DOES IT REALLY TRANSFER? AUDITING FROZEN EEG FOUNDATION MODELS ON SMALL EAR-EEG DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained foundation models are often evaluated by freezing them and training a small classifier on a new dataset. When the dataset is small and differs from the pretraining data, a high score can come from choices tuned on the test subjects or from cues unrelated to the task rather than from the representation. We propose a protocol that tells whether such a model really transfers. It records every setup choice, retests each result by repeating the search that produced it on permuted labels and by nested cross-validation, checks whether failed channels or age and sex alone predict the label, and injects known signals to confirm that real effects can be found. We apply it to eight EEG foundation models pretrained on scalp recordings and to SYNAPSE, a wearable ear-EEG dataset of 23 subjects with a passive auditory paradigm and clinical labels, which we release. A standard evaluation finds two models that appear to transfer. The protocol shows that BIOT-18ch's score (0.86 AUROC) comes from the channels that failed quality control and falls to 0.46 without them, and that LABRAM's (0.92) does not rely on failed channels but cannot be confirmed: it holds only in a narrow range of input windows and cannot be separated from sex and age at this sample size. On the same recordings, the protocol confirms a transfer of known size injected into a random group of subjects and confirms nothing without it. We release the protocol, code, data and all results.

1 INTRODUCTION

Evaluating frozen foundation models on small datasets. A common way to use a pretrained foundation model on a new task is to keep its weights frozen, train a simple classifier (a probe) on a small labeled dataset, and read the probe's accuracy as evidence that the pretrained representation transfers. When the new data differ from the pretraining data and are few, this evidence is easy to misread. The pipeline involves choices that papers rarely report, such as how the data are fed to the model, which segment of the signal is used and which classifier sits on top, and these choices are often tuned on the same subjects that are used for testing. The data can also contain cues that predict the label for reasons unrelated to the task, such as recording quality or the age and sex of the subjects. Either can give a high score to a model that does not transfer.

From scalp to ear EEG. Electroencephalography (EEG) records brain activity non-invasively with millisecond resolution and is used to diagnose epilepsy, sleep disorders and neuropsychiatric conditions. The standard device is a scalp cap of 19 to 64 electrodes that needs gel and trained staff, and recent EEG foundation models (Jiang et al., 2024; Yang et al., 2023; Kostas et al., 2021; Wang et al., 2024; 2025a; Guetschel et al., 2024; Ma et al., 2026) are pretrained on large corpora of such recordings. Ear EEG records from a few electrodes around or inside the ear (Debener et al., 2015). It is discreet, supports long recordings in daily life and suits home monitoring, sleep and hearing-related applications (Mikkelsen et al., 2017; Holtze et al., 2022), but ear-EEG datasets rarely exceed a few dozen subjects, too few to train large models on a noisy, non-stationary and person-specific signal. Reusing foundation models pretrained on scalp EEG is therefore the natural route, and its outcome depends on four differences between the settings, each an instance of a general evaluation problem. (i) Ear EEG has about a dozen closely spaced, non-standard electrodes, and (ii) wearable devices often record below the pretraining sampling rate, so the input must be adapted to the pretrained interface,

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.

including channel or position embeddings that cannot be reused. (iii) Electrodes around the ear lose contact more easily, and channels that fail quality control can form a group-correlated pattern, an acquisition artifact that a model can use as a shortcut. (iv) Cohorts are small and each subject carries a single label, so an evaluation rests on a few dozen samples, and the input window, configuration and probe are often chosen by trying many options on these same subjects, which inflates the reported score. The original model papers report only the configuration that worked. We therefore ask: *how can we tell whether a frozen foundation model really transfers to a small dataset that differs from its pretraining data?*

Protocol and findings. Our key insight is that each way an apparent transfer can arise has a direct test: a selection can be repeated on permuted labels, a shortcut can be removed, and a test can be checked on a signal of known size (§3). We build an evaluation protocol on this insight. It treats channel adaptation, sampling rate, embedding replacement, input window, failed-channel treatment and probe as explicit choices, and tests each result by repeating the search that produced it on permuted labels and by nested cross-validation, which makes the choice without the test subject. It checks whether the label can be predicted from the pattern of failed channels or from age and sex alone, reruns each model on the channels that pass quality control in every subject, injects a known signal to confirm that real effects are detected, and estimates by simulation how strong an effect must be to be detected. We apply it to eight EEG foundation models on SYNAPSE, a 16-channel cEEGrid ear-EEG cohort (n=23) recorded during passive auditory stimulation for sound-sensitivity-disorder screening (Baguley, 2003; Schecklmann et al., 2012), which we release as a testbed. A standard evaluation that reports each model's best score finds two models that appear to transfer, LABRAM at 0.92 AUROC and BIOT-18ch at 0.86. The protocol shows that BIOT-18ch's score comes from the channels that failed quality control: on the channels that pass in every subject it falls to 0.46. For LABRAM the protocol cannot reach a verdict. LABRAM does not rely on failed channels and stays significant after accounting for the choice among its 16 configurations, but it holds only in a narrow range of input windows, the stricter nested estimate is not significant, and because the control group is younger and all female, its score cannot be separated from sex and age (§5). To show that the protocol can also confirm a real transfer, we add a known signal to the recordings of a random group of subjects: the protocol confirms this transfer and confirms nothing when the signal is absent. A simulation shows that with 23 subjects only strong transfer can be confirmed. These failure modes are not specific to ear EEG. They arise wherever frozen models meet small, imperfectly recorded datasets, and §6 draws the general lessons.

Contributions. (1) A **protocol** that tells whether a frozen foundation model really transfers to a small, unusual dataset. It records every setup choice, retests each result by repeating its search on permuted labels, checks whether failed channels or age and sex alone predict the label, and injects known signals to confirm that real effects are found. On real recordings it exposes a model whose score comes from failed channels and confirms an injected transfer of known size (§5). (2) Three **lessons** that apply beyond ear EEG: (a) interpolating failed channels is not enough, because a model can still tell which channels failed, so models should be compared on channels that pass in every subject; (b) the largest source of inflated scores is trying many input windows and keeping the best; (c) a setting that matters on a small clinical task may not change what the model can read, since resampling moves the clinical score by up to 0.54 but detection of a known injected signal by at most 0.04 (§5, §6). (3) **Synapse**, the first public wearable ear-EEG dataset with passive auditory stimuli and clinical labels, released with all results and code (§4).

2 RELATED WORK

Evaluating whether pretrained models transfer. Three lines of work evaluate transfer, and none tells whether a frozen model transfers to a small, unusual dataset. EEG foundation models (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024; 2025a; Guetschel et al., 2024; Ma et al., 2026) and their benchmarks (Xiong et al., 2026; Lahiri et al., 2026; Aristimunha et al., 2025) fine-tune or probe on downstream datasets and report the best configuration. They show that preprocessing, classifier heads and protocol can reverse model rankings, but they fix the input adaptation (channel mapping, resampling, embedding replacement, input window) instead of testing it, and when a label is defined per subject they often count segments of one subject as independent samples (Appendix K). Evaluation audits in machine learning show that conclusions depend on

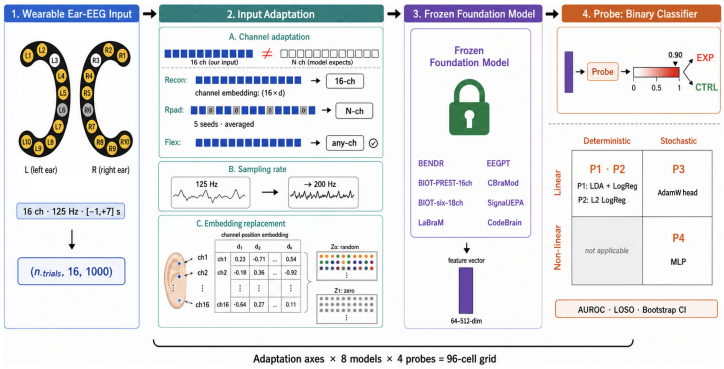

Figure 1: Overview of the protocol. **Stage 1**: wearable ear-EEG input (16-channel cEEGrid, 125 Hz). **Stage 2**: input adaptation, with channel adaptation (Recon, Rpad or Flex, fixed by architecture), sampling rate (S0, S1) and embedding replacement (Z0, Z1, and the control Zp in Appendix B), together with the input window and failed-channel treatment. **Stage 3**: eight frozen foundation models. **Stage 4**: four probes (P1–P4) and subject-level AUROC under LOSO. The resulting grid is then tested for selection, shortcuts and detectability (§3).

evaluation design (Liang et al., 2023; Bender & Friedman, 2018), that leakage (Kapoor & Narayanan, 2023) and model selection on the test data (Cawley & Talbot, 2010) inflate estimates, and that models exploit shortcuts, including acquisition artifacts in clinical data (Geirhos et al., 2020; DeGrave et al., 2021), but they diagnose these problems where test sets are large enough to measure them directly. Audits of clinical EEG foundation models test with permutation controls whether frozen features encode subject identity (Lin et al., 2026a) or dataset identity (Zare, 2026). They do not test failed-channel or demographic shortcuts or selection over configurations and windows, and they do not inject a known signal to show that a negative result is informative. With a few dozen subjects each of these gaps can decide the verdict, and our protocol addresses them together.

EEG and ear-EEG datasets. EEG foundation models pretrain on large scalp corpora such as the Temple University Hospital EEG corpus (Obeid & Picone, 2016) and are evaluated on public scalp datasets recorded with 19 or more electrodes. Ear EEG records from electrodes in the ear canal (Looney et al., 2012) or around the ear (Debener et al., 2015; Bleichner & Debener, 2017), and ear-EEG studies address auditory attention (Mirkovic et al., 2016; Holtze et al., 2022) and sleep staging (Mikkelsen et al., 2017) in healthy participants. We found no public ear-EEG dataset that combines a passive auditory paradigm with clinical labels (§4), and SYNAPSE fills this gap.

3 PROBLEM AND PROTOCOL DESIGN

Setting. A frozen foundation model maps a segment of EEG to a feature vector, and the features are used to separate two groups of subjects, such as patients and controls. The evaluator chooses how channels are adapted, whether the signal is resampled, what replaces embeddings that cannot be reused, which segment of the signal is used, how failed channels are handled and which probe is trained, and computes a subject-level AUROC for each choice by leave-one-subject-out cross-validation (LOSO). A transfer claim holds when the features separate the groups in new subjects and the separation comes from the EEG rather than from selection, acquisition artifacts or demographic differences. Four difficulties stand in the way of testing this on small out-of-distribution data, and Figure 1 shows where each check enters the protocol.

Difficulty 1: many choices, and the result depends on them. Prior work fixes one choice or reports the best one. On SYNAPSE, sliding the input window alone moves one model's AUROC between 0.30 and 0.96. *Design:* every choice is an explicit axis, and the whole grid is run and reported, not only its best cell.

Difficulty 2: choosing on the test subjects inflates the score, more so with fewer subjects. Reported scores, confidence intervals and per-cell permutation tests ignore that the score was selected.

Table 1: SYNAPSE against the pretraining data of the audited models. Channel count is out of distribution for six of the eight models, the other axes for all eight.

