PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 02 No. 04 (2026) Page: 78-90
E-ISSN: 3123-7827

Calibration-Light Cross-Subject Signal-Readiness Screening for Motor-Imagery BCI with Self-Supervised Conformers, Selective Prediction, and Evidence-Grounded LLM Feedback

Mason Robinson[1], Helen Sun[2], Amelia Clark[3]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[2]Department of Computer Science, University of Kentucky, Lexington, KY, USA
[3]Department of Computer Science, Washington State University, Pullman, WA, USA

Email: [1*]m_robinson_vt@gmail.com

Abstract

Low-cost electroencephalography can widen access to brain-computer interface research, but cross-subject deployment is limited by device heterogeneity, electrode saturation, and scarce calibration data. We examined the six continuous recordings in the OpenBCI multidirectional mouse-cursor collection as a signal-readiness benchmark. The repository describes four motor-imagery activities, eye blinking, and rest, yet the files contain no event boundaries and all marker values are zero. Accordingly, the study does not assign motor-imagery classes or report task-decoding accuracy. All 450,651 rows were audited, harmonized to eight channels at 125 Hz, and divided into 506 five-second windows. A deterministic target marked the upper within-subject artifact-burden quartile. Subject-held-out evaluation compared bandpower-LDA, descriptor-map CNN, descriptor Conformer, masked-descriptor self-supervision, covariance alignment, and eight-window target calibration. Unlabeled-target CORAL-logistic achieved 0.823 mean balanced accuracy. The supervised Conformer obtained 0.689, whereas masked pretraining decreased it to 0.579. At 80.44% realized coverage, confidence-based withholding reduced accepted-window error from 0.135 to 0.093. A persistent-rail hard gate prevented lower-relative-burden predictions from certifying a defective recording. The structured output layer rendered 18 rule-verified, nonmedical feedback cards from measured quality and confidence fields. The study establishes what these files support directly and identifies the event metadata required for valid activity decoding.

Keywords: Electroencephalography; Openbci; Signal Quality; Cross-Subject Transfer; Conformer; Self-Supervised Learning; Selective Prediction; Calibration; Nonmedical Feedback.

Abstrak

Elektroensefalografi berbiaya rendah dapat memperluas akses terhadap penelitian antarmuka otak-komputer (brain-computer interface), namun penerapan lintas subjek terkendala oleh heterogenitas perangkat, saturasi elektroda, dan minimnya data kalibrasi. Kami mengkaji enam rekaman kontinu dalam koleksi kursor mouse multidireksional OpenBCI sebagai tolok ukur kesiapan sinyal. Repositori tersebut mencakup empat aktivitas imajinasi motorik, kedipan mata, dan kondisi istirahat, namun berkas-berkasnya tidak memuat batas peristiwa dan semua nilai penanda (marker) bernilai nol. Oleh karena itu, studi ini tidak menetapkan kelas imajinasi motorik ataupun melaporkan akurasi dekode tugas. Sebanyak 450.651 baris data diperiksa, diselaraskan menjadi delapan saluran pada frekuensi 125 Hz, dan dibagi menjadi 506 jendela berdurasi lima detik. Target deterministik ditetapkan berdasarkan kuartil atas beban artefak dalam subjek (within-subject). Evaluasi dengan metode subject-held-out membandingkan pendekatan bandpower-LDA, CNN descriptor-map, Conformer descriptor, self-supervision dengan masked-descriptor, penyelarasan kovarians, dan kalibrasi target delapan jendela. Metode CORAL-logistik dengan target tanpa label mencapai rata-rata akurasi seimbang (balanced accuracy) sebesar 0,823. Model Conformer dengan pembelajaran terawasi (supervised) memperoleh skor 0,689, sedangkan pra-pelatihan dengan metode masking menurunkan skor tersebut menjadi 0,579. Pada tingkat cakupan aktual 80,44%, penyaringan berbasis tingkat keyakinan (confidence-based withholding) mengurangi tingkat kesalahan jendela yang diterima dari 0,135 menjadi 0,093. Mekanisme hard gate berbasis ambang batas (persistent-rail) mencegah prediksi dengan beban relatif rendah untuk memvalidasi rekaman yang cacat. Lapisan keluaran terstruktur menghasilkan 18 kartu umpan balik non-medis yang telah diverifikasi berdasarkan aturan, menggunakan parameter kualitas terukur dan tingkat keyakinan. Studi ini menetapkan kemampuan analisis langsung dari berkas-berkas tersebut serta mengidentifikasi metadata peristiwa yang diperlukan untuk dekode aktivitas yang valid.

Kata Kunci: Elektroensefalografi; OpenBCI; Kualitas Sinyal; Transfer Lintas Subjek; Conformer; Pembelajaran Mandiri (Self-Supervised Learning); Prediksi Selektif; Kalibrasi; Umpan Balik Non-medis.

Received: 01-08-2026 | Revised: 09-08-2026 | Published Online: 10-08-2026

A. INTRODUCTION

Motor imagery changes sensorimotor rhythms without overt movement and is a control paradigm for noninvasive brain-computer interfaces (BCIs) [2]. A BCI turns neural activity into communication or control [3], and noninvasive EEG has controlled two-dimensional movement under carefully designed protocols [4]. Accessible hardware still requires event timing, subject-held-out evaluation, and signal-quality control.

Calibration remains a barrier. First-session procedures can help untrained users [5], while transfer learning reuses source structure to limit target burden [16]. Riemannian Procrustes [17], Euclidean alignment [18], and adversarial invariance [19] require held-out-subject assessment because neighboring windows share acquisition conditions. MOABB likewise emphasizes consistent preprocessing and subject-wise comparisons [20].

Recent work sharpens this deployment question in three ways. A calibration-light subject-independent motor-imagery study combined self-supervised pretraining with a Conformer [30], while approximately shared-feature theory formalizes why only part of a representation may transfer across domains [31]. Cross-dataset wearable activity recognition likewise shows that adaptation gains can coexist with residual device and population shift [32]. Cross-cloud forecasting reports the same separation between reusable structure and target-specific calibration under workload and topology shift [33], and uncertainty-aware late fusion demonstrates that confidence estimates should influence how heterogeneous evidence is combined rather than merely annotate the final score [34]. These studies motivate the source-only, unlabeled-target, and few-shot pathways used below, but they do not supply labels for the present files.

Decoder development spans band-specific features and compact networks. Filter-bank common spatial patterns require labeled trials and electrode geometry [6], and no classifier family dominates every EEG setting [7]. ConvNets [8], EEGNet [9], EEG Conformer [10], and ATCNet [11] motivate the compact comparisons here, but the endpoint is observed artifact burden rather than activity.

Self-supervision can exploit unannotated EEG. Temporal-context learning [12], BENDR [13], BIOT [14], and LaBraM [15] motivate source-only masked-descriptor pretraining for unseen-subject artifact discrimination, not undocumented cognitive decoding.

Evidence from adjacent biomedical settings also cautions against treating pretraining as automatically beneficial. Lightweight medical foundation-model work uses multi-task pretraining and parameter-efficient transfer [35], whereas uncertainty-aware medical vision-language classification couples prediction with calibrated communication [36]. Their positive use cases depend on task-aligned supervision and evaluation; in contrast, the current masked objective reconstructs descriptor columns while the downstream endpoint is a hand-specified artifact-burden rule. We therefore treat the self-supervised result as an empirical ablation rather than an assumed improvement.

