TECHNICAL PREPRINT • VERSION 1.2

CORTEX

A three-year engineering study of low-cost EEG forecasting and personal brain–computer interface software

From predictive-signal experiments to a direct-Muse macOS research platform

Connor Colyer

Independent researcher and software engineering student, United Kingdom

ConnorColyer@icloud.com • github.com/ConnorColyer

Research period: November 2023 – present

Preprint issued: 27 July 2026

A note on the project

CORTEX has changed constantly over three years, so this paper follows the project as it actually developed instead of pretending it was one fixed experiment.

I started with a predictive EEG equation, then built experimental Python pipelines, several Swift prototypes and finally a native macOS research application. The current version connects directly to a Muse 2, checks signal quality, guides personal training, records sessions and replays them. Each stage taught me something different, and I have kept that progression visible throughout the paper.

The percentages in this paper come from different tests on different versions. I have kept them as separate snapshots and used the most complete tables from each report. I have not yet brought every original 2025 run file into one public release, so the paper says clearly which results come from those reports and which parts I can run directly in the current codebase.

WHAT CORTEX IS

CORTEX is practical BCI research: forecasting EEG patterns and recognising a small set of trained personal states. It is not a medical device, and I am not claiming that four electrodes can read arbitrary thoughts.

Contents

Abstract

CORTEX is a three-year independent research and software-engineering project investigating whether low-cost, four-channel electroencephalography (EEG) can support short-horizon signal forecasting and personalised brain–computer interaction. The project progressed from mathematical notes and waveform extrapolation experiments to four-state classification studies and a native macOS platform that acquires Muse 2 data directly over Bluetooth Low Energy. Different versions produced different results: an April 2025 table reported 92.6% ± 3.5% held-out four-state accuracy; a May 2025 engineering report recorded macro-F1 of 0.907 across ten stratified folds; and other versioned tests reported 94.6% and 95.0%. The current software adds direct four-channel acquisition, inertial and physiological context, causal windowing, configurable filtering, label-specific readiness, quality checks, an Unknown fallback, SQLite session storage and replay. Across the three years, the project moved from asking whether an EEG waveform could be predicted to building a complete environment in which predictions can be captured, tested, rejected when the signal is poor and replayed afterwards. This paper brings that work together and sets out the next experiments needed to test the current system across complete sessions.

Keywords: electroencephalography; brain–computer interface; Muse 2; wearable EEG; signal forecasting; personal classification; Swift; macOS; reproducibility.

Executive summary

CORTEX is an independent, longitudinal attempt to make low-cost EEG useful for anticipatory and responsive human–computer interaction.

The project began in November 2023 as a question: could patterns in an electroencephalographic stream be transformed into a representation that supports short-horizon forecasting rather than only retrospective classification? Over three years, that question produced mathematical notes, research papers, synthetic-signal experiments, Muse-based prototypes, comparative evaluations, an engineering report and a native macOS application.

The work uses a consumer Muse 2 headset with four principal EEG channels—TP9, AF7, AF8 and TP10. Earlier versions explored band-derived waveforms, smoothing, trend features, extrapolation and a proposed Colyer Equation. Later versions broadened the problem into four-state classification, latency and usability testing. The current software moves away from sweeping prediction claims and towards a reliability-first personal BCI: it acquires Muse packets directly over Bluetooth Low Energy, cleans the signal, extracts spectral and quality features, blocks inference when the data is unsuitable, trains user-specific classifiers and records sessions for replayable inspection.

The archived evaluation record contains several results from different versions. A March 2025 summary reports 94.6% two-second EEG forecasting accuracy. An April 2025 long-form paper contains a held-out classification table reporting 92.6% ± 3.5%, with traditional classifier baselines between 48.7% and 58.9%; narrative passages in the same paper refer to 95.0%. A May 2025 engineering evaluation reports mean macro-F1 of 0.907 across ten stratified folds on 120 labelled segments, plus a 30-trial latency series with a reported mean of 1,475.8 ms. These results are retained as separate snapshots because the models, datasets and protocols evolved.

CORE CONTRIBUTION

The project is bigger than one percentage. Its main contribution is the progression from an ambitious BCI idea into mathematics, signal experiments, comparison studies, native acquisition software, quality controls, replayable sessions and a clear route into the next round of testing.