Dimension	Foundation-model pretraining	SYNAPSE
Channel count	16–64	**16**
Sampling rate	128–256 Hz	**125 Hz** (below all pretrain f_s)
Channel positions	Standard 10-20 scalp	**Non-standard ear**
Sample size	10^2–10^3	n=23
Cohort	Healthy / single-disorder	**Sound-sensitivity disorder**

In our positive example, a label made of pure noise still reaches 0.90 as the best of 88 settings. *Design:* a result is tested by repeating the search that produced it on permuted labels (a max-statistic permutation test) and estimated by nested cross-validation, which makes the choice within the training subjects only. The test repeats exactly the choices that were made: the configuration and probe when the window was fixed, and the window as well when it was chosen.

Difficulty 3: cues that predict the label but have nothing to do with the brain. In a small cohort the groups easily differ in recording conditions and demographics. On SYNAPSE, controls have more failed channels and are all young women. A model can use such cues directly, and common fixes such as interpolating failed channels do not remove them. Prior work reports demographics or interpolates failed channels. *Design:* baselines test whether the cues alone predict the label (the failed-channel pattern alone, sex and age alone), and controls rerun the model with the cue removed (only channels that pass quality control in every subject, retraining within demographic subsets). A result that disappears once a cue is removed comes from a shortcut.

Difficulty 4: a failed test has no single explanation. The model may not transfer, the pipeline may be unable to read the signal, or the sample may be too small to show an effect. *Design:* a positive control injects a known signal to check that the pipeline reads it, a positive example injects a transfer of known size into a random group of subjects to check that the protocol can confirm it, and a simulation estimates how strong an effect must be to be detected at the given sample size.

Verdicts. A transfer claim is *confirmed* when it passes the permutation test that repeats its search, its nested estimate is significant, and it survives the removal of every shortcut. It is *rejected* when removing a shortcut makes it disappear, and *undetermined* otherwise.

4 SYNAPSE: A DOCUMENTED VALIDATION TESTBED

Why this dataset. Sound-sensitivity disorders, including hyperacusis, tinnitus and misophonia, are assessed mainly with questionnaires and loudness tests that depend on the patient's report (Baguley, 2003; Schecklmann et al., 2012). A passive paradigm, in which subjects only listen, needs no task performance, and ear EEG could make such recordings routine outside the clinic. SYNAPSE records this setting and differs from the pretraining data on several axes at once (Table 1).

Acquisition. Sixteen cEEGrid channels are recorded at 125 Hz, 8 around each ear (L01–L10 and R01–R10 without position 3, excluded because hair prevented reliable skin contact, and position 6, used for reference and ground). Each subject completes three auditory tasks, LET (listening effort, speech-in-noise), HLT (hearing loudness test) and AST (aversive sound test), plus a visual task (PMT, 4–5 trials) that we release but do not analyze. Epochs span -1 s to $+7$ s around stimulus onset (1000 samples, setup and timing in Appendix A).

Cohort. After per-ear quality control (each ear must have $\geq 5/8$ channels passing impedance, variance, and correlation gates), which excludes 17 of 40 recorded subjects (11 with all 16 channels flagged), we retain n=23 subjects: 16 EXP (sound-sensitivity disorder: hyperacusis, tinnitus, misophonia, or comorbid) and 7 CTRL. EXP averages 41.7 y (11M/5F) and CTRL 25.3 y (all female). Used as scores, sex alone gives 0.84 AUROC, age 0.74, and the rule "male or older than 35" 0.91, the level a demographic shortcut can reach (Lesson 4). Retained subjects still have failed channels (3.0

per CTRL and 1.9 per EXP subject). The task is subject-level EXP vs. CTRL classification, with trial features averaged within each auditory task and then across tasks.

Release. We release the signals with channel masks, labels, sex and 5-year age bins, the same trials on $[-5, +7]$ s, per-cell predictions for both channel treatments and all audit analyses, and reproduction scripts, with a datasheet (Appendix D) (Gebru et al., 2021) and Croissant metadata validated with `croissant-checker` (Akhtar et al., 2024), under CC-BY-4.0 for data and MIT for code at an anonymized repository (`https://huggingface.co/datasets/synapse-ear-eeg/synapse`). Preprocessing is a 1–50 Hz FIR bandpass and a 60 Hz notch. SYNAPSE is the first public wearable ear-EEG cohort with a passive-stimulus paradigm and clinical labels.[1] It is a testbed for evaluation methodology, not clinical screening evidence at n=23.

5 EXPERIMENTS

Setup. We audit eight frozen EEG foundation models, BENDR, BIOT-16ch, BIOT-18ch, LABRAM, EEGPT, CBRAMOD, SIGNALJEPA and CODEBRAIN, pretrained on 16 to 64 channels at 128 to 256 Hz (Table 5 in Appendix B). Channels are adapted as each architecture requires: five models are rebuilt for 16 channels (Recon), BIOT-18ch and BENDR place them in random input slots over five seeds (Rpad), and SIGNALJEPA takes any channel count (Flex). The signal is fed at 125 Hz (S0) or resampled to the pretraining rate (S1). Embeddings that the loaders replace are set randomly (Z0), to zero (Z1) or, as a control, back to their pretrained values where reusable (Zp). Four probes sit on the frozen features: LDA followed by logistic regression (P1), L2 logistic regression (P2), a linear head trained by AdamW (P3) and a one-hidden-layer MLP (P4), the last two over five seeds. Each model reads 4 s (5 s for BIOT) from -1 s, a window fixed after a smoke test on LABRAM that compared it with a post-onset window, and a sweep slides the window over $[-5, +7]$ s in 0.2 s steps. Failed channels are kept, replaced by the mean of the passing channels of the same ear (interpolated), or excluded by keeping the three channels that pass quality control in every subject. Crossing sampling rates, embeddings and probes gives 96 cells per channel treatment, scored by LOSO subject-level AUROC. Selection-aware tests use two candidate sets: set A, a model's grid cells at the grid window with all four probes, and set B, the sweep windows with the other grid configurations under P1 and P2. Appendix B gives loaders, input lengths, hyperparameters and the other statistics.

What the grid reports. With failed channels kept, the best cells report LABRAM at 0.921 and BIOT-18ch at 0.861, statistically tied (p=0.155), and five models fall below a per-channel band-power baseline (0.723, Table 2). This best-cell reading lacks checks that small-cohort evaluations of EEG foundation models also rarely apply (Appendix K).

BIOT-18ch is rejected. Its score comes from the failed channels. With them interpolated it falls to 0.696, and on the three channels that pass quality control in every subject to 0.455, never exceeding 0.55 at any window position (Table 3, Lesson 1). Its nested estimate over windows (0.75) is higher than LABRAM's (0.69), but on the three channels it stays at or below 0.55 at every window, so the higher estimate reflects the failed channels.

LaBraM is undetermined. It does not rely on failed channels: at its S1/Z1/P2 cell it keeps 0.875 on the three channels, the same as with interpolation, and its best interpolated cell (P3) reaches 0.954, above every other model (Appendix C). Whether 0.954 survives selection depends on how much of the choice is repeated (Table 3). Repeating the choice among its 16 configuration-probe cells at the grid window, it passes a max-statistic test (p=0.004), but nested LOSO gives 0.84 (p=0.07). Repeating also the smoke-test choice between two windows, it still passes (p=0.009, nested 0.84, p=0.11). Choosing among all 41 sweep windows with P1 and P2 (set B), the best candidate reaches 0.955 with a selection-aware p of 0.051, and the nested estimate is 0.69 (p=0.19, Lesson 2). No nested estimate reaches significance. Because the control group is younger and all female, a rule on sex and age alone reaches 0.906, and retraining on females only gives 0.80 (p=0.08, Lesson 4). The protocol therefore cannot decide on LABRAM.

[1] We searched OpenNeuro, PhysioNet, Zenodo, and IEEE DataPort for ear-EEG datasets combining passive auditory stimuli and clinical disorder labels and found no matching dataset as of April 2026.

Table 2: Best-configuration AUROC per model with failed channels kept and interpolated, and shortcut baselines. Best config and CI refer to the interpolated grid. CI: 95% bootstrap over subjects.

Model	Best config (interp.)	Kept	Interp.	95% CI	Cells
Pretrained foundation models (frozen + probe)					
LABRAM-base	Recon / S1 / Z1 / P3	**0.921**	**0.954**	[0.848, 1.000]	16/16
SIGNALJEPA-w/o-chans	Flex / S1 / – / P3	0.714	0.725	[0.515, 0.923]	8/8
BIOT-six-18ch	Rpad / S1 / Z1 / P4	0.861	0.696	[0.447, 0.902]	16/16
CODEBRAIN	Recon / S0 / – / P2	0.759	0.634	[0.384, 0.863]	8/8
BIOT-PREST-16ch	Recon / S0 / Z1 / P1	0.645	0.634	[0.389, 0.857]	16/16
CBRAMOD	Recon / S1 / Z0 / P3	0.607	0.562	[0.300, 0.800]	8/16
EEGPT	Recon / S1 / Z0 / P3	0.589	0.555	[0.288, 0.812]	16/16
BENDR	Rpad / S1 / – / P1	0.589	0.504	[0.286, 0.882]	8/8
Supervised baselines (end-to-end, 3 seeds) and spectral baseline					
SHALLOWFBCSPNET	—	0.467	0.533	—	—
EEGCONFORMER	—	0.503	0.481	—	—
EEGNET	—	0.452	0.464	—	—
Band power + LogReg	$\delta/\theta/\alpha/\beta$	0.723[‡]	0.527[‡]	—	—
Shortcut baselines (no EEG)					
QC mask alone (LOSO LogReg)	—	0.759	—	—	—
Sex / age / male or >35 y	score	0.844 / 0.737[§] / 0.906		—	—

[‡]Per channel with failed channels kept, per ear over passing channels when interpolated, on the 4 s grid window.
[§]Exact age. The released 5-year bins give 0.705 and reproduce the other two values.

Table 3: The protocol applied to the three highest-scoring models, failed channels interpolated unless stated. Set A: the grid cells at the grid window with all four probes (16 for LABRAM and BIOT-18ch, 8 for SIGNALJEPA), which produced the grid best. Set B: 41 or 36 sweep windows and the other grid configurations, each with P1 and P2. Selection-aware p repeats the choice on permuted labels. For reference, the QC mask alone gives 0.76 and the rule "male or older than 35" 0.91.

Check	LABRAM	SIGNALJEPA	BIOT-18ch
Grid			
Grid best, failed channels kept	0.921	0.714	0.861
Grid best, interpolated (set A)	0.954	0.725	0.696
Per-cell permutation p of the grid best	0.002	0.09	0.10
Shortcut controls			
Three channels passing in every subject (P2)	0.875	0.482	0.455
Retrained on females only, 5 vs. 7, S1/Z1/P2 (p)	0.80 (0.08)	0.23	0.54
Selection			
Set A: selection-aware p of the grid best	0.004	0.31	0.37
Set A: nested LOSO (p)	0.84 (0.07)	0.64 (0.30)	0.63 (0.21)
Set A at both smoke-test windows: p, nested (p)	0.009, 0.84 (0.11)	—	—
Set B: best candidate, selection-aware p	0.955, 0.051	0.893, 0.44	0.750, 0.32
Set B: nested LOSO (p)	0.69 (0.19)	0.52 (0.45)	0.75 (0.10)

No other model passes any test. With failed channels interpolated, SIGNALJEPA and BIOT-18ch give $p \geq 0.31$ over their grid cells, no other best cell reaches $p<0.05$ even per cell (smallest $p=0.09$), and the supervised baselines stay at chance.