Reliable deployment requires withholding uncertain decisions. Neural scores can be miscalibrated [21]; selective classification formalizes risk versus coverage [22], and SelectiveNet learns a reject option [23]. Conformal methods require exchangeability [24], [25], while adaptive inference addresses drift without eliminating subject shift [26].

Similar deployment architectures appear outside EEG. Probability-calibrated default tiering uses uncertainty-driven rejection [37], and a model-risk workflow combines calibration, distribution-free uncertainty, and explanations [38]. In infrastructure forecasting, calibrated capacity risk [39], conformal demand envelopes [40], multi-horizon operational risk curves [41], and few-shot cold-start adaptation [42] all separate point prediction from an action boundary. These are cross-domain methodological precedents, not evidence about neural physiology. Their relevance is the decision pattern: estimate uncertainty, defer when evidence is weak, and preserve an auditable route from measurement to action.

The version-2 OpenBCI collection is CC BY 4.0 and describes four imagined movements, eye blinking, and rest [1]. Its six continuous files lack event annotations and every marker is 0.0. Three declare eight-channel Cyton at 250 Hz; three declare 16-channel Cyton+Daisy at 125 Hz. We therefore analyze signal readiness under leave-one-subject-out (LOSO) evaluation and make no activity claim.

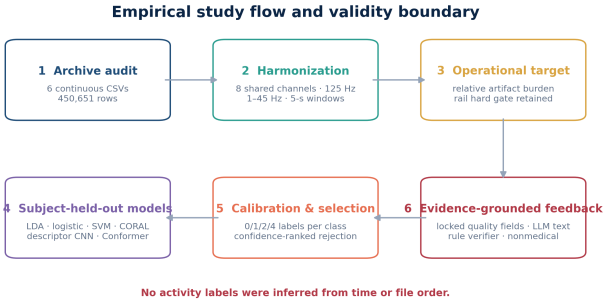

Figure 1 End-To-End Study Flow. The Red Validity Boundary Prevents Activity Labels From Being Inferred From Time Order.

B. METHODS

Dataset and integrity audit

Each CSV was parsed from its own header schema. Subjects 01-03 contain 86,217, 89,300, and 88,899 rows with eight EXG channels at 250 Hz; Subjects 04-06 contain 63,108, 62,278, and 60,849 rows with 16 channels at 125 Hz. The 450,651 rows equal 2,547.544 s at declared rates. Validation covered field count, numeric and finite values, timestamps, sample index, markers, and rail equality. Every marker was zero and no event file was present.

The analysis used EXG channels 0-7, the only channel indices shared across all subjects. The last eight Daisy channels were retained for archive-level quality reporting

but were not supplied to the cross-subject models. This avoids giving one device group a larger input space and prevents a model from identifying the board solely from channel count. Several full-record channels spent substantial time at the Cyton rail, including near-continuous saturation in parts of the Daisy recordings. Rail occupancy was therefore calculated on unfiltered voltage values and was never concealed by subsequent centering or filtering.

Table 1 Recording inventory measured from the six CSV files. Duration is sample-count equivalent, not trial duration.

Subject	Declared board	EXG	Hz	Rows	Nominal s	5-s windows
S01	Cyton	8	250	86,217	344.868	68
S02	Cyton	8	250	89,300	357.200	71
S03	Cyton	8	250	88,899	355.596	71
S04	Cyton+Daisy	16	125	63,108	504.864	100
S05	Cyton+Daisy	16	125	62,278	498.224	99
S06	Cyton+Daisy	16	125	60,849	486.792	97

Table 2 Marker And Saturation Audit. Every Row In Every Recording Has Marker Value 0.0.

Subject	Nonzero markers	Worst shared rail	Worst extra rail	Measured implication
S01	0	EXG 0: <0.01%	not present	Three rail samples; one all-zero EXG row
S02	0	0%	not present	No rail samples; one all-zero EXG row
S03	0	EXG 6: 6.28%	not present	Intermittent shared-channel clipping
S04	0	0%	EXG 14: ≈100%	Daisy-only channels 14–15 unusable
S05	0	EXG 3: ≈100%	EXG 10: ≈100%	Shared EXG 3 and three Daisy channels unusable
S06	0	0%	EXG 10: ≈100%	Daisy-only clipping; host timestamps bursty

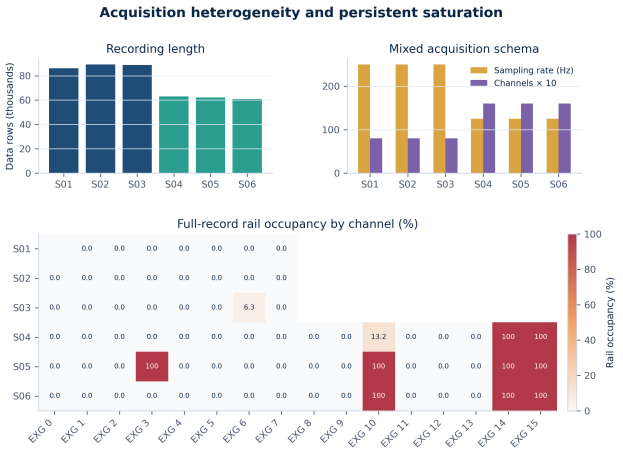

Figure 2 Measured Acquisition Schema And Rail Occupancy. Blank Heatmap Cells Denote Channels Absent From The Eight-Channel Files.

Harmonization And Five-Second Windows

At every time point, the median of the eight shared channels was subtracted as a common reference. A fourth-order zero-phase Butterworth filter retained 1-45 Hz. The first three recordings were then resampled from 250 to 125 Hz with a polyphase anti-aliasing filter; the remaining recordings retained their native rate. Raw-amplitude, repetition, and rail indicators were calculated before filtering. For learned models, descriptor scaling parameters were estimated on source-training subjects in each fold and applied unchanged to source validation and the held-out target.

Recordings were divided chronologically into nonoverlapping five-second windows, yielding 625 samples per channel. An incomplete terminal segment was excluded. Windows were never shuffled across the boundary used for target-subject calibration and testing. The five-second duration is long enough to estimate broad spectral and temporal quality characteristics while remaining short enough for actionable reacquisition feedback.

The operational target was binary high-artifact signal quality, not motor imagery. Six indicators were computed per window: maximum raw rail fraction, maximum exact-repeat fraction, the across-channel 90th percentile of filtered RMS, the 90th percentile of absolute excess kurtosis, the 90th percentile of 50 Hz power ratio, and the 90th percentile of 35-45 Hz power ratio. Within each subject, every indicator was converted to a nonnegative robust z score using its median and median absolute deviation and clipped at 12. The burden score was the weighted sum with weights 0.25, 0.15, 0.20, 0.15, 0.15, and 0.10 in the order listed. Windows at or above the subject-specific upper-quartile burden threshold were labeled high relative burden; the remainder were labeled lower relative burden. The rules were fixed before model fitting and applied identically to all six subjects. These labels encode relative acquisition burden, not physiology or absolute cleanliness. Per-indicator values and burden scores were retained so the target remained auditable. Independently,

any shared channel with at least 50% recording-level rail occupancy activated a hard failure that no lower-burden window prediction could override.

Table 3 Window Counts And Operational Target. Class 0 Is Relative Within-Recording Burden And Is Not An Absolute Clean-Signal Certificate.