Current project at a glance

Table 1. Current implementation summary.

Area	Current position
Platform	Native macOS application written in Swift and SwiftUI
Headset	Muse 2 via direct Bluetooth Low Energy
EEG	Four channels: TP9, AF7, AF8 and TP10
Additional streams	Accelerometer, gyroscope, telemetry and up to three PPG channels
Windowing	1.0 s windows with 250 ms hop updates
Processing	Detrend, configurable notch, band-pass, normalisation and real FFT features
Training	Guided, user-specific state and intent labels with readiness tracking
Safety behaviour	Hard and soft quality blockers; Unknown fallback when confidence is not earned
Storage	.musesession packages with SQLite, CSV/JSON exports and replay
Maturity	Working research software; not a medical device or clinically validated decoder

1. Origin, problem and scope

1.1 The original problem

Brain–computer interfaces translate measured brain activity into a communication or control channel, with performance depending on the signal, training method, translation algorithm and evaluation procedure [1]. Many practical BCI systems are reactive: they detect or classify a signal after a sufficiently clear pattern has developed. In assistive interaction, even modest delays can make control feel disconnected. CORTEX began by asking whether the recent history of an EEG stream contains enough structure to estimate a near-future continuation, or at least to recognise a preparatory cognitive state early enough to improve response.

That question is difficult because scalp EEG is low-amplitude, non-stationary and contaminated by eye movement, muscle activity, motion, contact changes and mains interference. Motor-imagery paradigms often rely on event-related synchronisation and desynchronisation in rhythmic activity [2], but a four-channel consumer headset adds severe spatial constraints. CORTEX therefore treats the problem as an engineering system rather than an equation in isolation: acquisition, time alignment, filtering, feature extraction, prediction, confidence, user calibration and output gating all affect whether a result is meaningful.

1.2 Scope

- Low-cost, non-invasive EEG acquired from a Muse 2 headset.
- Short-horizon waveform forecasting and personalised cognitive-state classification.
- Real-time or near-real-time operation on ordinary Apple hardware.
- User-specific calibration rather than an assumption of universal brain signatures.
- Local-first recording, replay and audit of what the system received and inferred.
- Assistive and human–computer interaction research, not diagnosis or treatment.

1.3 Non-claims

I am not claiming that four scalp electrodes can read arbitrary thoughts, or that the project has clinical or population-wide results. “Predicting thought before it occurs” captures the ambition that started CORTEX, but the technical goal is more specific: forecast short-term signal structure and recognise a small set of deliberately trained states or intents.

2. Research questions and design principles

2.1 Research questions

12. Can a compact representation of recent EEG history support useful short-horizon prediction?
13. Can a consumer four-channel headset distinguish a small set of deliberately trained cognitive tasks for one user?

14. Which preprocessing and quality checks are necessary before a prediction should be shown or acted upon?
15. Can the complete loop—headset, acquisition, processing, inference, interface and recording—operate responsively on a normal Mac?
16. How should confidence, uncertainty and Unknown states be exposed so the system declines unreliable inferences?
17. What evidence package is required to separate a promising personal experiment from a generalisable scientific claim?

2.2 Design principles

Table 2. Principles distilled across the research and current codebase.

Principle	Engineering consequence
Direct acquisition	Capture the Muse protocol directly where practical; retain packet-level evidence.
Causal processing	A prediction at time t may use only information available at or before t.
Quality before confidence	Block inference when signal, packet or motion conditions are unsuitable.
Personal calibration	Treat readiness as label- and user-specific rather than assumed.
Replayability	Record enough raw and derived state to reproduce an inference path.
Modality honesty	Keep trusted measures separate from experimental proxies.
Conservative output	Use Unknown when the system has not earned a stable label.
Versioned evidence	Attach every metric to its dataset, protocol, code version and evaluation date.

3. Three-year development chronology

CORTEX progressed through overlapping research and software tracks; version labels indicate artefact lineage, not a single linear production release.

Table 3. Consolidated chronology. Dates derive from document metadata, source headers and project structure.