Can the frozen models read a stimulus-locked signal? Telling the post-onset from the pre-onset window of the same trial across held-out subjects gives 0.50–0.54 for every model and for band power, and the grand-average evoked response is weak, its largest across-subject t statistic lying at the 95th percentile of pre-stimulus fluctuations (Appendix N). When a synthetic evoked response is added to the recorded trials, seven models detect it at 0.71–0.91 for an N1 peak of twice the median trial standard deviation, whereas BENDR, whose features vary almost entirely along one direction, barely registers it (0.54, Figure 2A). The pipelines can read such a signal, but the recorded one is too weak to detect trial by trial.

Table 4: The protocol confirms a transfer of known size: LABRAM audited against a pseudo-label that carries an injected evoked response of the given amplitude (median trial standard deviations). Entries are AUROC with permutation p in parentheses. Set A and set B report the best candidate with its selection-aware p, and nested LOSO.

Amp.	Grid cell	Three channels	Set A best	Set A nested	Set B best	Set B nested
0	0.59 (0.23)	0.61 (0.21)	0.67 (0.49)	0.32 (0.63)	0.90 (0.21)	0.83 (0.13)
0.5	0.75 (0.046)	0.72 (0.09)	0.93 (0.016)	0.79 (0.10)	1.00 (0.006)	0.87 (0.09)
1	0.89 (0.005)	0.88 (0.014)	1.00 (0.002)	0.96 (0.020)	1.00 (0.005)	0.93 (0.050)
2	0.98 (0.002)	0.98 (0.001)	1.00 (0.006)	0.99 (0.020)	1.00 (0.006)	0.97 (0.020)

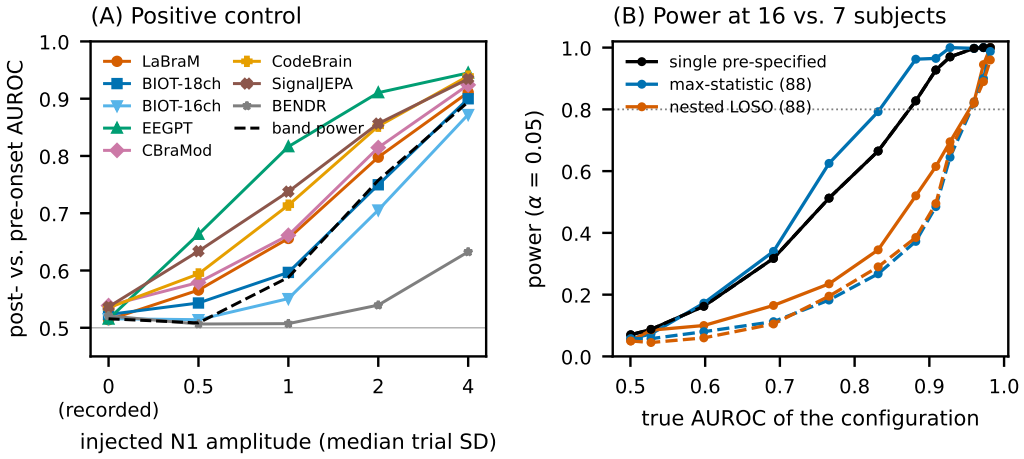

Figure 2: **(A)** Positive control: trial-level AUROC for telling the post-onset from the pre-onset window across held-out subjects, as a synthetic evoked response of increasing amplitude (in median trial standard deviations) is added to the recorded trials. Amplitude 0 is the recorded data, and BENDR uses slot seed 0. **(B)** Simulated power at 16 vs. 7 subjects against the true AUROC of a P2 probe trained on 22 subjects, for a single pre-specified test and for selection over 88 correlated candidates by a max-statistic test or nested LOSO. Solid: every candidate carries the effect. Dashed: only 8 adjacent candidates do, and the single test, pre-specified inside them, is unchanged.

Can the protocol confirm a real transfer? Sixteen subjects receive a pseudo-label chosen to be uncorrelated with group, sex, age and each QC-mask channel, a synthetic evoked response of known size is added to their trials only, and LABRAM goes through the full protocol against this label (Table 4, details in Appendix N). Without injection no test passes, although the best of 88 window candidates reaches 0.90 and nested LOSO 0.83, which shows how far selection alone lifts AUROC at this sample size. At half a median trial standard deviation the max-statistic tests pass and nested LOSO does not, the pattern LABRAM shows on the clinical label. From one standard deviation every test passes, and the protocol confirms the transfer.

How strong must a transfer be to be confirmed? At 16 vs. 7 subjects, a single pre-specified test reaches 80% power at a true AUROC of 0.87, a max-statistic test over 88 correlated candidates at 0.83 when every candidate carries the effect and at 0.96 when only 8 adjacent ones do, and nested selection at 0.95–0.96 (Figure 2B). An undetermined verdict at this sample size is therefore expected unless transfer is strong.

6 LESSONS FOR EVALUATING FROZEN MODELS

The first three lessons are the contributions of §1, and the last two concern demographics and single choices. Each states a convention, what SYNAPSE shows, and what to do instead.

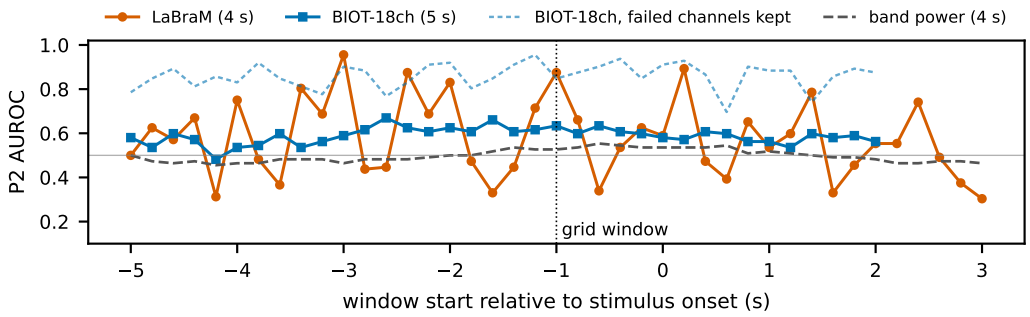

Figure 3: P2 AUROC as the input window slides in 0.2 s steps, failed channels interpolated and other configuration choices fixed. The dashed blue curve keeps failed channels. Band power is the per-ear spectral baseline of Table 2 on the same windows. Dotted line: the grid window.

Lesson 1: interpolating failed channels does not remove their shortcut. **Convention:** release raw signals with a QC mask, keep or interpolate failed channels, and expect a frozen model to treat them as noise. **Finding:** when failures correlate with the label, the model reads them. On SYNAPSE the 16-dimensional mask alone predicts the label at 0.76 (p=0.015), because two channels fail in 4 of 7 controls and 1 of 16 patients. Kept failed channels lift BIOT-18ch to 0.86 and above 0.77 at 33 of 36 window positions (Figure 3, dashed) and create a spurious gap of 0.21 over its channel-count-matched sibling BIOT-16ch (p=0.048, 0.05 after interpolation). Interpolated channels duplicate the ear mean, so their positions stay recoverable, and only the control restricted to channels that pass in every subject removes the shortcut (§5, Appendix I). *Check that the QC mask cannot predict the label, and compare models on channels that pass quality control in every subject.*

Lesson 2: the input window is the main source of inflated scores. **Convention:** fix one input window per dataset, treat it as preprocessing, and report the best configuration, as in the downstream protocols of LABRAM (Jiang et al., 2024), EEGPT (Wang et al., 2024) and CBRAMOD (Wang et al., 2025a). **Finding:** the window moves results more than any other choice. As it slides in 0.2 s steps, LABRAM ranges from 0.30 to 0.96, changing by a median of 0.17 per step, and every model averages 0.44 to 0.62 over positions (Figure 3, Table 9). Selection over configurations and probes at a fixed window leaves LABRAM's grid best significant under a max-statistic test, whereas selection over windows makes the same result borderline (p=0.051, Monte Carlo standard error 0.007) and lowers its nested estimate to 0.69, and selection alone lifts a noise label to 0.90 across 88 window candidates (§5, Appendix J). *Pre-specify the window, or choose it by nested cross-validation and test the result by repeating the choice on permuted labels.*

Lesson 3: a setting that matters on the clinical label may not change what the model reads. **Convention:** judge an adaptation choice by its effect on the target score. Resampling to the pretraining rate is the default in the audited codebases, none of which ablates it (Jiang et al., 2024; Yang et al., 2023; Wang et al., 2024; 2025a; Ma et al., 2026). **Finding:** on the clinical label, resampling changes P2 AUROC by between -0.14 (CODEBRAIN) and $+0.54$ (LABRAM) (Table 7 in Appendix E), and with time content matched the directions persist (Table 11). On the injected signal, where 1760 trials resolve small differences, the same choice changes detection by -0.04 to $+0.02$, and LABRAM's by $+0.002$ (Appendix N). Embedding choices behave the same way: Z1 gains up to $+0.17$ over Z0 on the clinical label and changes detection of the injected signal by less than 0.02. *Judge adaptation choices on an endpoint with a known signal and enough samples, not only on the small target label.*

Lesson 4: at small n, demographics bound what a dataset can show. **Convention:** attribute group separation to the disorder once demographics are reported. **Finding:** on SYNAPSE the controls are all female and younger than the patients, and a rule on sex and age alone reaches 0.91, inside LABRAM's CI. Retrained within demographic subsets, LABRAM reaches 0.80 on females only (p=0.08) and 0.73 on an age-matched subcohort (p=0.12), and decoding sex or age within the patient group has too little power at n=16 to separate the explanations (Appendices F and J). *Report the AUROC of demographic variables next to every model, and retrain within matched subsets.*

Lesson 5: a single embedding or probe choice can mislead. **Convention:** replace embeddings without testing the alternative and report one linear probe, as in self-supervised vision (Chen et al., 2020; He et al., 2020). **Finding:** Z1 is at least as good as Z0 at each model's better sampling rate with P2, yet Z0 is higher in 14 of 32 paired cells. Keeping the reusable pretrained embeddings (Zp) changes no model's standing (Appendix M), and at S1/Z1 the probe alone moves AUROC by up to 0.14 (Appendix L). Pooled LOSO AUROC is also biased low at small n (Airola et al., 2011), and leave-pair-out AUROC raises most models' cells by 0.01 to 0.09 (Table 12). *Report every embedding option and probe, and add leave-pair-out AUROC or permutation tests at small n.*

7 Discussion and Limitations

Implications. The undetermined verdict on LaBraM states what data would settle it: controls matched in sex and age, a window and configuration fixed in advance, and enough subjects to detect a moderate effect. LaBraM passes the max-statistic tests but not nested selection, the pattern of the positive example with a weak injected transfer and not of the one without. Its result is consistent with a real but weak group difference whose source, disorder or demographics, this cohort cannot resolve, and it merits validation on such a cohort. The checks are not designed around EEG. Selection, acquisition artifacts and demographic differences can inflate scores wherever frozen models meet small, imperfectly recorded datasets with many configuration choices, and the same controls and injection tests apply there.