Subject	Windows	Lower relative	Upper quartile	Burden cutoff	Independent hard gate
S01	68	51	17	1.540	None
S02	71	53	18	2.400	None
S03	71	53	18	1.916	None
S04	100	75	25	0.902	None
S05	99	74	25	0.532	Persistent rail
S06	97	72	25	1.112	None

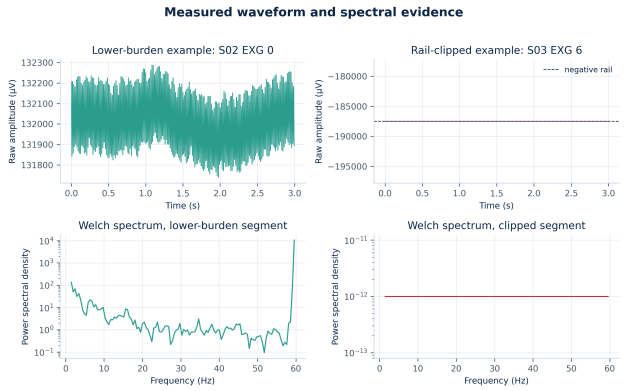

Figure 3 Raw Waveform And Welch Spectrum From A Lower-Burden Segment And A Rail-Clipped Channel. The Clipped Trace Shows Why Filtering Cannot Repair Saturated Acquisition.

Models And Representation Learning

Each filtered window was converted to an 8 by 12 descriptor map. Per-channel descriptors were log relative power in 1-4, 4-8, 8-13, 13-30, and 30-45 Hz; log activity; Hjorth mobility and complexity; log RMS; log line length; excess kurtosis; and skewness. The five bandpower descriptors were averaged across channels and standardized for shrinkage LDA. Additional classical comparisons flattened all 96 descriptors for balanced logistic regression and an RBF support-vector machine. These baselines preserve interpretable spectral comparisons [6] while adding nonlinear and all-feature controls.

The EEGNet-style neural baseline accepted the 8 by 12 descriptor map as a small two-dimensional input. An eight-filter 1 by 3 convolution operated along the descriptor axis, followed by batch normalization and a depthwise convolution across all eight channels with depth multiplier two. Exponential-linear activation, 1 by 2 average pooling, a 16-filter separable 1 by 3 convolution, 0.30 dropout, and a softmax head completed the model. It retains EEGNet's compact spatial-feature motivation [9] but operates on measured descriptors rather than raw samples.

The compact descriptor Conformer treated the 12 descriptor columns as tokens, each carrying the eight-channel pattern for one feature. A learned 16-dimensional embedding preceded two blocks. Each block used layer normalization, parameter-free normalized dot-product self-attention, a residual connection, and residual convolutional mixing with kernel sizes three and one. Global average pooling, 0.25 dropout, and a softmax head completed the classifier. The attention layer had no learned query, key, or value projections. This design retains the local-plus-global principle of EEG Conformer [10] while limiting capacity for 506 windows.

For self-supervision, one randomly positioned span of four contiguous descriptor columns was masked across all eight channels, and the encoder reconstructed the 32 normalized values by masked mean-squared error. Pretraining used only the four source-training subjects inside each LOSO fold; the held-out target and source-validation subject contributed no gradient or scaling statistics. The source-subject SQI head was then trained with class-weighted cross-entropy. This separation follows the label-efficient motivation of prior EEG self-supervision [12]-[15] and prevents target leakage.

CORAL aligned source descriptors to unlabeled target covariance before logistic classification and was reported separately because it observes the target distribution. Classical few-shot logistic models added the earliest one, two, or four target windows per class; their combined target weight equaled one mean source subject. Neural few-shot adaptation used the earliest four per class, froze the pretrained encoder, and fitted only its head for 15 epochs. Calibration windows were removed from evaluation. Chronological selection simulated an initial quality check rather than a favorable random draw. Source-only, CORAL, and few-shot settings therefore answer different deployment questions.

The comparison was therefore organized by target-information budget rather than by model family alone. The direct motor-imagery study [30], approximately shared-feature analysis [31], and cross-cloud transfer results [33] all indicate that source-only performance can conceal target-domain mismatch. Strict zero-shot models, unlabeled CORAL, and labeled target adaptation are consequently reported in separate rows. This prevents an adaptation method from being described as subject-independent merely because target labels were few.

Table 4 The 12 Measured Descriptors Forming Each 8 × 12 Neural Input Map.

Group	Per-channel descriptor	Definition	Quality role
Spectral	log-relative δ	log[P(1–4 Hz)/P(1–45 Hz)]	slow-band balance
Spectral	log-relative θ	log[P(4–8 Hz)/P(1–45 Hz)]	slow-band balance
Spectral	log-relative α	log[P(8–13 Hz)/P(1–45 Hz)]	alpha-band balance
Spectral	log-relative β	log[P(13–30 Hz)/P(1–45 Hz)]	beta-band balance

Group	Per-channel descriptor	Definition	Quality role
Spectral	log-relative γ	log[P(30–45 Hz)/P(1–45 Hz)]	high-band balance
Amplitude	log activity	log variance	dispersion
Hjorth	mobility	sqrt[var(Δx)/var(x)]	relative derivative scale
Hjorth	complexity	mobility(Δx)/mobility(x)	waveform complexity
Amplitude	log RMS	log root-mean-square	signal magnitude
Morphology	log line length	log mean |Δx|	rapid change
Morphology	excess kurtosis	fourth standardized moment − 3	transients/heavy tails
Morphology	skewness	third standardized moment	asymmetry

Table 5 Model Configurations Fixed For The Six LOSO Folds. Neural Parameter Counts Come From The Saved Training Log.

Model	Input	Core configuration	Target access	Parameters
Bandpower-LDA	5 averaged bands	standardization; shrinkage LDA	none	closed form
All-feature logistic	96 descriptors	C=1; balanced source classes	none	97 + intercept
RBF-SVM	96 descriptors	C=1; gamma=scale; probability fit	none	support vectors
CORAL-logistic	96 descriptors	Ledoit–Wolf covariance whitening/recoloring	unlabeled covariance	97 + intercept
Descriptor CNN	8 × 12 map	8 filters; depth multiplier 2; 16 separable filters	none	570
Descriptor Conformer	12 tokens	d=16; 2 blocks; parameter-free attention	none	2,418
Masked SSL Conformer	12 tokens	4 contiguous columns masked; 35 pretrain epochs	none	21,074
Masked SSL +4/class	12 tokens	same encoder; classifier-head update, 15 epochs	8 target labels	21,074

Training And Subject-Held-Out Evaluation

Each subject served once as the outer test subject. Classical models used all five non-target subjects for fitting. For neural models, the next subject cyclically served as source validation and the other four supplied training windows. Model selection, early stopping, descriptor scaling, and class weighting used source data only; no subject crossed training and validation partitions. Neural models minimized class-weighted cross-entropy with AdamW, fixed method-specific learning rates, and source-validation early stopping. Training preserved chronological order. Fold-derived seeds controlled initialization and masking. Parameter count, selected epoch, calibration indices, and elapsed time were stored. An independent deterministic rerun produced byte-identical predictions. Eight target windows were reserved before neural evaluation so all neural variants shared an identical test remainder; only the four-shot head used them for fitting.

Balanced accuracy was primary because the upper-quartile target is imbalanced. Secondary measures were macro-F1, accuracy, class recall, and Cohen's kappa. Probability quality used negative log likelihood (NLL), Brier score, and expected calibration error (ECE). Subject-level values were summarized by mean, standard deviation, and range; paired fold differences were retained because six folds do not justify strong asymptotic claims.