Period	Stage	What changed	Evidence
Nov 2023	Concept and research start	Predictive EEG question, early literature review and mathematical framing.	CORTEX Summary
Oct–Dec 2024	Visualisation and v0.1	Swift visualisation experiments and first CoreBluetooth headset work.	EEGVisualization; Cortex v0.1

Period	Stage	What changed	Evidence
Jan 2025	v0.2–v0.3	Muse SDK integration experiments and device/data display work.	Cortex v0.2/v0.3 source
Feb–Mar 2025	v0.4 and equation paper	Waveform smoothing, predictive transformation, equations and early forecasting claims.	CORTEX Summary; v0.4
Apr 2025	Long-form evaluation	Held-out classifier table, baselines, per-task metrics, AUC and signal-error comparison.	CORTEX FuLL copy
May 2025	Engineering evaluation	Acceptance tests for four-state classification, latency, setup, stability and privacy.	Blank 4 copy
Jun–Jul 2025	CTX expansion	Broader architecture, ethics, multimodal concepts and a two-second Swift simulation.	CTX paper and code
2026	Native research platform	Direct BLE, EEG/IMU/PPG/telemetry, session database, replay, quality gates and Prediction Stack V3.	MuseScope/MuseWire
2026 onward	Formal generalisation	A broader causal forecasting system and thesis workspace, with experiments still to be completed.	Colyer Thesis Workspace

3.1 What continuity means

The name CORTEX connects the artefacts, but continuity does not mean implementation identity. The early forecasting equation, the April classifier experiments, the May acceptance tests, the July CTX simulation and the current Prediction Stack V3 are related research iterations. Their metrics cannot be transferred between versions without rerunning the relevant experiment.

4. Hardware and direct data acquisition

4.1 Muse 2

The current hardware target is the Muse 2. CORTEX uses four principal EEG positions—TP9, AF7, AF8 and TP10—with a nominal 256 Hz stream described in the archived reports. Published work has shown that Muse-class hardware can capture quantifiable event-related potentials, while also noting constraints from electrode location, signal contact and Bluetooth timing [3]. Comparative testing has also found lower test–retest reliability and greater artefact susceptibility in consumer systems than in medical-grade equipment [4]. The headset additionally exposes inertial, telemetry and photoplethysmography streams under suitable presets. These modalities are valuable chiefly because they help explain contamination and context: motion can invalidate an EEG window; PPG and telemetry may support session quality inspection; neither should be silently interpreted as direct evidence of a thought.

4.2 Acquisition lineage

The earliest Swift prototype scanned for a named Muse peripheral using CoreBluetooth and attempted to subscribe to an EEG characteristic. Subsequent versions explored the vendor SDK. The current implementation returns to direct BLE access on macOS, inventories relevant characteristics, applies known-safe control writes, subscribes to streaming characteristics and logs unknown packets in Explorer mode without blind undocumented writes.

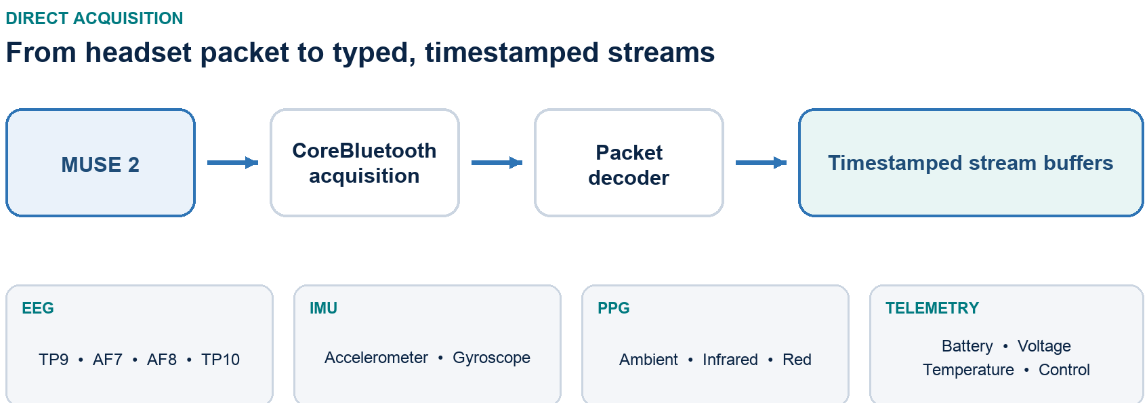


Figure 1. Current direct-acquisition path and decoded Muse 2 streams.