Demographics. Controls are all female and younger than patients (Lesson 4), so absolute AUROC values on Synapse can partly reflect demographics, which the subset analyses address only in part.

Frozen features only. We evaluate frozen backbones to isolate representation quality, since fine-tuning would conflate it with optimization quality. End-to-end training stays at chance on Synapse (Table 2), and adapter-tuning (LoRA) is outside scope.

Post hoc findings. Lessons 1 and 2 arose during the audit rather than from a pre-registered plan. We test Lesson 1 with channel controls and Lesson 2 with selection-aware tests and the positive example, and the remaining p values are uncorrected.

One dataset and small n. Synapse shifts five axes at once, so a failure cannot be attributed to a single axis, and channel adaptation is not varied within a model. A second dataset, such as ear EEG from another cohort or scalp EEG reduced to 16 channels at 125 Hz, is needed to separate the axes. With n=23, CIs are wide and pooled LOSO AUROC is biased low, which leave-pair-out AUROC partly addresses.

Model loading. The verdicts concern these public models loaded as described in §5 and Appendix B, not whether any model works on wearable EEG. Keeping the reusable pretrained embeddings changes none of the grid-window results (Appendix M), but other mappings, such as giving each ear electrode the embedding of the nearest scalp electrode, were not tested.

8 Conclusion

We propose a protocol that tells whether a frozen foundation model really transfers to a small, unusual dataset. It records every setup choice, retests each result by repeating the search that produced it on permuted labels, checks whether failed channels or age and sex alone predict the label, and injects known signals to confirm that real effects can be found. Applied to eight EEG foundation models on the ear-EEG dataset Synapse, it shows that of the two models a standard evaluation reports as transferring, one reads the channels that failed quality control and the other cannot be confirmed at this sample size, while it confirms an injected transfer of known size on the same recordings. Evaluations of frozen models on small datasets should report the whole configuration grid, test each result against the search that produced it, and rule out shortcuts before claiming transfer.

AI USE STATEMENT

We used a generative AI assistant to help with coding, experiments, writing and figures. The authors verified all results and take full responsibility for the content.

ETHICS STATEMENT

All recordings were collected under IRB approval with informed consent, which covers the release of de-identified EEG, group labels, age and sex (Appendix D). Subject identifiers are pseudonymized study codes, free-text demographic fields are excluded, and age is released in 5-year bins to limit re-identification. Because the labels encode a clinical condition, the release is intended for methodology research, not for clinical screening or diagnosis. The age and sex imbalance of the cohort (§7) affects absolute AUROC values, and Appendix F reports an age-matched reanalysis.

REPRODUCIBILITY STATEMENT

The dataset (the analyzed $[-1, +7]$ s epochs and the same trials on $[-5, +7]$ s), the per-cell LOSO prediction files for all 96 grid cells under both channel treatments, the supervised baselines, the window sweep, the channel controls, the permutation, nested and leave-pair-out analyses, the embedding control, the positive control and the power simulation, and the scripts that regenerate every table and figure are available at an anonymized repository (`https://huggingface.co/datasets/synapse-ear-eeg/synapse`). Section 5 and Appendices B and J specify the loader, input adaptation, input window, probe and selection-aware tests for each model, Appendix D documents collection and preprocessing, and stochastic cells report five seeds.

REFERENCES

Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis*, 55(4):1828–1844, 2011.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, et al. Croissant: A metadata format for ML-ready datasets. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 82133–82148, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/9547b09b722f2948ff3ddb5d86002bc0-Abstract-Datasets_and_Benchmarks_Track.html`.

Bruno Aristimunha, Dung Truong, Pierre Guetschel, Seyed Yahya Shirazi, Isabelle Guyon, Alexandre R Franco, Michael P Milham, Aviv Dotan, Scott Makeig, Alexandre Gramfort, Jean-Rémi King, Marie-Constance Corsi, et al. EEG foundation challenge: From cross-task to cross-subject EEG decoding. *arXiv preprint arXiv:2506.19141*, 2025.

David M. Baguley. Hyperacusis. *Journal of the Royal Society of Medicine*, 96(12):582–585, 2003. URL `https://doi.org/10.1177/014107680309601203`.

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Martin G. Bleichner and Stefan Debener. Concealed, unobtrusive ear-centered EEG acquisition: cEEGrids for transparent EEG. *Frontiers in Human Neuroscience*, 11:163, 2017. doi: 10.3389/fnhum.2017.00163.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Stefan Debener, Reiner Emkes, Maarten De Vos, and Martin Bleichner. Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Scientific reports*, 5(1): 16743, 2015.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2021.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. NeurIPT: Foundation model for neural interfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.16548.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention. In *Proceedings of the 9th Graz Brain-Computer Interface Conference*, pp. 11–16. Verlag der Technischen Universität Graz, 2024. URL `https://doi.org/10.3217/978-3-99161-014-4-003`.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Björn Holtze, Marc Rosenkranz, Manuela Jaeger, Stefan Debener, and Bojana Mirkovic. Ear-EEG measures of auditory attention to continuous speech. *Frontiers in Neuroscience*, 16:869426, 2022. doi: 10.3389/fnins.2022.869426.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, and Roger Wattenhofer. EEG-Bench: A benchmark for EEG foundation models in clinical applications. In *NeurIPS Workshop on Foundation Models for the Brain and Body*, 2025. arXiv:2512.08959.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Tiantian Feng, and Shrikanth Narayanan. A multi-dimensional framework for evaluating generalization in eeg foundation models. *arXiv preprint arXiv:2605.28563*, 2026.

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.

Jeet Bandhu Lahiri, Parshva Runwal, Arvasu Kulkarni, Mahir Jain, Aditya Ray Mishra, Siddharth Panwar, and Sandeep Singh. PRISM: Exploring heterogeneous pretrained EEG foundation model transfer to clinical differential diagnosis. *arXiv preprint arXiv:2603.02268*, 2026.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Jun-You Lin, Ying Choon Wu, and Tzyy-Ping Jung. The identity trap in EEG foundation models: A diagnostic audit. *arXiv preprint arXiv:2606.06647*, 2026a.

Maggie Lin, Chung-Lin Hou, and Tzyy-Ping Jung. Leveraging a foundation model for the EEG-based diagnosis of Alzheimer's disease. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2026b. arXiv:2608.27719.

David Looney, Preben Kidmose, Cheolsoo Park, Michael Ungstrup, Mike Lind Rank, Karin Rosenkranz, and Danilo Mandic. The in-the-ear recording concept: User-centered and wearable brain monitoring. *IEEE Pulse*, 3:32–42, 2012. doi: 10.1109/MPUL.2012.2216717.

Ziling Lu, Zongsheng Li, Xinke Shen, Kexin Lou, Yingyue Xin, Xiaoqi Chen, Shinan Wang, Xiang Chen, Jiahao Fan, Chenyu Huang, Xin Xu, Zhoujie Hou, Chen Wei, and Quanying Liu. OmniEEG-Bench: A standardized evaluation benchmark for EEG foundation models. *arXiv preprint arXiv:2606.00815*, 2026.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *International Conference on Learning Representations (ICLR)*, 2026.

Kaare B. Mikkelsen, David Bové Villadsen, Marit Otto, and Preben Kidmose. Automatic sleep staging using ear-EEG. *BioMedical Engineering OnLine*, 16:111, 2017. doi: 10.1186/s12938-017-0400-5.

Bojana Mirkovic, Martin G. Bleichner, Maarten De Vos, and Stefan Debener. Target speaker detection with concealed EEG around the ear. *Frontiers in Neuroscience*, 10:349, 2016. doi: 10.3389/fnins.2016.00349.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196.

Martin Schecklmann, Veronika Vielsmeier, Thomas Steffens, Michael Landgrebe, Berthold Langguth, and Tobias Kleinjung. Relationship between audiometric slope and tinnitus pitch in tinnitus patients: insights into the mechanisms of tinnitus generation. *PLoS ONE*, 7(4):e34878, 2012.

Deeksha M. Shama, Punnisa Amornsirikul, and Archana Venkataraman. EEG-PRISM: Physiologically-grounded interpretability of predictions by EEG foundation models. *arXiv preprint arXiv:2608.13676*, 2026.

Fanqi Shen, Enhong Yang, Jiahe Li, Junru Hong, Xiaoran Pan, Zhizhang Yuan, Meng Li, and Yang Yang. Brain4FMs: A benchmark of foundation models for electrical brain signal. *arXiv preprint arXiv:2602.11558*, 2026.

Toyotaro Suzumura, Hiroki Kanezashi, and Shotaro Akahori. Graph adapter of EEG foundation models for parameter efficient fine tuning. In *AAAI Workshop on Health Intelligence (W3PHIAI)*, 2025. arXiv:2411.16155.

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 (NeurIPS)*, 2024.

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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025a.

Siwen Wang, Shitou Zhang, Wan-Lin Chen, Dung Truong, and Tzyy-Ping Jung. From theory to application: Fine-tuning large EEG model with real-world stress data. In *International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 752–755, 2025b. doi: 10.1109/NER61569.2025.11588937.

Yihe Wang, Nan Huang, Nadia Mammone, Marco Cecchi, and Xiang Zhang. LEAD: An EEG foundation model for Alzheimer's disease detection. *arXiv preprint arXiv:2502.01678*, 2025c.

Yihe Wang, Zhiqiao Kang, Bohan Chen, Yu Zhang, and Xiang Zhang. Benchmarking ERP analysis: Manual features, deep learning, and foundation models. *IEEE Transactions on Biomedical Engineering*, 2026. doi: 10.1109/TBME.2026.3686229.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2508.17742.

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 (NeurIPS)*, volume 36, pp. 78240–78260, 2023.

Marzieh Zare. A negative-control protocol for clinical EEG foundation-model benchmarks: Dataset identity and external-cohort stress testing. *arXiv preprint arXiv:2607.24519*, 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 (NeurIPS)*, 2025. arXiv:2506.23075.

A Experimental Setup and Trial Timing

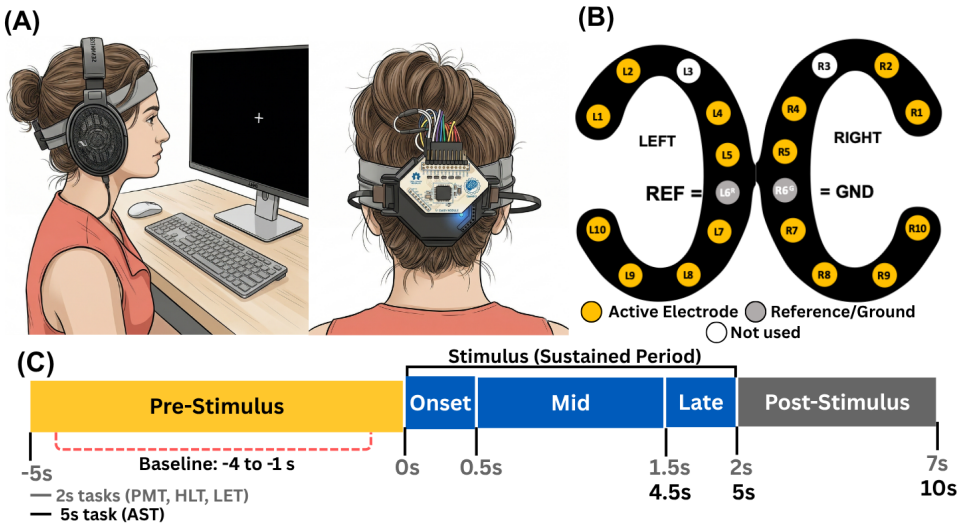

Figure 4: **(A)** Experimental setup: participant passively listens to auditory stimuli while wearing cEEGrid ear-EEG and over-ear headphones in a sound-attenuated booth. **(B)** cEEGrid layout: 16 active electrodes (L1–L10 left, R1–R10 right) with shared reference (L6) and ground (R6). **(C)** Trial timing from the original experimental protocol: the $[-4, -1]$ s period (red dashed) is the pre-stimulus silence, and stimulus onset is at 0 s. We release $[-1, +7]$ s epochs (and $[-5, +7]$ s for the window sweep), and each model ingests the first 4–5 s from -1 s (§B).