Probability Assessment And Selective Prediction

NLL, Brier score, and ten-bin ECE assessed the probability outputs [21]. Predictions were then ranked within each held-out subject by maximum class probability. At target coverages from 60% through 100%, the highest-confidence fraction returned a lower-burden/high-burden decision and the remainder was withheld. Coverage is the accepted fraction; selective risk is the error fraction among accepted windows [22]. Risk-coverage curves and area under the risk-coverage curve were evaluated in addition to full-coverage accuracy. This retrospective ranking quantifies the attainable tradeoff at a chosen coverage; it does not supply a universal confidence threshold for a future subject. No conformal coverage guarantee was attached because the strong subject and hardware shift conflicts with simple exchangeability [24]-[26].

To avoid translating cross-domain precedents into an unsupported guarantee, the study uses calibration metrics and retrospective risk-coverage curves rather than claiming conformal validity. The rejection logic parallels calibrated default-risk tiering [37] and conformal infrastructure control [40], but the held-out subjects are neither exchangeable replicates nor a stationary stream. The reported thresholds are therefore observed operating points, not clinical or universal acceptance criteria.

Constrained Language-Feedback Interface

The feedback layer was designed as a renderer over structured evidence, not as another EEG decoder. Its input fields were limited to deterministic quality flags, model probability, acceptance status, recording-level hard-gate status, dominant technical failure mode, and a permitted action code. Raw EEG and free-form user history were excluded. A lower-burden prediction could yield a short continuation message only if no hard gate was active; a confident high-burden decision could recommend checking contact, reducing movement, or repeating the segment; a withheld prediction could state that confidence was insufficient and request another five-second sample.

Evidence-constrained generation was selected because retrieval alone does not guarantee faithful output. Word-level hallucination detection and provenance tracing [43], deterministic evidence-chain explanations [44], abstention-aware RAG for unanswerable questions [45], and private-operations RAG [46] all emphasize explicit links between output claims and source evidence. OpsLLM-style incident triage extends this pattern to alert summarization [47]. In

the present interface, the evidence object is smaller and more restrictive: it contains only measured quality fields and an approved action, so the renderer cannot retrieve or infer an activity label.

The evaluated renderer used gpt-5.6-terra on August 5, 2026, with one fixed instruction and a small approved action vocabulary. Eighteen locked evidence records were submitted without raw EEG or activity labels, and the generated text was saved before verification. A deterministic checker then rejected diagnostic language, claims about attention or intention, treatment advice, confidence mismatches, exposed action codes, unsupported actions, and statements that a mental task had been recognized. This follows the CRED-nf emphasis on explicit feedback signal, modality, schedule, and outcome reporting [27], the NIST generative-AI emphasis on mapped risks and human oversight [28], and WHO principles of autonomy, transparency, and accountability for health-adjacent AI [29]. The interface is nonmedical and its technical messages are not evidence that language feedback improves BCI learning.

Safety is enforced at two layers. Response-level self-verification [48], staged refusal with utility preservation [49], and continual guardrail updates under new jailbreaks [50] motivate checking rendered text after generation rather than trusting the prompt alone. Human-facing risk-card research similarly stresses visible uncertainty, provenance, and bounded next actions for LLM agents [51], medical-image interfaces [52], patient-facing safety communication [53], and distributed-log incident displays [54]. Design-rationale interfaces [55], LLM-as-design-critic evaluation [56], and capacity-dashboard visual briefs [57] further support separating machine-readable evidence from user-facing hierarchy. These studies inform interface structure only; they do not validate the wording or efficacy of EEG feedback.

Data minimization is equally important. Confidence-gated conversational memory [58] shows that persistent state should be written, retrieved, updated, and forgotten under explicit controls, while hybrid-cloud LLM deployment [59] highlights placement and cost/privacy tradeoffs between local and remote inference. The current evaluation avoids those additional risks by sending no raw EEG and retaining no user history; a deployment that adds longitudinal memory would require consent, deletion, and access controls beyond the present experiment...

C. RESULTS AND DISCUSSION

Corpus-Level Findings And Signal-Quality Burden

All six files were read successfully under their declared schemas, accounting for all 450,651 data rows. Harmonization produced 506 complete five-second windows: 68, 71, 71, 100, 99, and 97 for Subjects 01-06. The audit placed 128 windows (25.30%) in the upper-quartile high-burden class. Per-subject counts were 17, 18, 18, 25, 25, and 25. The near-quarter prevalence follows from the operational definition; the more informative evidence of subject heterogeneity is that the burden cutoff ranged from 0.532 to 2.400 and the six indicator distributions differed substantially. A random window split would mix these subject conditions and overstate deployment performance.

The raw audit exposed defects that a band-pass filter alone would not resolve. Subject 03 channel 6 occupied the device rail for about 6.28% of its recording. Among the 16-channel files, Subject 04 channel 10 was saturated for about 13.16% and channels 14-15 were fully saturated; Subject 05 channels 3, 10, 14, and 15 were saturated for essentially the entire record; Subject 06 channels 10 and 14-15 were likewise fully saturated. Restricting the learned models to the shared first eight channels removes several unusable Daisy-only channels but retains Subject 05 channel 3. Consequently, Subject 05 fails the persistent-rail hard gate even when a window has lower burden relative to that recording. The window classifier ranks local burden; it cannot certify a defective session.

Cross-Subject Discrimination

The classical zero-shot models used every held-out window. Bandpower-LDA attained mean LOSO balanced accuracy of 0.673 and macro-F1 of 0.687. The all-descriptor logistic model reached 0.717 and 0.658, while the RBF-SVM reached 0.773 and 0.731, respectively. Neural variants were compared on the same 458-window subset after reserving eight calibration windows per target. The 570-parameter descriptor-map CNN obtained balanced accuracy 0.650 ± 0.205 and macro-F1 0.575 ± 0.267. The 2,418-parameter supervised descriptor Conformer reached 0.689 ± 0.154 and 0.667 ± 0.126.

Masked-descriptor pretraining did not improve transfer. Its 21,074-parameter training graph obtained balanced accuracy 0.579 ± 0.203 and macro-F1 0.566 ± 0.211; masked validation MSE was 0.799 ± 0.337. Relative to the supervised Conformer, balanced accuracy decreased by 0.110, and all six fold differences were negative, ranging from -0.349 to -0.010. An exact paired sign-flip test gave two-sided $p = 0.03125$. The result indicates that reconstructing four withheld descriptor columns was not aligned with cross-subject discrimination of the weighted artifact-burden target at this sample size.

The negative masked-pretraining result is therefore not inconsistent with positive reports in task-labeled motor-imagery BCI [30] or multi-task biomedical transfer [35]. Pretraining helps when the surrogate objective preserves factors that matter to the downstream label; here, reconstructing neighboring descriptor columns may encode board regime or general signal morphology more strongly than the weighted burden rule. The approximately shared-feature perspective [31] predicts this possibility: some structure transfers while task-relevant components remain domain-specific.

The fold spread is at least as important as the mean. Supervised Conformer balanced accuracy ranged from 0.505 to 0.929 across held-out subjects. This variability is consistent with mixed boards and concentrated saturation. A model can learn cross-channel descriptor relations, but a

representation selected from five source subjects cannot assume that the sixth shares their quality distribution. The result supports calibration-light language, not a claim that individual calibration has been eliminated.