4.3 Captured sample artefact

The current project directory includes a direct EEG CSV containing 7,644 samples over approximately 29.9 seconds, with host time, packet index, within-packet sample index and the four EEG channels. This artefact confirms that the modern toolchain is not limited to synthetic UI data. It is not, by itself, a labelled evaluation dataset.

CURRENT WORKING EVIDENCE

The project includes a direct EEG capture, the current Swift source, compiled application bundles and the complete session and test code. The older percentages are taken from the versioned reports described in Section 7.

5. Signal-processing pipeline

5.1 Historical pipeline

The archived engineering report describes a real-time pipeline comprising BLE ingest, a two-second ring buffer, preprocessing, feature extraction, prediction, a confidence gate, display/API output and an artefact-removal side path. The signal-processing flow image below is preserved as a historical architecture figure; the current application has a different internal implementation.

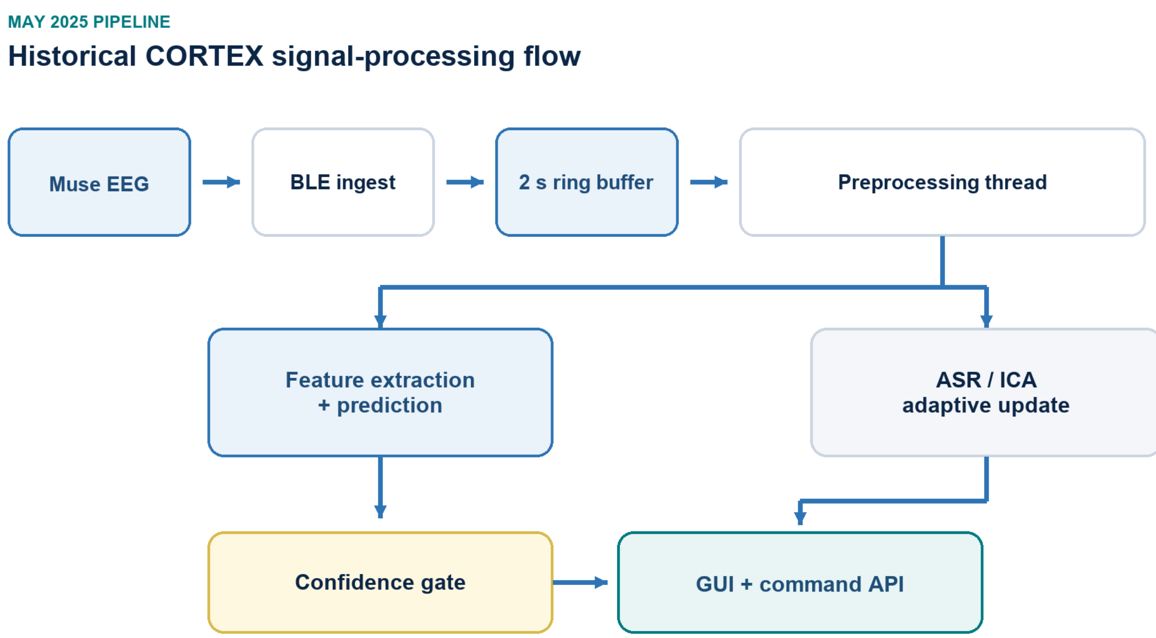


Figure 2. Historical CORTEX EEG signal-processing flow from May 2025.

5.2 Current Prediction Stack V3

Prediction Stack V3 works on 1.0-second EEG windows with 250 ms hop updates. The preprocessor detrends each channel, applies line-noise suppression at 50 or 60 Hz, filters the usable EEG range, normalises the window and computes a real Fourier spectrum. Feature extraction combines spectral information with quality and available modality context. State and intent heads are trained separately, and each label carries readiness state rather than becoming active simply because a model can emit a probability.

18. Schedule a complete 256-sample window and advance by a stable 64-sample hop.
19. Detrend and filter each available EEG channel.
20. Estimate band power and other time/frequency features.
21. Evaluate per-channel and whole-window quality.
22. Reject or downgrade windows affected by hard or soft blockers.