B Implementation Details for Nonstandard Model Loaders

Table 5: Eight EEG foundation models considered. TUEG = Temple University EEG Corpus, STFT = Short-Time Fourier Transform, JEPA = Joint-Embedding Predictive Architecture, AR = autoregressive, attn. = attention.

Model	Venue	Pretrain channels	Pretrain f_s	Params	Architecture
BENDR (Kostas et al., 2021)	Frontiers'21	20 (TUEG)	256 Hz	157.1M	conv-encoder
BIOT-PREST-16ch (Yang et al., 2023)	NeurIPS'23	16 (bipolar scalp)	200 Hz	3.2M	STFT-token
BIOT-six-18ch (Yang et al., 2023)	NeurIPS'23	18	200 Hz	3.2M	STFT-token
LaBraM-base (Jiang et al., 2024)	ICLR'24	64 (multi-source)	200 Hz	5.8M	patch-token
EEGPT (Wang et al., 2024)	NeurIPS'24	62 (1005)	256 Hz	25.3M	AR transformer
CBraMod (Wang et al., 2025a)	ICLR'25	22 (TUEG)	200 Hz	5.0M	criss-cross attn.
SignalJEPA-w/o-c (Guetschel et al., 2024)	arXiv'24	flexible	128 Hz	3.5M	JEPA
CodeBrain (Ma et al., 2026)	ICLR'26	19 (TUEG)	200 Hz	106.5M	state-space

Adaptation axes. *Channel adaptation* is fixed by each architecture. Five models are rebuilt for 16 channels and load every pretrained weight except the replaced embeddings (Recon), the fixed-input BIOT-18ch and BENDR place the 16 channels at random slots of their N-channel input and zero-pad the rest over five seeds (Rpad), and SIGNALJEPA takes any channel count (Flex). *Sampling rate*: S0 feeds 125 Hz directly, and S1 resamples to the rate the loader expects (the pretraining f_s of Table 5, except 250 Hz for BENDR and EEGPT). *Input window*: each model ingests D seconds from -1 s (D=5 for the BIOT variants and 4 otherwise), so the grid uses $[-1, +4]$ s and $[-1, +3]$ s. We fixed this window after a smoke test in which LABRAM, with failed channels kept, scored 0.54 on a post-onset window and 0.83 on $[-1, +3]$ s. Trials are z-scored over their own window before any truncation. *Failed channels*: the QC mask flags channels that fail impedance, variance or correlation gates. Interpolation replaces each by the mean of the passing channels of the same ear, and the three-channel control keeps only L01, L08 and L09, which pass in every subject.

Statistics. For each model the grid crosses its sampling-rate, embedding and probe levels ($2\times2\times4$=16 cells with channel embeddings, 2×4=8 for BENDR, SIGNALJEPA and CODEBRAIN), and CBRAMOD's S0 cells are not run (see Input length), which leaves 96 cells. Evaluation is leave-one-subject-out with subject-level AUROC, five seeds for stochastic configurations, bootstrap 95% CIs over subjects, paired bootstrap between models and per-cell label-permutation tests. Nested LOSO chooses the candidate by inner LOSO within each training fold, and max-statistic permutation tests repeat the choice on permuted labels. Leave-pair-out AUROC (Airola et al., 2011) checks the negative bias of pooled LOSO AUROC at small n. Positive controls compare post- and pre-onset windows of 1760 trials across subjects, on the recorded trials and with a synthetic evoked response added (Appendix N). Three supervised baselines (EEGNET, SHALLOWFBCSPNET, EEGCONFORMER) are trained end to end by LOSO over three seeds with AdamW and a cosine learning-rate schedule for 20 epochs.

Table 5 lists the audited checkpoints. Three foundation models required nonstandard handling that we document for reproducibility. EEGPT and CBRAMOD are provided in `braindecode` 1.4 only, which broke LABRAM's API, so we maintain two virtual environments and route feature extraction through the appropriate one, sharing a single feature-cache directory for the downstream probe. SIGNALJEPA-without-chans has `max_abs_coordinate`=0 in the released checkpoint, which causes division-by-zero NaN in the position encoder, so we monkey-patch the position-encoding function to return zeros when $x_{\min} = x_{\max}$. CODEBRAIN is not in any released `braindecode` version, so we adapt the official GitHub repository (`https://github.com/jingyingma01/CodeBrain`) and load the HuggingFace checkpoint `YjMajy/CodeBrain`.

Embedding replacement. The LABRAM, BIOT and EEGPT loaders drop learned embeddings from the checkpoint by name: LABRAM's position embedding (65 pretrained slots, of which the input indexing reads the first 17) and temporal embedding (9 pretrained positions, of which a 4 s input uses 4), BIOT's channel tokens (16 or 18, the same count as its input) and EEGPT's channel table (62×512, of which the input indexing reads the first 16 rows). BIOT's sinusoidal positional encoding and EEGPT's rotary encoding are fixed functions and are unaffected. Z0 keeps the framework's random initialization and Z1 sets the dropped parameters to zero. CBRAMOD stores no such embedding, because a convolution computes its position encoding from the input, so its Z0 and Z1 are identical. The control Zp loads the dropped values back: LABRAM's temporal embedding sliced to the patch count, as its repository does, and the position slots its indexing reads, BIOT's channel tokens and EEGPT's table. Zt keeps only LABRAM's temporal embedding and sets its position embedding to zero, since scalp positions have no ear-EEG counterpart. Appendix M reports both.

Input length. At S0 the 4 s window has 500 samples. LABRAM and CODEBRAIN use 200-sample patches and truncate to 400 samples ($[-1, +2.2]$ s), while CBRAMOD, which has the same patch size, does not truncate, so its S0 cells are not run.

Probe hyperparameters. All probes standardize features within each training fold. P1: LDA (singular-value solver, no shrinkage) projecting to one dimension, then logistic regression (class-balanced, 5000 iterations). P2: L2 logistic regression (C=1, class-balanced, 5000 iterations). P3: linear head trained full-batch by AdamW (learning rate 10^{-3}, weight decay 10^{-3}, 200 epochs, class-balanced cross-entropy). P4: MLP with one hidden layer of 128 units, ReLU and dropout 0.3,

trained by AdamW (learning rate 10^{-3}, weight decay 10^{-3}, 300 epochs). With 22 training subjects and feature dimensions of 64 to 512, P1 operates in the $p>n$ regime, which is one reason we report all four probes.

C PAIRWISE PAIRED-BOOTSTRAP SIGNIFICANCE

Table 6 compares LABRAM and BIOT-18ch with the other models by paired bootstrap, failed channels interpolated. p values are two-sided, from 2000 paired bootstrap reps over subjects. The 21 pairs that do not involve LABRAM all have $p > 0.28$.

Table 6: Pairwise paired-bootstrap significance (2000 reps), failed channels interpolated. Δ = median paired-bootstrap AUROC difference on seed-averaged probabilities (row $-$ col). Bold = significant at $p<0.05$.

Comparison	Δ AUROC	p-value
LABRAM vs. SIGNALJEPA	$+0.217$	$\mathbf{0.028}$
LABRAM vs. BIOT-18ch	$+0.263$	$\mathbf{0.017}$
LABRAM vs. BIOT-16ch	$+0.313$	$\mathbf{0.004}$
LABRAM vs. CODEBRAIN	$+0.316$	$\mathbf{0.004}$
LABRAM vs. CBRAMOD	$+0.400$	$< \mathbf{0.001}$
LABRAM vs. EEGPT	$+0.393$	$\mathbf{0.002}$
LABRAM vs. BENDR	$+0.363$	$\mathbf{0.029}$
BIOT-18ch vs. SIGNALJEPA	-0.033	0.823 (n.s.)
BIOT-18ch vs. BIOT-16ch	$+0.049$	0.672 (n.s.)
BIOT-18ch vs. CODEBRAIN	$+0.053$	0.700 (n.s.)
BIOT-18ch vs. CBRAMOD	$+0.133$	0.342 (n.s.)
BIOT-18ch vs. EEGPT	$+0.127$	0.430 (n.s.)
BIOT-18ch vs. BENDR	$+0.095$	0.448 (n.s.)

D FULL GEBRU-STYLE DATASHEET FOR SYNAPSE

We follow the structure of Gebru et al. (2021). This appendix complements the documentation overview in §4.

Motivation. *For what purpose was the dataset created?* As a documented out-of-distribution validation testbed for auditing EEG foundation-model transfer to passive wearable clinical conditions. The dataset is also independently useful as the first publicly released wearable ear-EEG cohort with clinical sound-sensitivity-disorder labels collected under a passive-stimulus paradigm. *Who created the dataset and on behalf of which entity?* Anonymized for review. *Who funded the creation?* Anonymized for review.

Composition. *What do the instances represent?* Trial-level cEEGrid recordings (16 channels, 1000 samples at 125 Hz, $[-1, +7]$ s peri-stimulus window) for n=23 adult subjects across three auditory tasks and one visual task. *How many instances are there in total?* 1,874 trial-level recordings (575 LET, 575 HLT, 610 AST, 114 PMT), 23 subject-level feature vectors after trial→task→subject aggregation. *Does the dataset contain all possible instances or is it a sample?* A sample from a single recruitment site over a single recruitment window. *What data does each instance consist of?* Raw 16-channel waveform, per-channel QC mask, task label, subject-level binary EXP/CTRL label, and subject age (5-year bins) and sex. *Is there a label or target?* Yes, subject-level binary EXP vs. CTRL. *Is any information missing?* Continuous questionnaire scores are not part of this release, and free-text demographic fields are excluded under IRB. Used as scores against the label, sex gives 0.84 AUROC, exact age 0.74 (0.705 from the released 5-year bins) and the rule "male or older than 35" 0.91 (Lesson 4). *Are relationships between individual instances made explicit?* Yes: trials are nested within tasks within subjects, and the released aggregation script makes this hierarchy explicit. *Are there recommended data splits?* Yes: leave-one-subject-out (LOSO) is the recommended evaluation split, used throughout the paper. *Are there errors, sources of noise, or redundancies?* Per-channel

impedance/variance/correlation QC failures are flagged in the released mask, and subjects failing per-ear QC were excluded (17 of 40, 11 of them with all 16 channels flagged). Failed channels remain in the released signals and the mask is group-correlated (0.76 AUROC from the mask alone), so models should be compared on the three channels that pass QC in every subject (L01, L08, L09) or with failed channels dropped, since interpolation leaves the failure pattern recoverable (Lesson 1). *Is the dataset self-contained?* Yes, all signals and labels are bundled in the released archive. *Does the dataset contain confidential or PII data?* No directly identifying information. Subject IDs are pseudonymized study codes (group prefix and enrollment number) that carry no personal information. *Sensitive content?* Clinical disorder status is sensitive, and all data, including age and sex, are released under CC-BY-4.0 with pseudonymized subject IDs.