The classical baseline remains informative even where a neural model ranks first. Its spectral summaries trace decisions to measured acquisition properties, whereas neural encoders can exploit relations among all 12 descriptors. Conversely, performance on this SQI target does not imply motor-imagery decoding. Filter-bank CSP, EEGNet, and Conformer were developed with task labels [6], [9], [10]; here the output head predicts only relative artifact burden.

Table 6 Mean ± Subject SD For Six Held-Out Subjects. Neural Methods Share The Same Post-Calibration Test Subset; Classical Rows Use All Eligible Target Windows Unless Stated Otherwise.

Model	Target access	n/fold	Balanced acc.	Macro-F1	NLL	Brier	ECE
Bandpower-LDA	none	84.3	0.673 ± 0.097	0.687 ± 0.109	0.487	0.150	0.111
All-feature logistic	none	84.3	0.717 ± 0.167	0.658 ± 0.242	1.081	0.230	0.224
RBF-SVM	none	84.3	0.773 ± 0.139	0.731 ± 0.179	0.476	0.151	0.122
CORAL-logistic	unlabeled target	84.3	0.823 ± 0.099	0.818 ± 0.085	0.406	0.111	0.089
Descriptor CNN	none	76.3	0.650 ± 0.205	0.575 ± 0.267	0.597	0.207	0.221
Descriptor Conformer	none	76.3	0.689 ± 0.154	0.667 ± 0.126	0.553	0.183	0.148
Masked SSL Conformer	none	76.3	0.579 ± 0.203	0.566 ± 0.211	0.566	0.194	0.174
Masked SSL +4/class	8 labels	76.3	0.580 ± 0.186	0.578 ± 0.188	0.541	0.183	0.141

Table 7 Balanced Accuracy By Held-Out Subject. The Neural Columns Use The Common Target Subset After Eight Early Windows Were Reserved.

Subject	LDA	RBF	CORAL	CNN	Conformer	SSL 0	SSL +4
S01	0.824	0.931	0.951	0.500	0.798	0.671	0.671
S02	0.750	0.879	0.823	0.837	0.643	0.633	0.643
S03	0.647	0.877	0.925	0.908	0.929	0.893	0.857
S04	0.660	0.613	0.793	0.690	0.690	0.341	0.375
S05	0.573	0.633	0.746	0.364	0.505	0.383	0.376
S06	0.586	0.706	0.698	0.599	0.571	0.553	0.558

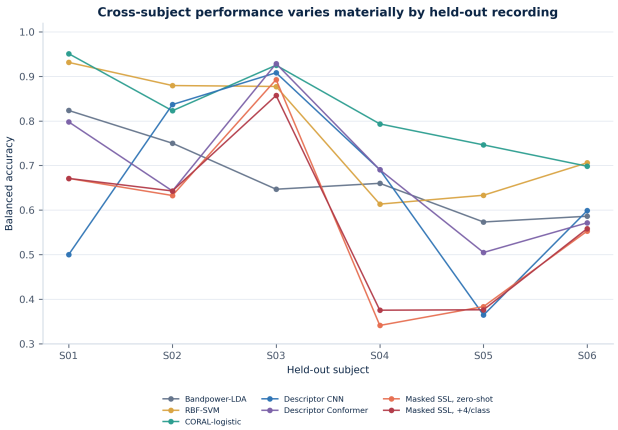

Figure 4 Balanced Accuracy For Seven Model Settings Across The Six Held-Out Subjects. Neural Methods Use The Matched Post-Calibration Target Subset.

Domain Alignment And Eight-Window Calibration

CORAL-logistic obtained 0.823 mean balanced accuracy after observing unlabeled target-window covariances, 0.150 above bandpower-LDA. The improvement was positive in all six folds and ranged from 0.073 to 0.278; an exact subject-level sign-flip test gave two-sided $p = 0.03125$ across the 64 possible sign patterns. Covariance matching can therefore help when change is expressed in second-order descriptor structure, but it cannot repair a rail-clipped channel. Because CORAL accesses the target distribution, it is best understood as unsupervised session adaptation rather than strict subject-independent inference.

CORAL's consistent improvement is also compatible with cross-dataset adaptation evidence in wearable sensing [32] and cross-cloud forecasting [33], where representation-level shift can be reduced without assuming identical targets. The hard-gate exception remains essential: covariance alignment can make distributions more comparable, but it cannot restore information lost to persistent saturation.

For the masked-pretrained neural model, fitting the head on the earliest four target windows from each SQI class yielded balanced accuracy 0.580 ± 0.186 and macro-F1 0.578 ± 0.188. The balanced-accuracy gain over its zero-shot counterpart was only 0.001, although macro-F1 increased by 0.012, NLL decreased by 0.025, Brier score decreased by 0.011, and ECE decreased by 0.033. Classical four-per-class logistic adaptation was stronger at 0.814 ± 0.136 balanced accuracy and 0.821 ± 0.107 macro-F1. All eight calibration windows were excluded from testing. These results concern quality-burden calibration only and do not estimate labeled motor-imagery calibration needs.

Table 8 Target Calibration Dose. Every Labeled Target Window Was Excluded From The Corresponding Test Set.

Family	Labels/class	n/fold	Balanced acc.	Macro-F1	NLL	Brier	ECE
Logistic	0	84.3	0.717 ± 0.167	0.658 ± 0.242	1.081	0.230	0.224
Logistic	1	82.3	0.739 ± 0.172	0.724 ± 0.162	0.623	0.137	0.123
Logistic	2	80.3	0.773 ± 0.165	0.765 ± 0.145	0.451	0.115	0.109
Logistic	4	76.3	0.814 ± 0.136	0.821 ± 0.107	0.350	0.089	0.072
Masked SSL	0	76.3	0.579 ± 0.203	0.566 ± 0.211	0.566	0.194	0.174
Masked SSL	4	76.3	0.580 ± 0.186	0.578 ± 0.188	0.541	0.183	0.141

Table 9 Self-Supervision Ablation. Masked Reconstruction Converged, But Downstream Cross-Subject Discrimination Did Not Improve.

Ablation	Masked pretrain	Target labels	Parameters	Masked MSE	Balanced acc.	Macro-F1	Δ vs sup.
Supervised Conformer	none	0	2,418	—	0.689	0.667	reference
Masked SSL, zero-shot	35 epochs	0	21,074	0.799	0.579	0.566	-0.110
Masked SSL, +4/class	35 epochs	8	21,074	0.799	0.580	0.578	-0.109

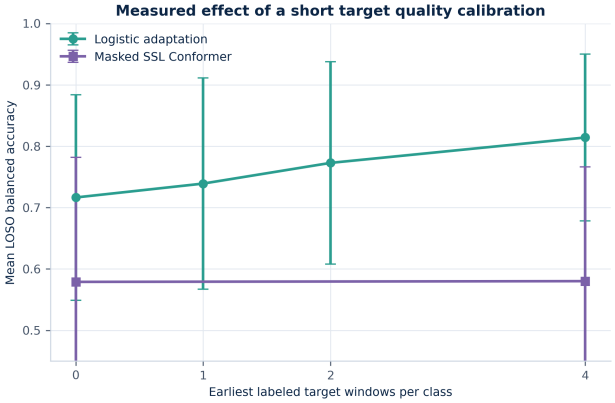

Figure 5 Target Quality-Calibration Dose. Error Bars Show One Subject Standard Deviation; Masked Pretraining Did Not Yield A Discrimination Gain.