23. Construct the feature vector using only available modalities.
24. Evaluate trained state and intent models where readiness permits.
25. Return Unknown or an unlocked result when evidence is insufficient.
26. Record the event and allow explicit confirm, relabel or discard feedback.

5.3 Quality reasons

Table 4. Current quality-control model.

Category	Examples	Effect
Hard blockers	Low signal; packet dropout	The window cannot produce a lock-eligible inference.
Soft blockers	Line noise; motion contamination; camera unreliability; artefact burst	Confidence or eligibility is reduced according to label policy.
Label-aware exceptions	For example, a blink label may permit a burst that a resting-state label rejects.	Quality policy follows what the label is intended to capture.
Availability flags	EEG, IMU, camera, PPG and telemetry tracked separately	Feature extraction degrades gracefully when optional streams are absent.

6. The Colyer predictive framework

6.1 Historical formulation

The historical Colyer Equation was presented as a transformation of multi-band EEG into a smoother, lower-dimensional waveform followed by trend-sensitive extrapolation. Drafts describe adaptive feature weighting, temporal error correction, normalisation by signal energy and cubic-spline interpolation. The intended purpose was to generate a continuous short-horizon signal estimate without the computational cost of a large deep network.

The equation changed as the project developed, and that is worth showing rather than hiding. For the next release I will choose one canonical form, define its inputs and units, implement it in a small reference package and tie each new result to the exact code version that produced it.

6.2 Present interpretation

The broader 2026 Colyer thesis workspace generalises the original idea from EEG into causal forecasting of ordered numerical streams. It introduces signal space, latent state, memory, transition operators, rollout branches, baseline envelopes, regret and explicit limits on universal prediction claims. Its experiments chapter is presently a scaffold, so it should be treated as a current formalisation programme rather than evidence for the older EEG metrics.

SCOPE

This paper keeps the EEG project at the centre. The broader Colyer System is included as the direction the theory is now moving in, separate from the completed CORTEX experiments.

7. Experimental programme and versioned findings

The project contains several evaluation snapshots because the system, dataset and question changed as the work progressed.

7.1 Snapshot A — early two-second waveform forecasting

A March 2025 summary reports 94.6% accuracy in forecasting EEG values two seconds before their recording and contrasts the result with an approximately 55% contemporary baseline. The document explains the conceptual transformation and contains evaluation code fragments, but the rendered summary does not include a complete metric definition, dataset manifest or raw output sufficient to reproduce the percentage.

HOW I USE THIS RESULT

The 94.6% figure belongs to the early forecasting version. I keep it here as part of the project history rather than using it as the headline result for the current application.

7.2 Snapshot B — held-out classification and baseline comparison

The April 2025 long-form paper evaluates a Colyer-derived predicted signal mapped to four mental states through a classifier. Its main held-out results table reports 92.6% accuracy with a standard deviation of 3.5 percentage points. Traditional baselines are substantially lower in the same table.

Table 5. April 2025 held-out classification table as archived.

Method	Accuracy	Standard deviation
Colyer Equation + classifier	92.6%	3.5
LDA	55.2%	7.1
SVM	58.9%	6.5
Naive Bayes	48.7%	9.0
Chance level	25.0%	0.0

Per-task precision, recall and F1 values in the following table range from 0.93 to 0.97. A sentence immediately before the table incorrectly says that F1 ranges from 0.68 to 0.79; the table and the task-by-task prose agree with one another, so the lower range is treated as a drafting error.

Table 6. April 2025 per-task classification metrics.

Mental task	Precision	Recall	F1
Rest	0.96	0.97	0.96
Imagine left hand	0.94	0.93	0.93
Imagine right hand	0.95	0.94	0.94
Mental counting	0.95	0.96	0.95

The same paper reports pairwise AUC values from 0.92 to 0.99 for the Colyer-based method and 0.60 to 0.75 for the SVM baseline. One narrative sentence describes 0.88 as the highest Colyer AUC, which conflicts with the table and is marked for correction.

Table 7. April 2025 pairwise AUC table.