Collection process. *How was the data acquired?* cEEGrid (16-electrode around-the-ear array) recordings during three passive auditory tasks (LET, HLT, AST) and one visual task (PMT). *What mechanisms or procedures were used?* Standardized stimulus presentation, controlled-volume calibration, IRB-approved consent. *Who was involved?* Trained research-assistant operators under PI supervision (anonymized for review). *Over what timeframe?* Single recruitment window (anonymized for review). *Were ethical review processes used?* Yes, IRB approval (anonymized for review), with informed consent from all subjects. *Did individuals consent to inclusion?* Yes, and release of de-identified EEG, group labels, age and sex is covered by the consent form. *Provided a mechanism to revoke consent?* Yes, subjects may withdraw and have their records removed from future releases. *Analysis of potential impact on data subjects?* Discussed in §7 (cohort imbalance disclosure and intended-use disclaimer).

Preprocessing/cleaning/labeling. *Preprocessing performed?* 1–50 Hz FIR Hamming bandpass and 60 Hz notch with harmonics, then interpolation of failed channels from passing channels of the same ear and per-trial z-score normalization, both at load time (raw signals stored, no re-referencing). *Was raw data saved alongside the preprocessed data?* Yes. *Software available for cleaning?* Yes, full reproduction scripts are released.

Uses. *Has the dataset been used for any tasks already?* Yes: this paper's 96-cell foundation-model transfer audit. *Is there a repository linking to all uses?* Yes (anonymized for review, linked at camera-ready). *What other tasks could the dataset be used for?* Auditory ERP analysis, representation-learning pretraining ablations, and transfer-learning research from scalp to ear. *Is there anything about its composition or its collection that might impact future uses?* Yes: cohort imbalance in age and sex, a group-correlated QC mask (§7), and $n{=}23$, which does not support population-level disorder generalization claims. *Are there tasks for which the dataset should not be used?* It should not be used as standalone evidence for clinical screening of sound-sensitivity disorders, nor for any actuarial or insurance-related decision-making.

Distribution. *Will the dataset be distributed to third parties?* Yes, public release. *How will it be distributed?* Hugging Face, through an anonymized repository during review, with the analyzed $[-1, +7]$ s epochs and the same trials on $[-5, +7]$ s for the window sweep. *When?* During review, through the anonymized repository. *Subject to copyright or other IP-based restrictions?* Raw EEG signals, derived features and per-cell results are released under CC-BY-4.0, and code under MIT. *Export controls or regulatory restrictions?* None known.

Maintenance. *Who is supporting/hosting/maintaining?* The authors (anonymized for review). *How can the curator be contacted?* Project repository issue tracker, and an email address at camera-ready. *Is there an erratum?* Yes, maintained alongside the public release. *Will the dataset be updated?* Yes, planned: additional subjects to address cohort imbalance and cross-task augmentations. *Will older versions be supported / hosted?* Yes, versions are tagged in the release archive. *Mechanism for others to extend / augment / build on?* Yes: dataset format and reproduction scripts released, and pull requests are welcome.

E SAMPLING-RATE AND EMBEDDING ABLATIONS

Table 7 gives the grid-window ablations behind Lessons 3 and 5.

Table 7: Ablations for Lessons 3 and 5, failed channels interpolated. *Top*: sampling rate (P2/Z1). *Bottom*: embedding replacement (P2, better sfreq for Z1). Bold = preferred per row. Δ is the difference of mean per-seed AUROCs, and the CIs in Insights 4 and 5 are paired bootstrap CIs (2000 reps over subjects) of the same statistic.

Model	**S0**	**S1**	$\Delta_{\text{S1-S0}}$
LaBraM	0.330	**0.875**	**+0.54**
SignalJEPA	0.438	**0.589**	+0.15
BENDR	0.307	**0.446**	+0.14
BIOT-18ch	0.616	**0.634**	+0.02
BIOT-16ch	**0.518**	0.491	−0.03
EEGPT	**0.527**	0.500	−0.03
CodeBrain	**0.634**	0.491	−0.14
Model	**Z0**	**Z1**	$\Delta_{\text{Z1-Z0}}$
LaBraM	0.705	**0.875**	**+0.17**
BIOT-18ch	0.500	**0.634**	+0.13
BIOT-16ch	0.496	**0.518**	+0.02
EEGPT	0.525	**0.527**	+0.00
CBraMod	0.536	0.536	0.00

F Robustness check: methodology insights on age-matched subcohort

To address the cohort imbalance threat (§7), we recompute Lessons 3 and 5 on an age-matched subcohort: EXP subjects with age ≤ 35 y (n=9, mean 25.6 y) and the full CTRL cohort (n=7, mean 25.3 y), for a total of n=16. Age distributions are well-matched (EXP range 18–31 y, CTRL range 20–34 y, from exact ages, and the released 5-year bins reproduce every subset used here). The sex imbalance is reduced relative to the full cohort (EXP 6M/3F, CTRL 7F), though not eliminated, since CTRL remains all female, so this check tests age only. Appendix J retrains LaBraM on females only (5 vs. 7).

Table 8 reports per-model Δ values on the age-matched subcohort alongside full-cohort values, failed channels interpolated. Eight of eleven sign comparisons are preserved. The three sign changes, BIOT-18ch Lesson 3 ($+0.02 \rightarrow 0.00$), EEGPT Lesson 3 ($-0.03 \rightarrow +0.10$) and BIOT-16ch Lesson 5 ($+0.02 \rightarrow 0.00$), all start from full-cohort effects of at most 0.03. LaBraM's gains from resampling and zero-init, the findings behind its configuration, grow on the matched subcohort ($+0.57$ and $+0.23$).

Table 8: Age-matched subcohort robustness check (n=16: 9 EXP $\leq$35 y + 7 CTRL). Full-cohort Δ is from Table 7, and matched Δ uses the same statistic on the same subject-level LOSO predictions, restricted to the 16-subject subcohort. P2 LogReg probe throughout.

Lesson	**Model**	**Full Δ**	**Matched Δ**	**Sign preserved?**
3: S1−S0	LaBraM	+0.54	+0.57	✓
	SignalJEPA	+0.15	+0.22	✓
	BENDR	+0.14	+0.10	✓
	BIOT-18ch	+0.02	0.00	× (full CI includes 0)
	BIOT-16ch	−0.03	−0.05	✓
	EEGPT	−0.03	+0.10	× (full CI includes 0)
	CodeBrain	−0.14	−0.17	✓
5: Z1−Z0	LaBraM	+0.17	+0.23	✓
	BIOT-18ch	+0.13	+0.15	✓
	BIOT-16ch	+0.02	0.00	× (full CI includes 0)
	EEGPT	+0.00	+0.00	✓

Table 8 restricts full-cohort predictions. Retrained within the subcohort, LaBraM reaches 0.73 (p=0.12), and Appendix J gives the retrained female-only analyses.

Interpretation. All three sign changes have full-cohort bootstrap 95% CIs that include zero ($[-0.29, +0.35]$, $[-0.21, +0.15]$ and $[-0.10, +0.17]$), so they reflect sampling noise rather than a change of direction. Every effect of at least 0.10 keeps its sign.

G STABILITY OF THE LABRAM RESULT: SUBJECT-LEVEL LOSO PREDICTIONS AND JACKKNIFE ANALYSIS

Figure 5 addresses the concern that 1–2 subjects may dominate the 23-fold LOSO mean. Because each LOSO fold withholds exactly one subject (yielding a single predicted probability per fold), AUROC cannot be computed within a single fold. We therefore present two complementary views.

Panel (A) plots each subject's seed-averaged predicted P(EXP) against rank order, coloured by true label, failed channels interpolated. For LABRAM, 13 of 16 EXP subjects lie above 0.6 and five of seven CTRL below 0.2, with no CTRL above 0.38. BIOT-18ch, shown for contrast, separates the groups less cleanly (three of seven CTRL below 0.2 and three above 0.99).

Panel (B) reports jackknife AUROC: for each of the 23 subjects, we recompute AUROC on the remaining 22 LOSO probabilities. This directly answers whether any single subject drives the mean. LABRAM jackknife AUROCs span $[0.948, 0.981]$ (std $= 0.008$), with a maximum single-subject leverage of 0.026, and BIOT-18ch spans $[0.646, 0.750]$. Both panels use seed-averaged probabilities, whose AUROC (0.955 for LABRAM) differs slightly from the mean of per-seed AUROCs in Table 2.

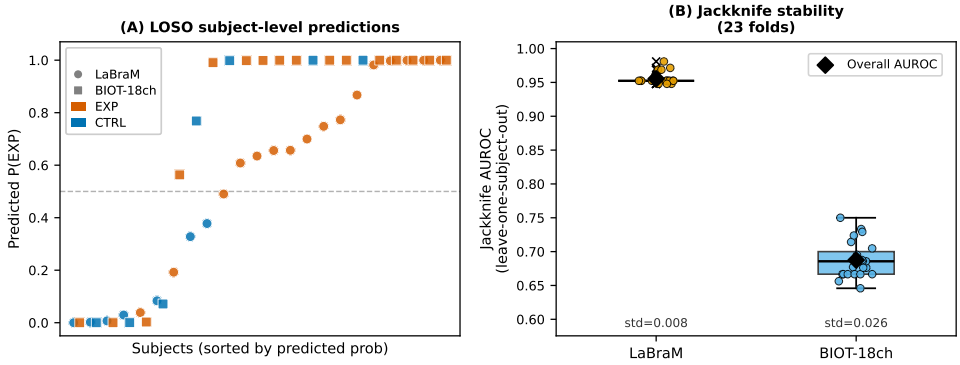

Figure 5: Stability of LABRAM (Recon/S1/Z1/P3), with BIOT-18ch (Rpad/S1/Z1/P4) for contrast, failed channels interpolated. **(A)** Seed-averaged LOSO predicted P(EXP) for all 23 subjects sorted by predicted value, coloured by true label. **(B)** Jackknife AUROC distribution over 23 leave-one-subject-out folds, with the overall AUROC marked.

H WINDOW SWEEP FOR ALL EIGHT MODELS

Each model keeps its loader length D and the other choices of its best P2 configuration with failed channels interpolated, uses the P2 probe, and slides its window over $[-5, +7]$ s in 0.2 s steps. Windows that start before -1 s use the released $[-5, +7]$ s epochs, with failed channels interpolated in the same way. The per-window predictions are released. At the grid window each model reproduces its cell of the 96-cell grid to within 0.01. CBRAMOD uses the loader of the main grid, where Z0 and Z1 give identical features.