Probability Quality And Selective Operation

CORAL-logistic had mean NLL 0.406, Brier score 0.111, and ECE 0.089, compared with 0.476, 0.151, and 0.122 for the zero-shot RBF-SVM. These averages complement discrimination but do not imply that one probability threshold is calibrated for every subject [21]. The fold-level CORAL ECE and Brier values varied with acquisition condition, reinforcing the need to inspect reliability by target subject.

At a target coverage of 80%, realized mean coverage was 80.44%, selective balanced accuracy was 0.874, and accepted-window error was 0.0926, compared with 0.1352 error at full coverage. Mean area under the risk-coverage curve was 0.0705. At 60.48% realized coverage, accepted error fell further to 0.0778 and balanced accuracy rose to 0.895. Withholding low-confidence decisions therefore offered a measured tradeoff: fewer automatic decisions in return for lower accepted risk. A withheld decision can trigger another five-second recording or the eight-window calibration pathway rather than silently accepting an uncertain quality state. Because the cutoffs were obtained by ranking each target fold, they describe operating characteristics rather than a precommitted threshold for new deployments.

Operationally, the policy resembles systems that turn uncertainty into bounded escalation rather than free-form confidence language. Calibrated credit workflows [37], [38] and risk-bounded capacity controls [39], [40], [41], [42] all distinguish prediction quality from decision cost. The reacquisition action here is deliberately narrower: it changes data collection, not diagnosis, treatment, or resource allocation.

Table 10 CORAL-Logistic Selective Operation, Averaged Across Subject Folds. Thresholds Are Fold-Specific Retrospective Operating Points.

Target coverage	Realized	Mean threshold	Accepted risk	Balanced acc.
100%	100.0%	0.529	0.135 ± 0.058	0.823
90%	90.5%	0.683	0.110 ± 0.047	0.853
80%	80.4%	0.784	0.093 ± 0.053	0.874
70%	70.4%	0.866	0.085 ± 0.055	0.882
60%	60.5%	0.917	0.078 ± 0.055	0.895

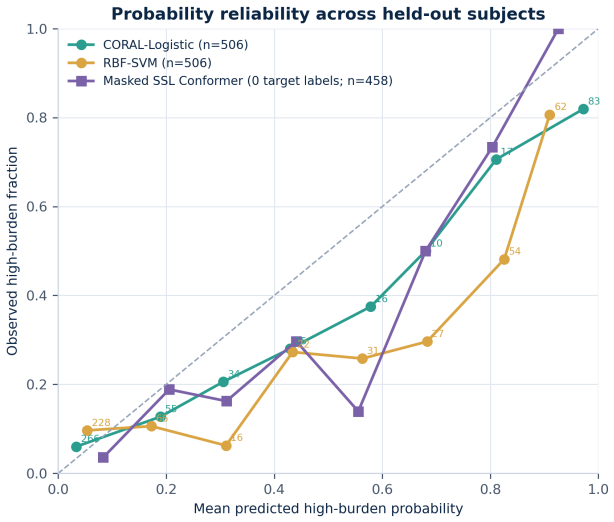

Figure 6 Pooled Reliability Curves. Numbers Beside Classical Points Indicate Windows In Each Probability Bin.

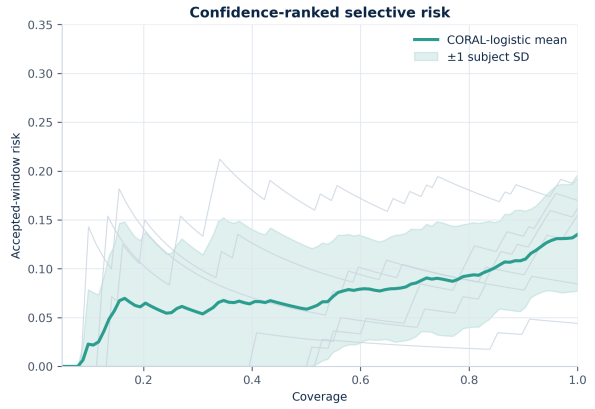

Figure 7 Subject-Level And Mean Risk–Coverage Curves For CORAL-Logistic. Lower Coverage Retained Only Higher-Confidence Decisions.

Feedback Scope And Study Limitations

The interface created 18 structured feedback cards, three per subject: one confident lower-burden prediction, one confident high-burden prediction, and one low-confidence candidate. Each record contained observed quality flags, hard-gate status, model confidence, and an approved action. The persistent-rail gate changed all three Subject 05 actions to contact checking and reacquisition, including its lower-burden and low-confidence cases. All 18 LLM-rendered messages preserved the rounded confidence value, expressed the encoded action in natural language, and passed the rule checker on the first pass; none was blocked. This is a safety and traceability test, not an evaluation of neurofeedback efficacy. A future user study would need prespecified feedback timing, controls, behavioral outcomes, and adverse-event monitoring consistent with reporting guidance [27].

Several anomaly-detection studies provide useful structural analogies for this narrow renderer. Retrieved normal prototypes can ground OpenStack failure explanations [60], while retrieval-grounded HDFS narratives preserve a deterministic path from anomaly evidence to text [61]. LogBERT-style masked pretraining offers a session-level representation baseline [62], and a command-safety-checked DevOps copilot demonstrates that retrieval still needs policy verification [63]. Conformal alert control can couple a score to an evidence-grounded ticket [64], whereas language-guided feature selection in intrusion detection keeps model inputs auditable [65]. System-call analysis can localize suspicious windows before producing a forensic narrative [66], and evidence-ranked claim verification provides a further example of separating claim selection from prose [67]. None of these results transfers performance to EEG; together they motivate the same architectural discipline: localize the measured problem first, then render a constrained explanation whose claims can be checked.

Rubric-grounded educational feedback [68], cost-aware routing across parsing, diagnosis, and summarization tasks [69], and counterfactual credit explanations [70] are more distant domain examples, but each reinforces a general requirement: a generated rationale should be tied to an explicit score, evidence set, or action policy. In this study, the burden indicators and approved action vocabulary play that role, and the verifier rejects language that implies unsupported cognition, diagnosis, or treatment.

The main limitation is decisive: activity boundaries are absent. The dataset description establishes that six activities occurred [1], but neither the all-zero marker column nor the continuous samples identifies when they occurred. Dividing a recording into six assumed blocks would manufacture labels and could confound time, drift, and activity. The present labels are instead derived from recorded signal properties and support only technical readiness conclusions. Electrode locations are also not encoded in the files, so spatial claims about sensorimotor cortex are unwarranted. Finally, six subjects provide a stringent LOSO design but limited population coverage. Confidence intervals and fold-level results should accompany averages, and external validation is required before deployment.

The SQI target has a second limitation: it is an operational rule set rather than an independent expert consensus label. Learned models approximate that rule set across hardware and subjects from compact descriptor input, and selective prediction adds a safety valve where the approximation is uncertain. High balanced accuracy would not prove that every accepted segment suits every downstream BCI algorithm. Applications may tolerate different drift, noise, or clipping levels. Per-rule burden, raw traces, and fold predictions should therefore accompany aggregate scores, allowing later threshold changes or comparison with blinded human ratings.

Table 11 Acquisition-Regime And Within-Recording Time Audit. Time Tertiles Were Computed Separately For Each Subject.