Pair	Colyer	SVM baseline
Rest vs left hand	0.98	0.72
Rest vs right hand	0.99	0.75
Rest vs counting	0.97	0.69
Left vs right hand	0.92	0.65
Left hand vs counting	0.93	0.60
Right hand vs counting	0.96	0.68

Signal-prediction error is tabulated as MSE 0.015 and RMSE 0.122 for the Colyer Equation, versus MSE 0.021 and RMSE 0.145 for an autoregressive model. A later paragraph gives MSE 0.005 and RMSE 0.071. I have used the table values here because they form the complete side-by-side comparison, and I will check the later pair when I rebuild that experiment.

Table 8. Signal-error comparison reported in the April 2025 experiment.

Method	MSE	RMSE
Colyer Equation	0.015	0.122
Autoregressive model	0.021	0.145

WHAT SURVIVES FROM THIS VERSION

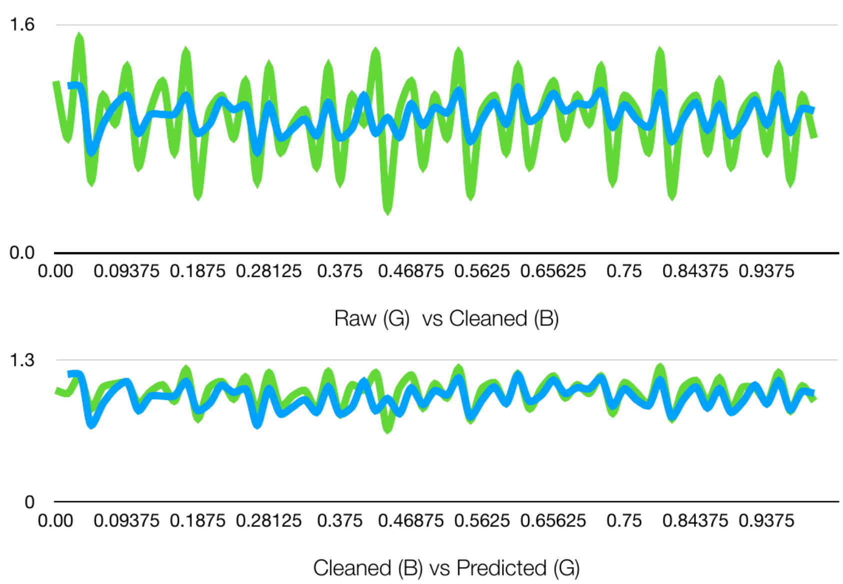
The April paper contains the full comparison tables and figures. The next step is to rerun the same comparison through the current experiment runner and publish the generated outputs beside it.

7.3 Snapshot C — 95.0% narrative result and waveform comparison

Discussion and conclusion sections of the April paper refer to 95.0% classification accuracy at a two-second horizon. This may represent a later run than the 92.6% table, but the document does not label the version boundary. Until the run is recovered, 95.0% should be described as a later reported result rather than silently substituted into Table 5.

underlying EEG patterns associated with different cognitive states, even when predicting several seconds into the future. Furthermore, the precision, recall, and F1-scores for individual mental tasks were consistently high, ranging from 0.93 to 0.97 for the Colyer Equation, indicating a very balanced and robust performance in identifying each state.

The AUC-ROC analysis for pairwise comparisons further supported the model's excellent discriminative power, consistently achieving values above 0.92 for the Colyer Equation, demonstrating a strong ability to distinguish between different mental states. While a specific prediction latency was not explicitly calculated due to the 2-second anticipatory nature of the prediction, the high accuracy achieved at this horizon underscores the model's capability for near real-time control. Finally, the lower signal prediction error (MSE of 0.005 and RMSE of 0.071) compared to a traditional AR model further indicates that the Colyer Equation is highly accurate in forecasting the EEG signal itself during the anticipatory window. These compelling findings collectively demonstrate that the Colyer Equation holds exceptional promise for real-time thought prediction using a low-cost, accessible EEG device like the Muse 2 headset.



68

Figure 3. Raw, cleaned and predicted waveform comparison from the April 2025 CORTEX experiment.

7.4 Snapshot D — ten-fold classification and engineering acceptance tests

The May 2025 engineering report documents a controlled calibration dataset of 120 labelled 200 ms EEG segments across Rest, Left-Hand Motor Imagery, Right-Hand Motor Imagery and Mental Arithmetic. It reports ten-fold stratified cross-validation with the same preprocessing and feature pipeline in every fold. The fold macro-F1 scores average 0.907.