Averaged over positions no model exceeds 0.62, and only LABRAM (7 positions) and SIGNALJEPA (4) pass the permutation threshold anywhere. The grid window is the third best of 41 positions for LABRAM and the 40th for CBRAMOD, so a fixed window flatters some models and penalizes others. With failed channels kept (Figure 3, dashed), BIOT-18ch ranges from 0.70 to 0.96 and passes the threshold at 33 of 36 positions, the signature of a static shortcut rather than a stimulus response.

Table 9: Window sweep with the P2 probe, failed channels interpolated. Step: median absolute AUROC change between adjacent 0.2 s windows. Grid rank: rank of the grid window among all positions. Above 0.77: positions above a fixed threshold of 0.77, slightly above the per-cell label-permutation 95th percentile (0.76 for the Z1 P2 cells tested), about 2 of 41 expected by chance.

Model	D (s)	Positions	Range	Mean	Step	Grid rank	Above 0.77
CBraMod	4	41	0.53–0.75	0.62	0.07	40/41	0
BIOT-18ch	5	36	0.48–0.67	0.59	0.03	3/36	0
CodeBrain	4	41	0.48–0.69	0.58	0.02	6/41	0
LaBraM	4	41	0.30–0.96	0.58	0.17	3/41	7
BIOT-16ch	5	36	0.46–0.62	0.55	0.03	28/36	0
SignalJEPA	4	41	0.12–0.89	0.53	0.14	15/41	4
EEGPT	4	41	0.40–0.62	0.51	0.04	14/41	0
BENDR	4	41	0.18–0.61	0.44	0.11	20/41	0
Band power (4 s)	4	41	0.46–0.55	0.50	0.01	–	0
Band power (5 s)	5	36	0.46–0.52	0.49	0.00	–	0

I Grid with failed channels kept and a passing-channel control

The full 96-cell grid with failed channels kept is released alongside the interpolated one, and Table 2 gives each model's best AUROC on it. Table 10 evaluates each model at the grid window, with the P2 probe and its best Z1 configuration with failed channels kept, under three channel treatments: failed channels kept, interpolated, and only the three channels that pass QC in every subject (L01, L08, L09), which no failed channel can reach. BIOT-18ch loses its advantage under both controls, while LaBraM improves under both.

Table 10: Grid-window P2 AUROC under three channel treatments. Band power is per channel when channels are kept or restricted and per ear on passing channels when interpolated. Kept and interpolated values are the grid cells, and the three-channel values come from the channel-control run. The QC mask alone gives 0.759.

Model	**Kept**	**Interpolated**	**Three passing channels**
LaBraM	0.830	0.875	0.875
BIOT-18ch	0.848	0.616	0.455
CodeBrain	0.688	0.634	0.688
CBraMod	0.607	0.536	0.661
EEGPT	0.589	0.527	0.527
SignalJEPA	0.571	0.438	0.482
BENDR	0.529	0.307	0.516
BIOT-16ch	0.464	0.491	0.420
Band power	0.723	0.527	0.688

J Selection-aware analyses, time-matched resampling and leave-pair-out AUROC

Selection over windows (set B). LaBraM's candidates are its 41 sweep windows (S1/Z1) and the three other grid configurations at the grid window (S0/Z1, S0/Z0, S1/Z0), each with P1 and P2, 88 in total, with failed channels interpolated. In nested LOSO, every candidate is scored by LOSO on the 22 training subjects of each outer fold, and the best one is refit on them and predicts the held-out subject. The inner selection chose $[-3, +1]$ s with P2 in 17 of 23 folds, and the outer predictions give 0.69. Repeating the nested procedure on 200 label permutations gives a null 95th percentile of 0.86 and p=0.19. The max-statistic test permutes labels 1000 times and rescores all candidates by LOSO: the null maximum has a 95th percentile of 0.947, just below the best observed candidate (0.955), and p=0.051. In the same permutations the number of windows (P2) above 0.77 has mean 1.7 and 95th percentile 7, against 7 observed (p=0.057). With 1000 permutations the Monte Carlo standard error of a p value near 0.05 is about 0.007.

Selection over grid cells with all four probes (set A). Set A of Table 3 repeats the choice that produced each model's grid best: its configurations at the grid window, each with P1–P4, scored as in the grid (P3, P4 and seed-dependent features averaged over five seeds). P3 and P4 are retrained on every permutation, all LOSO folds at once with one parameter tensor per fold, which is equivalent to training each fold separately because AdamW updates each parameter independently. This implementation draws its initializations differently from the grid run. P3 is sensitive to them: over 50 initialization seeds LABRAM's S1/Z1/P3 cell has mean 0.943 (SD 0.021, range 0.87–0.98), so the five-seed mean of the grid (0.954) and of this run (0.921) are draws from the same randomized probe, and p values refer to the grid value. For LABRAM, the null maximum over the 16 cells has a 95th percentile of 0.870, the grid best gives p=0.004 (1000 permutations, and 0.015 for this run's own five-seed value), and nested LOSO gives 0.84, with P3 chosen in 17 folds, P1 in 5 and P2 in 1, all at S1/Z1, against a nested null 95th percentile of 0.84 (p=0.07, 100 permutations). For SIGNALJEPA (8 cells) and BIOT-18ch (16 cells) the grid best gives p=0.31 and 0.37 and nested LOSO 0.64 (p=0.30) and 0.63 (p=0.21). Adding LABRAM's Zt and Zp cells (32 cells) gives p=0.012 and a nested estimate of 0.66 (p=0.30). The grid window itself was chosen in a smoke test between $[0, +4]$ s and $[-1, +3]$ s. Adding LABRAM's 16 cells at $[0, +4]$ s to the candidates repeats that choice as well: the null 95th percentile rises to 0.895, the grid best gives p=0.009, and nested LOSO gives 0.84 (p=0.11), with every fold choosing the $[-1, +3]$ s window.

Time-matched resampling. At S0, LABRAM and CODEBRAIN see 3.2 s while S1 sees 4 s. Table 11 compares S0 and S1 on the same 8 s epoch (1000 and 1600 samples) and on nearly matched windows (S0 on $[-1, +2.2]$ s, S1 truncated to $[-1, +2.0]$ s).

Table 11: Resampling with matched time content, failed channels interpolated, AUROC with P2 (P3 in parentheses) and paired bootstrap 95% CI of the P2 difference (2000 reps over subjects). Inputs are z-scored over the matched interval, whereas the grid S0 cell is z-scored over 4 s before truncation. The exact-match LABRAM S0 value equals the grid S0 cell (0.330) by coincidence, with different subject-level predictions.

Model	Match	S0	S1	$\Delta_{\text{S1-S0}}$	95% CI (P2)
LABRAM	exact, $[-1, +7]$ s	0.330 (0.398)	0.598 (0.662)	$+0.27$ ($+0.26$)	$[-0.17, +0.66]$
LABRAM	near, 3.2 vs. 3.0 s	0.268 (0.302)	0.741 (0.746)	$+0.47$ ($+0.44$)	$[+0.10, +0.80]$
CODEBRAIN	exact, $[-1, +7]$ s	0.554 (0.598)	0.482 (0.514)	-0.07 (-0.08)	$[-0.25, +0.06]$
CODEBRAIN	near, 3.2 vs. 3.0 s	0.616 (0.598)	0.411 (0.446)	-0.21 (-0.15)	$[-0.39, -0.06]$

Leave-pair-out AUROC and demographic subsets. Leave-pair-out AUROC (Airola et al., 2011) holds out one EXP and one CTRL subject per fit and counts correctly ordered pairs, which avoids the negative bias of pooling LOSO predictions. Within-EXP decoding applies the P2 probe by LOSO to LABRAM S1/Z1 features of the 16 EXP subjects, with sex or age above 35 as the target. Table 12 reports leave-pair-out AUROC for each model's P2 cell (averaged over the five slot seeds for BENDR, whereas BIOT-18ch Rpad/Z1 features do not depend on the seed, because zero channel tokens make the slots interchangeable) and for LABRAM retrained within demographic subsets, with label-permutation p values (1000 permutations) for the subsets.

K How small clinical cohorts are evaluated: a survey of 16 studies

We surveyed recent studies that apply EEG foundation models to subject-level clinical labels (arXiv, OpenAlex citation lists of the audited models, and the benchmarks they cite), read each full text, and verified every entry against arXiv and the publisher record. Table 13 lists what each study states. Thirteen of the 16 evaluate a label with fewer than 100 subjects. Held-out test sets can be very small: the MDD split introduced by CBRAMOD (Wang et al., 2025a) and reused by later models (Zhou et al., 2025; Fang et al., 2025) tests on 10 subjects scored as 5 s samples. One study selects hyperparameters by nested cross-validation (Shama et al., 2026), two tune them on a separate subject-disjoint validation set (Xiong et al., 2026; Shen et al., 2026), and none selects the input window by nested cross-validation. Three headline the best of several configurations, split seeds or classifiers (Wang et al., 2025b; Lin et al., 2026b; Zare, 2026). Two report permutation-based

Table 12: Pooled LOSO and leave-pair-out (LPO) AUROC with the P2 probe, failed channels interpolated.

Model or subset	LOSO	LPO	Permutation p
LABRAM (S1/Z1)	0.875	0.875	0.005
CODEBRAIN (S0)	0.634	0.670	—
BIOT-18ch (S1/Z1)	0.634	0.643	—
SIGNALJEPA (S1)	0.589	0.625	—
CBRAMOD (S1/Z1)	0.536	0.625	—
EEGPT (S0/Z1)	0.527	0.545	—
BIOT-16ch (S0/Z1)	0.518	0.554	—
BENDR (S1)	0.446	0.457	—
LABRAM, females only (5 vs. 7)	0.800	0.800	0.079
LABRAM, age-matched (9 vs. 7)	0.730	0.762	0.117
LABRAM, females $\leq$35 y (3 vs. 7)	0.571	0.619	0.331

tests (Zare, 2026; Lin et al., 2026a) and none reports an age or sex baseline. Two audits show what these choices cost: protocol differences reverse rankings by up to 24 points (Lahiri et al., 2026), and a subject-disjoint rerun of a best-seed result falls from 0.90 to 0.43–0.50 (Lin et al., 2026a).

Table 13: Evaluation practice in 16 studies of EEG foundation models on subject-level clinical labels, as stated in each paper. Unit: whether segments or trials of one subject are scored as independent samples (seg.) or one prediction is made per subject or recording (subj.). Selection: nested cross-validation (nested), a separate validation set for hyperparameters (val.), validation only for early stopping or checkpoints (stop), best of several reported (best), or not stated (n.s.). Perm.: permutation-based test. No study reports an age or sex baseline. ?: unclear.

