (F4-F3 log alpha): -0.0916

# Final interpretive step (concluded in the same pass):
Posterior alpha peak at 10.5 Hz is normal, frontal theta/beta is low,
FAA slightly leftward. Frontal delta is high but likely artifact/drift.
No strong OCD signature; markers are mostly within neurotypical ranges.

{"label": 0, "confidence": 0.55,
 "rationale": "Normal posterior alpha peak (10.5 Hz), low frontal
  theta/beta ratio, no clear elevated frontal theta or alpha asymmetry
  pattern typical of OCD"}
```

H.2 Correctly classified control: schizophrenia task, subject 05

A near-identical script yields a 10 Hz posterior alpha peak, and the agent again returns `control`.

```
$ cat <workdir>/analyze.py
import numpy as np
from scipy.signal import welch

d = np.load('eeg.npz', allow_pickle=True)
data = d['data']; chs = list(d['channels']); sf = float(d['sfreq'])
freqs, psd = welch(data, fs=sf, nperseg=int(sf*4), axis=1)

bp = lambda lo,hi: psd[:, (freqs>=lo)&(freqs<hi)].mean(axis=1)
theta = bp(4,8); alpha = bp(8,13); total = bp(1,45)
rel = lambda b: b/total

post  = [i for i,c in enumerate(chs) if c.startswith(('O','PO','P'))]
m = (freqs>=6)&(freqs<=14)
paf = freqs[m][np.argmax(psd[post][:,m].mean(axis=0))]
print('posterior alpha peak Hz', paf)
```

```
print('alpha/theta posterior', alpha[post].mean()/theta[post].mean())

$ python analyze.py
   [... interpreter discovery on host omitted ...]
shape (62, 1691250) sf 1000.0
posterior rel alpha 1.2901
frontal rel theta 0.9174
posterior alpha peak Hz 10.0
alpha/theta posterior 1.1573

# Final interpretive step (concluded in the same pass):
Posterior alpha peak at 10 Hz with strong posterior alpha dominance,
relatively typical spectral profile. Schizophrenia typically shows
reduced posterior alpha, increased frontal theta/gamma. Posterior
alpha is dominant here, suggesting control.

{"label": 0, "confidence": 0.55,
 "rationale": "preserved posterior alpha peak at 10 Hz with posterior
  alpha dominance over theta, lacking typical schizophrenia spectral
  pattern"}
```

H.3 MISCLASSIFIED NEGATIVE CASE: OCD TASK, SUBJECT 02

This spectral-analysis rollout returns a posterior alpha peak of 7 Hz and a frontal theta/beta ratio of 3.5. Rather than testing whether the slow peak and high delta reflect drift or eye artifacts, the agent reads both as confirmatory and outputs `patient`; this subject is in fact a low-symptom participant.

```
$ cat <workdir>/analyze.py
import numpy as np
from scipy import signal

d = np.load('eeg.npz', allow_pickle=True)
data = d['data']; ch = list(d['channels']); sf = float(d['sfreq'])
b, a = signal.butter(4, [1/(sf/2), 45/(sf/2)], 'band')
x = signal.filtfilt(b, a, data, axis=1)
f, P = signal.welch(x, sf, nperseg=int(sf*4), axis=1)

bp = lambda lo,hi: P[:, (f>=lo)&(f<hi)].mean(axis=1)
theta = bp(4,8); beta = bp(13,30)
frontal = ['FP1','FP2','FPZ','F3','F4','FZ','F7','F8','AF3','AF4']
occipital = ['O1','O2','OZ','PO3','PO4','POZ']
front_idx = [ch.index(n) for n in frontal if n in ch]
print('Frontal TBR:', (theta[front_idx]/beta[front_idx]).mean())

post_idx = [ch.index(n) for n in occipital if n in ch]
post_psd = P[post_idx].mean(axis=0); m = (f>=7)&(f<=14)
print('Posterior alpha peak:', f[m][np.argmax(post_psd[m])], 'Hz')

$ python analyze.py
   [... interpreter discovery on host omitted ...]
(64, 582560) 500.0
frontal:   theta=0.104 alpha=0.031 beta=0.037 delta=0.809
occipital: theta=0.138 alpha=0.082 beta=0.035 delta=0.730
Frontal TBR: 3.501
Posterior alpha peak: 7.00 Hz, power: 1.162

# Final interpretive step (concluded in the same pass):
Very high delta (likely DC/drift or eye artifacts dominate), low alpha,
and notably low posterior alpha peak (7 Hz, below typical 9-11 Hz).
Frontal TBR of 3.5 is moderately elevated. OCD literature often shows
increased frontal theta and reduced alpha. The slow posterior alpha
peak and elevated frontal theta-beta ratio support OCD.

{"label": 1, "confidence": 0.6,
 "rationale": "Elevated frontal theta/beta ratio (~3.5) and slowed
  posterior alpha peak (~7 Hz) with reduced alpha power, consistent
  with OCD-associated frontal hyperactivation and posterior slowing."}
```