Table 9. May 2025 ten-fold macro-F1 results; mean 0.907.

Fold	Macro-F1	Fold	Macro-F1
1	0.93	6	0.91
2	0.91	7	0.92
3	0.89	8	0.90
4	0.92	9	0.88
5	0.90	10	0.91

The same report records 30 latency observations from 1,355 ms to 1,599 ms and reports a mean of 1,475.8 ms. Direct calculation from the listed observations confirms the stated minimum, maximum and mean to rounding. The report labels 1,599 ms as the 95th percentile; under common percentile definitions the precise value should be recalculated from the raw timestamps, because 1,599 ms is also the maximum.

Table 10. May 2025 engineering acceptance tests.

Acceptance test	Reported result	What I have
End-to-end latency	30 trials; mean 1,475.8 ms; max 1,599 ms	The complete 30-value series is recorded in the report
Four-state classification	Mean macro-F1 0.907	All ten fold scores are recorded in the report
Setup	Five trials; mean 278.6 s	Values and mean internally consistent
One-hour stability	99.64% reported	The calculation contains a factor-of-ten error; rerun planned
Privacy checklist	All items passed	Checklist recorded in the engineering report
Unit-test coverage	Just over 87% reported	Coverage figure recorded in the engineering report

STABILITY CALCULATION

While consolidating the report I found a factor-of-ten error: a 200 ms interval over 3,600 seconds gives about 18,000 expected events, not 1,800. I have therefore left 99.64% out of the headline results and will replace it with a new timestamped soak test.

7.5 Snapshot E — current implementation

The current native application is the strongest engineering version so far, but I have not attached a new held-out percentage to it yet. It gives the next experiment a much better base: direct packet acquisition, deterministic preprocessing, explicit quality policy, model readiness, persistent session data and XCTest coverage of key behaviours. I can now follow the route from packet to output instead of treating the model as a black box.

8. Baselines, comparisons and interpretation

8.1 Why the baselines matter

A high percentage is only useful if the comparison is fair. The nominal chance rate is 25% for a balanced four-class task, but statistical significance depends on sample count and evaluation design rather than the nominal rate alone [5]. LDA, SVM and Naive Bayes test whether a more elaborate predictive representation earns its complexity. An autoregressive model tests whether short-horizon waveform performance exceeds a simple time-series continuation. Future evaluations should also include persistence, local linear extrapolation and a modern regularised classifier using identical causal splits.

8.2 What I take from the results

Table 11. Interpretation boundary.

What the results show	What I have not tested yet
Personal four-state results beat chance in several versions.	Performance across different people.
The April model beat the recorded conventional classifiers.	A fresh comparison inside one modern pipeline.
The May result held across ten folds.	Completely separate recording sessions.
The graph shows the cleaned signal and predicted continuation.	The same test rebuilt in the current codebase.
The current code rejects poor windows and can return Unknown.	A held-out benchmark for Prediction Stack V3.

8.3 Leakage and independence

The main methodological risk is not necessarily overlap of rows between folds; it is dependence between nearby segments from the same recording. If adjacent 200 ms windows from one session are randomly distributed across training and testing folds, the model may recognise session-specific structure rather than generalise to a future session. Model selection performed against the same cross-validation estimate can also bias the reported error; nested or genuinely independent evaluation is needed when tuning is involved [6]. The next evaluation should split by complete session or trial, fit preprocessing only on training data, then report both within-session and across-session performance.

9. Current native macOS implementation

9.1 Software scale and structure

The present MuseScope/MuseWire codebase contains approximately 10,786 lines of Swift across the principal source and test files inspected for this preprint. MuseScope is the graphical research application; MuseWire is a focused command-line capture utility. The design is deliberately headset-first and local.

Table 12. Current software components.

Component	Responsibility
MuseWire	Scanning, connecting and exporting direct four-channel Muse EEG to CSV.
MuseScope app	Discovery, acquisition profiles, live panels, calibration, recording, replay and prediction UI.
MuseMonitorModel	Central orchestration of BLE, decoded streams, calibration, storage and app state.
PredictionCore	Window scheduling, preprocessing, quality evaluation, feature extraction, training and inference.
MuseSessionStore	SQLite-backed recording, exports and replay frame reconstruction.
Prediction Store V3	Versioned labels, samples, normalisers, state/intent models and UI settings.
MuseScopeTests	Deterministic tests for scheduling, filtering, quality, readiness, persistence and replay.