Study	Subjects (label)	Split	Unit	Selection	Perm.
Wang et al. (2025a)	64 (MDD)	fixed subject split 43/9/10	seg.	stop	no
Zhou et al. (2025)	64 (MDD)	same split	seg.	stop	no
Wang et al. (2025c)	23–189 (AD, 5 sets)	repeated subject splits, LOSO	both	stop	no
Wang et al. (2025b)	18 (stress)	recording-level random split	seg.	best	no
Lin et al. (2026b)	206 (dementia)	subject 5-fold	subj.	best	no
Suzumura et al. (2025)	63 (MDD)	10-fold, grouping ?	seg.	n.s.	no
Kastrati et al. (2025)	39–120 (OCD, SZ, PD, mTBI)	fixed subject split	subj.	n.s.	no
Xiong et al. (2026)	88 (AD/FTD)	fixed subject split	seg.	val.	no
Lahiri et al. (2026)	200 (epilepsy)	fixed subject split	?	?	no
Shen et al. (2026)	28–122 (MDD, SZ, PD, AD, ADHD)	subject 5-fold	?	val.	no
Lu et al. (2026)	31–121 (PD, MDD, AD, ADHD)	repeated subject splits	seg.	n.s.	no
Kommineni et al. (2026)	64 (MDD)	subject 5-fold	seg.	stop	no
Wang et al. (2026)	50–147 (ADHD, PD)	repeated subject splits	seg.	stop	no
Lin et al. (2026a)	17–65 (4 sets)	subject 5-fold	subj.	stop	yes
Zare (2026)	23–65, 1187 rec.	subject 5-fold, LOSO	seg.	best	yes
Shama et al. (2026)	177 (ASD)	subject 5-fold	subj.	nested	no

L PROBE SENSITIVITY

Table 14 gives the four probes at each model's S1/Z1 configuration, failed channels interpolated. The gap is $\max(\text{P3}, \text{P4}) - \max(\text{P1}, \text{P2})$.

M KEEPING THE REUSABLE PRETRAINED EMBEDDINGS

Table 15 compares the two replacements of the main grid (Z0, Z1) with the control that loads the reusable pretrained values back (Zp) and, for LABRAM, with the variant that keeps only the temporal embedding (Zt), failed channels interpolated (Appendix B lists what each loader replaces). No model changes its standing. LABRAM stays high at S1 and low at S0 under every replacement, and its Zp/S1/P2 cell (0.884) passes a per-cell permutation test (p=0.010) like its Z1 cell (0.875, p=0.005). Under Zp, BIOT-16ch, BIOT-18ch and EEGPT stay between 0.46 and 0.62. The replacement therefore does not explain the negative results. For LABRAM, Zp matches Z1 at P2 and falls below it at the best probe, and for BIOT-18ch Z1 is higher in all four comparisons.

Table 14: Probe sensitivity (S1/Z1 per model, failed channels interpolated). Bold = best probe per row.

Model	P1	P2	P3	P4	Gap
LABRAM	0.839	0.875	**0.954**	0.866	+0.08
SIGNALJEPA	0.527	0.589	**0.725**	0.571	+0.14
BIOT-16ch	0.482	0.491	0.480	**0.618**	+0.13
CODEBRAIN	0.295	0.491	0.504	**0.586**	+0.09
BIOT-18ch	0.562	0.634	0.650	**0.696**	+0.06
EEGPT	0.446	0.500	**0.550**	0.426	+0.05
CBRAMOD	0.518	0.536	**0.552**	0.491	+0.02
BENDR	**0.504**	0.446	0.275	0.407	−0.10

Table 15: Embedding replacement against keeping the pretrained values, failed channels interpolated. Each entry is P2 / best of P1–P4 (seed-averaged for P3, P4 and seed-dependent features). Z0: random, Z1: zero, Zt: pretrained temporal embedding only (LABRAM), Zp: every reusable pretrained embedding.

Model	f_s	Z0	Z1	Zt	Zp
LABRAM	S0	0.332 / 0.393	0.330 / 0.330	0.259 / 0.259	0.188 / 0.321
LABRAM	S1	0.705 / 0.773	0.875 / 0.954	0.848 / 0.893	0.884 / 0.884
BIOT-16ch	S0	0.496 / 0.541	0.518 / 0.634	—	0.545 / 0.580
BIOT-16ch	S1	0.516 / 0.588	0.491 / 0.618	—	0.527 / 0.616
BIOT-18ch	S0	0.468 / 0.555	0.616 / 0.645	—	0.575 / 0.586
BIOT-18ch	S1	0.500 / 0.650	0.634 / 0.696	—	0.543 / 0.607
EEGPT	S0	0.525 / 0.525	0.527 / 0.527	—	0.500 / 0.500
EEGPT	S1	0.516 / 0.555	0.500 / 0.550	—	0.455 / 0.489

N POSITIVE CONTROL AND POWER SIMULATION

Positive control. Every trial of the three auditory tasks gives two windows of the model's input length D, one before onset ($[-5, -5+D]$ s) and one starting at onset ($[0, D]$ s), taken from the released $[-5, +7]$ s epochs with failed channels interpolated and z-scored per window as in the grid. A P2 probe on trial-level frozen features separates the two windows by LOSO over the 23 subjects (1760 trials, 3520 windows), and the score is the AUROC over the held-out subject's windows, averaged over subjects, with a bootstrap CI over subjects. The endpoint needs no clinical label and no demographic variable. On the recorded data every model and configuration scores between 0.50 and 0.54, and band power 0.52 (Table 16). The grand-average evoked response is weak: over channels and latencies from 50 ms to 1 s, its largest across-subject t statistic (2.23) lies at the 95th percentile of the same statistic on pre-stimulus windows (2.19). For the spike-in, a P1–N1–P2 waveform (Gaussian peaks at 50, 100 and 180 ms with relative amplitudes 0.3, -1 and 0.6) is added to every channel of every trial before z-scoring, with the N1 amplitude a multiple of the median trial standard deviation over $[-1, +3]$ s ($9.2\,\mu$V). Pre-onset windows are unchanged. BENDR's trial-level features vary almost only along their first principal component (over 99% of the standardized variance, against 19–26% for LABRAM and EEGPT), which is consistent with its barely registering the injected response.

Subject-level positive example. A pseudo-label assigns 16 of the 23 subjects to a positive group. Among 200,000 random 16/7 splits, it is the one whose largest absolute correlation with the true label, sex, age above 35 and each QC-mask channel is smallest (0.14). Sex, age, the "male or older than 35" rule and the true label score it at 0.43–0.54, and the QC mask alone gives a LOSO AUROC of 0.11, a below-chance value of the kind LOSO produces when features carry no label information. The synthetic response of the positive control is added to every trial of the positive subjects, at 0 to 2 median trial standard deviations, and LABRAM goes through the audit against the pseudo-label: the grid-window S1/Z1/P2 cell, the three-channel control, set A with all four probes (500 permutations, 50 for nested LOSO) and set B over 41 windows (1000 and 100). Table 4 lists the results. Nested p values have floors of 0.020 (set A, 50 permutations) and 0.010 (set B, 100). Under the verdict rule

Table 16: Positive control: post- vs. pre-onset AUROC across held-out subjects, by injected N1 amplitude (multiples of the median trial standard deviation). Amplitude 0 is the recorded data. Configuration: each model's sweep recipe. BENDR, whose feature extraction is slow on CPU, uses slot seed 0 throughout (0.519 on the recorded data over five seeds).

Model	Config	0	0.5	1	2	4
LABRAM	S1/Z1	0.514	0.566	0.656	0.798	0.911
BIOT-18ch	S1/Z1	0.523	0.543	0.597	0.750	0.900
BIOT-16ch	S0/Z1	0.516	0.514	0.551	0.706	0.872
EEGPT	S0/Z1	0.515	0.663	0.816	0.910	0.945
CBRAMOD	S1/Z1	0.539	0.579	0.661	0.815	0.924
CODEBRAIN	S0	0.535	0.594	0.714	0.852	0.940
SIGNALJEPA	S1	0.537	0.634	0.738	0.857	0.935
BENDR	S1	0.521	0.507	0.507	0.540	0.633
Band power (4 s)	—	0.516	0.509	0.588	0.758	0.894

Table 17: Sampling rate and embeddings on the spike-in endpoint (amplitude 2), with paired bootstrap 95% CIs over subjects. Z comparisons use the sampling rate of Table 7. —: not applicable (no learned embedding). CBRAMOD is omitted, since its S0 cells are not run and its Z0 and Z1 coincide.

Model	S0	S1	$\Delta_{\text{S1-S0}}$	$\Delta_{\text{Z1-Z0}}$	$\Delta_{\text{Zp-Z1}}$
LABRAM	0.796	0.798	$+0.002$ [-0.018, $+0.020$]	-0.003 [-0.012, $+0.004$]	-0.009 [-0.022, $+0.004$]
BIOT-18ch	0.738	0.750	$+0.012$ [-0.014, $+0.037$]	-0.015 [-0.035, $+0.005$]	-0.004 [-0.015, $+0.009$]
BIOT-16ch	0.706	0.725	$+0.020$ [$+0.004$, $+0.036$]	-0.002 [-0.014, $+0.010$]	-0.006 [-0.013, $+0.001$]
EEGPT	0.910	0.871	-0.039 [-0.056, -0.022]	-0.001 [-0.005, $+0.001$]	-0.002 [-0.013, $+0.007$]
CODEBRAIN	0.852	0.844	-0.008 [-0.027, $+0.011$]	—	—
SIGNALJEPA	0.861	0.857	-0.005 [-0.020, $+0.011$]	—	—
BENDR	0.544	0.540	-0.004 [-0.039, $+0.028$]	—	—

of §5, the protocol does not confirm the pseudo-label at amplitude 0, leaves it undetermined at 0.5, where the max-statistic tests pass but nested LOSO does not, and confirms it at 1 and 2, with the window-level nested test at the edge (p=0.050) at amplitude 1.

Table 18: True AUROC at which each test reaches 80% power at 16 vs. 7 subjects (linear interpolation between simulated effect sizes), 16-dimensional features, 8 dimensions in parentheses. Columns: the effect is present in all 88 candidates or in 8 adjacent ones, and the single candidate is pre-specified inside them.

Test	Effect in all 88	Effect in 8 adjacent
Single pre-specified candidate	0.87 (0.86)	0.87 (0.86)
Max-statistic over 88 candidates	0.83 (0.81)	0.96 (0.94)
Nested LOSO over 88 candidates	0.95 (0.94)	0.96 (0.96)

Power simulation. Synthetic cohorts of 16 EXP and 7 CTRL subjects have 16-dimensional features for K=88 candidates, close to the effective dimension of the frozen subject-level features (participation ratio 1–7 across models). Noise follows an AR(1) chain over candidates with coefficient 0.97, chosen in a calibration run (600 null cohorts) whose null 95th percentile of the maximum LOSO AUROC over candidates matched the one observed for LABRAM's set B (0.947). A group effect shifts one feature direction by δ, either in every candidate or in 8 adjacent ones. Each cohort is analysed as in the paper (P2 under LOSO), and thresholds come from simulated nulls of 1000 cohorts (200 for nested LOSO): 0.759 for a single pre-specified candidate (0.759 observed for the Z1 P2 cells), 0.964 for the maximum over candidates and 0.858 for nested LOSO (0.86 observed). The 95th percentile of a maximum over AUROCs moves in steps of 1/112, and the calibration (0.947 with 600 cohorts) and the main run (0.964 with 1000) differ by two steps, which makes the simulated max-statistic power slightly conservative. The true AUROC of a configuration is the expected AUROC on new subjects of a P2 probe trained on 22 subjects, from 4000 fresh subjects per training set. Power uses 400 cohorts per effect size (200 for nested LOSO), and the batched logistic regression matches scikit-learn to 10^{-6} in decision values. Table 18 gives the true AUROC at which

each test reaches 80% power. With 8-dimensional features the values change by at most 0.03. With 64 isotropic dimensions the probe cannot exceed a true AUROC of about 0.83 at any effect size, because it fits noise in the dimensions without signal.