Stratum	n	Upper quartile	CORAL bal. acc.	Error	Mean burden
Cyton	210	25.2%	0.899	0.095	1.024
Cyton+Daisy	296	25.3%	0.746	0.176	0.719

Stratum	n	Upper quartile	CORAL bal. acc.	Error	Mean burden
Time: early	169	36.1%	0.817	0.166	1.077
Time: middle	166	19.3%	0.870	0.114	0.792
Time: late	171	20.5%	0.728	0.146	0.670

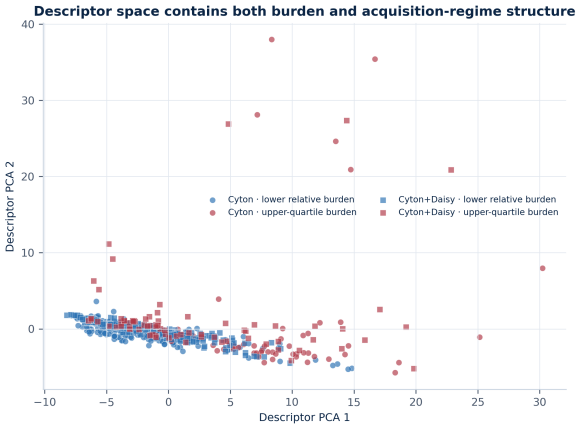

Figure 8 PCA Projection Of All 506 Descriptor Maps, Showing Both Burden And Board-Regime Structure. Axes Have No Task-Semantic Interpretation.

Table 12 Language-Feedback Test Using Gpt-5.6-Terra. All 18 Messages Preserved The Evidence Percentage, Used The Encoded Action, And Passed The Content Rules.

Approved action	Cards	Hard-gated	Verifier pass	Saved LLM output
Continue	5	0	yes	The recording-quality confidence is 100%; this segment can continue.
Remain still & repeat	4	0	yes	The recording-quality confidence is 100%. Please remain still, check the headset, and repeat the segment.
Low-confidence repeat	5	0	yes	The recording-quality confidence is 52%, below the acceptance threshold. Please record another five-second segment.
Check contact	4	3	yes	The recording-quality confidence is 100%. Please check electrode contact and record another five-second segment.

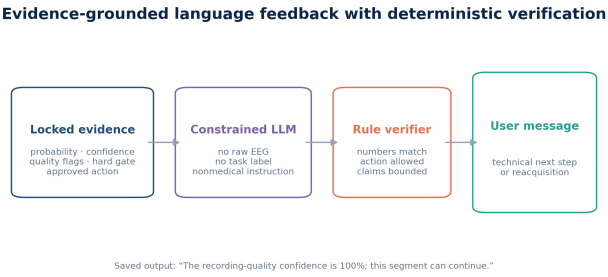

Figure 9 The LLM Received A Locked Evidence Object; A Deterministic Verifier Checked Numerical, Action, And Scope Fidelity Before Release.

D. CONCLUSIONS

The OpenBCI multidirectional cursor collection supports a complete empirical study of cross-subject EEG signal readiness, but not a defensible six-activity decoding experiment in its current form. Audit of all 450,651 rows revealed mixed eight- and 16-channel acquisition schemas, zero-valued markers, and concentrated rail saturation. Harmonization to eight channels at 125 Hz and five-second windows enabled consistent LOSO comparison of a classical model, an EEGNet-style descriptor CNN, a compact Conformer, masked self-supervision, CORAL, and eight-window fine-tuning. Calibration and selective prediction convert uncertainty into an explicit reacquisition or calibration action, while the structured language layer keeps feedback technical, traceable, and nonmedical. The resulting benchmark separates three questions that should not be conflated: whether a segment is technically usable, whether a model generalizes to a new subject, and whether a cognitive task can be decoded. Only the first two are testable from these files. Event timestamps or a documented task schedule would be necessary to address the third.

E. REFERENCES

[1] R. Akther, S. Rafique, Z. Saif, S. T. Roja, and K. Islam, "EEG MI-BCI multidirectional mouse cursor dataset (Open-BCI Ultra cortex mark-IV)," Mendeley Data, ver. 2, Aug. 5, 2025, doi: 10.17632/rz2jhpvhp3.2.

[2] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," Proc. IEEE, vol. 89, no. 7, pp. 1123-1134, Jul. 2001, doi: 10.1109/5.939829.

[3] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," Clin. Neurophysiol., vol. 113, no. 6, pp. 767-791, Jun. 2002, doi: 10.1016/S1388-2457(02)00057-3.

[4] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans," Proc. Natl. Acad. Sci. U.S.A., vol. 101, no. 51, pp. 17849-17854, Dec. 2004, doi: 10.1073/pnas.0403504101.

[5] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects," NeuroImage, vol. 37, no. 2, pp. 539-550, Aug. 2007, doi: 10.1016/j.neuroimage.2007.01.051.

[6] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI Competition IV datasets 2a and 2b," Front. Neurosci., vol. 6, Art. no. 39, Mar. 2012, doi: 10.3389/fnins.2012.00039.

[7] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update," J. Neural Eng., vol. 15, no. 3, Art. no. 031005, Jun. 2018, doi: 10.1088/1741-2552/aab2f2.

[8] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," Hum. Brain Mapp., vol. 38, no. 11, pp. 5391-5420, Nov. 2017, doi: 10.1002/hbm.23730.

[9] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," J. Neural Eng., vol. 15, no. 5, Art. no. 056013, Oct. 2018, doi: 10.1088/1741-2552/aace8c.

[10] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG Conformer: Convolutional transformer for EEG decoding and visualization," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 31, pp. 710-719, 2023, doi: 10.1109/TNSRE.2022.3230250.

[11] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for EEG-based motor imagery classification," IEEE Trans. Ind. Informat., vol. 19, no. 2, pp. 2249-2258, Feb. 2023, doi: 10.1109/TII.2022.3197419.

[12] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical EEG signals with self-supervised learning," J. Neural Eng., vol. 18, no. 4, Art. no. 046020, 2021, doi: 10.1088/1741-2552/abca18.

[13] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," Front. Hum. Neurosci., vol. 15, Art. no. 653659, Jun. 2021, doi: 10.3389/fnhum.2021.653659.

[14] C. Yang, M. B. Westover, and J. Sun, "BIOT: Biosignal transformer for cross-data learning in the wild," in Advances in Neural Information Processing Systems 36, 2023, pp. 78240-78260, doi: 10.52202/075280-3420.

[15] W.-B. Jiang, L. Zhao, and B.-L. Lu, "Large brain model for learning generic representations with tremendous EEG data in BCI," in Proc. 12th Int. Conf. Learn. Representations (ICLR), 2024. [Online]. Available: https://openreview.net/forum?id=QzTpTRVtrP

[16] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," IEEE Comput. Intell. Mag., vol. 11, no. 1, pp. 20-31, Feb. 2016, doi: 10.1109/MCI.2015.2501545.

[17] P. L. C. Rodrigues, C. Jutten, and M. Congedo, "Riemannian Procrustes analysis: Transfer learning for brain-computer interfaces," IEEE Trans. Biomed. Eng., vol. 66, no. 8, pp. 2390-2401, Aug. 2019, doi: 10.1109/TBME.2018.2889705.

[18] H. He and D. Wu, "Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach," IEEE Trans. Biomed. Eng., vol. 67, no. 2, pp. 399-410, Feb. 2020, doi: 10.1109/TBME.2019.2913914.

[19] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Learning invariant representations from EEG via adversarial inference," IEEE Access, vol. 8, pp. 27074-27085, 2020, doi: 10.1109/ACCESS.2020.2971600.

[20] V. Jayaram and A. Barachant, "MOABB: Trustworthy algorithm benchmarking for BCIs," J. Neural Eng., vol. 15, no. 6, Art. no. 066011, Dec. 2018, doi: 10.1088/1741-2552/aadea0.

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Mach. Learn., vol. 70, 2017, pp. 1321-1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[22] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," J. Mach. Learn. Res., vol. 11, pp. 1605-1641, 2010. [Online]. Available: https://jmlr.org/papers/v11/el-yaniv10a.html

[23] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in Proc. 36th Int. Conf. Mach. Learn., vol. 97, 2019, pp. 2151-2159. [Online]. Available: https://proceedings.mlr.press/v97/geifman19a.html

[24] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," Found. Trends Mach. Learn., vol. 16, no. 4, pp. 494-591, 2023, doi: 10.1561/2200000101.

[25] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," in Advances in Neural Information Processing Systems 33, 2020, pp. 3581-3591. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

[26] I. Gibbs and E. Candès, "Adaptive conformal inference under distribution shift," in Advances in Neural Information Processing Systems 34, 2021, pp. 1660-1672. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html

[27] T. Ros et al., "Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist)," Brain, vol. 143, no. 6, pp. 1674-1685, Jun. 2020, doi: 10.1093/brain/awaa009.

[28] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, and K. Roberts, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. Gaithersburg, MD, USA: National Institute of Standards and Technology, Jul. 2024, doi: 10.6028/NIST.AI.600-1.

[29] World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland: World Health Organization, 2021, ISBN 978-92-4-002920-0. [Online]. Available:

https://www.who.int/publications/i/item/9789240029200

[30] Q. Wu, G. Mi, and D. Wood, "Calibration-Light Subject-Independent Motor Imagery BCI via Self-Supervised Pretraining and Conformer," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239-262, Apr. 2025, doi: 10.51903/jtie.v5i1.493.

[31] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 258, pp. 559-567, May 2025.

[32] J. Zhang, "From General Human Activity Recognition to Volleyball-Oriented Wearable Transfer Learning: Cross-Dataset Evidence from UCI HAR and WISDM for Domain Adaptation and Edge Deployment," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 263-283, Apr. 2025, doi: 10.51903/jtie.v4i1.524.

[33] S. He, X. Chang, and E. Sun, "Cross-Cloud Transfer Learning for AI Training Capacity Forecasting under Workload and Topology Distribution Shift," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 100-120, Jan. 2024, doi: 10.69987/JACS.2024.40108.

[34] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215-238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[35] G. Mi, T. Ye, and D. Wood, "A Lightweight Medical Foundation Model for Cross-Modal Multi-Task Pretraining and Parameter-Efficient Few-Shot Transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572-589, 2025, doi: 10.51903/jtie.v4i3.492.

[36] S. Lu and T. Zou, "Uncertainty-Aware Medical Vision-Language Classification on a Lightweight MedMNIST-Compatible Biomedical Patch Benchmark," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 1-19, 2026, doi: 10.51903/jtie.v5i2.530.

[37] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 31-47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[38] J. Jin, T. Huang, and S. Lu, "A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 74-85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

[39] S. He, H. Tu, and I. Liu, "Safe PD Capacity Forecasting with Time-Series Foundation Models and Calibrated Uncertainty for Heterogeneous GPU Clusters," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 48-66, Apr. 2023, doi: 10.69987/JACS.2023.30404.

[40] S. Chen, S. He, and E. Sun, "Risk-Bounded GPU Resource Oversubscription via Conformal Demand Envelopes in Production AI Clusters," J. Adv. Comput. Syst., vol. 4, no. 5, pp. 119-134, May 2024, doi: 10.69987/JACS.2024.40509.

[41] S. Zhao, J. Bai, and D. Roberson, "Multi-Horizon GPU Demand Forecasting with Workload Semantics and Operational Risk Curves: An Empirical Study on Alibaba Clusterdata GPU Trace," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 544-571, Dec. 2025, doi: 10.51903/jtie.v4i3.498.

[42] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306-324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[43] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[44] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520-533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[45] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502-520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[46] B. Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[47] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 97-111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[48] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 67-82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[49] D. Zheng and C. Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 83-99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[50] D. Zheng, C. Li, and H. Davidson, "Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation," J.

Adv. Comput. Syst., vol. 3, no. 2, pp. 35-49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[51] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[52] Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-Aware Medical Image Explanation Cards: LLM-Generated Visual Explanations for AI-Assisted Radiology Interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415-436, Oct. 2025, doi: 10.51903/ijgd.v3i2.3616.

[53] B. Zhou, H. Wang, and X. Chang, "Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Design Framework for Rubric-Based Medical Risk Communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3696.

[54] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179-185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[55] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216-240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[56] Y. Li, S. Lu, and L. Zhao, "LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[57] G. Liu, S. He, and H. Wong, "LLM-Compatible Visual Brief Cards for AI Infrastructure Capacity Dashboards: A UI/UX Framework for Turning Forecast Risk into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-213, May 2025, doi: 10.51903/ijgd.v3i1.3723.

[58] X. Sun, Y. Lu, and J. Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting," J. Adv. Comput. Syst., vol. 3, no. 8, pp. 9-24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[59] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," J. Inf. Syst. Informatics, vol. 7, no. 3, pp. 2182-2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.

[60] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, Nov. 2025, doi: 10.18196/eist.v6i2.31232.

[61] X. Sun, Z. S. Zhong, and Q. Wu, "Retrieval-Grounded HDFS Log Anomaly Detection and Deterministic Failure Narrative Generation," J. Comput. Syst. Appl., vol. 3, no. 1, pp. 15-30, Jun. 2026, doi: 10.64229/j6d7fr94.

[62] Q. Xin, "Self-Supervised Log Anomaly Detection with LogBERT-Style Transformers: Full Empirical Evaluation on a Reproducible SynHDFS Benchmark," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 23-35, Jun. 2026, doi: 10.54732/jeecs.v11i1.3.

[63] B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-Style DevOps Copilot for Cloud-Native Troubleshooting: Retrieval-Augmented Runbook Generation and Command-Safety Evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104-118, Aug. 2026, doi: 10.51903/jtie.v5i2.534.

[64] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," AVITEC, vol. 8, no. 2, pp. 247-264, May 2026, doi: 10.28989/avitec.v8i2.3974.

[65] Y. Li and S. Lu, "Language-Guided Feature Selection for DDoS and Intrusion Detection on CICIDS2017," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 284-305, Apr. 2025, doi: 10.51903/jtie.v4i1.531.

[66] Q. Xin, "Host-Based Intrusion Detection with System Call Sequences: Window Localization and Forensic Narratives," AVITEC, vol. 8, no. 2, pp. 325-334, Jun. 2026, doi: 10.28989/avitec.v8i2.3973.

[67] W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327-340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[68] Q. Xin, "Auditable Automated Essay Scoring and Formative Feedback: A Rubric-Grounded Pipeline for Secondary and Higher Education," J. Appl. Artif. Intell. Educ., vol. 2, no. 1, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[69] C. Li, G. Liu, and Z. Zhao, "Cost-Aware LLM-Style Routing for AIOps Log Analysis: Log Parsing, Anomaly Detection, Fault Diagnosis, and Incident Summarization on LogEval Task Files," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 91-103, Jun. 2026, doi: 10.51903/jtie.v5i2.538.

[70] Q. Xin, "Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," J-INTECH, vol. 14, no. 2, pp. 215-231, Jun. 2026, doi: 10.32664/j-intech.v14i02.2228.