9.2 Headset and physiology panels

The interface exposes raw and cleaned EEG, per-stream packet activity, channel health, band estimates, motion state, PPG waveforms, contact quality, telemetry and characteristic inventory. Physiology outputs are divided into trusted and experimental tiers. This separation is important: a computed heart-rate or interval measure with adequate signal support should not be visually equated with a speculative arousal-style proxy.

9.3 Guided capture and personal models

Training is guided rather than passively harvested. Built-in and custom labels define prompts, modality requirements and quality policies. The capture coordinator can retry steps when too few acceptable windows are gathered. Labels progress through untrained, training, ready and experimental states. The user can confirm, relabel or discard a recent prediction, keeping feedback explicit.

9.4 Current automated tests

- Stable 250 ms window-hop scheduling.
- Suppression of 60 Hz line noise while preserving alpha and beta energy.
- Hard and soft quality blockers, including label-aware artefact policy.
- Prediction Store V3 encode/decode round trip.
- Readiness blocking for under-trained labels and Unknown fallback.
- PPG and telemetry session replay round trip.
- Graceful feature extraction when PPG and telemetry are absent.

VERIFICATION NOTE

On 27 July 2026, MuseScope built successfully and all nine XCTest cases passed with zero failures under Xcode 27.0 on macOS. The run covered window scheduling, filtering, quality policy, readiness, persistence, replay and graceful feature degradation. This verifies the current test target; it does not reproduce the historical performance experiments.

10. Recording, replay and data architecture

A research system is easier to trust when an inference can be reconstructed. MuseScope records .musesession packages containing a SQLite database plus export artefacts. The store can retain raw packets, characteristic inventory, control responses, decoded EEG/IMU/PPG frames, telemetry, headset quality, physiology snapshots and user markers. Replay feeds recorded frames back through the application for inspection.

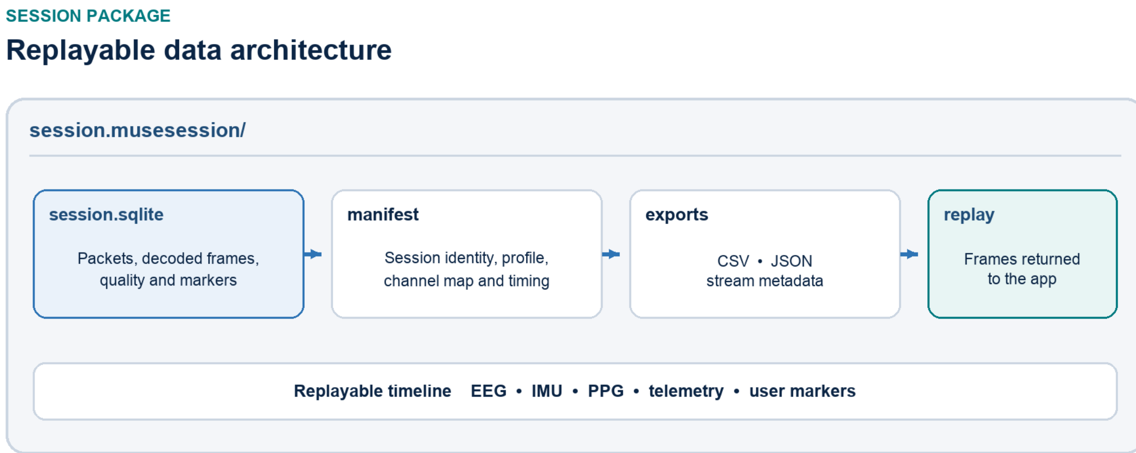


Figure 4. MuseScope session package and replayable data path.

For a public research release, each evaluated session should gain an immutable manifest: pseudonymous subject/session identifier, headset, firmware if available, acquisition profile, channel map, sampling cadence, task protocol, code commit, preprocessing configuration, split membership and checksum. The model output should be reproducible from the manifest and stored stream without manual UI actions.

11. Quality, safety and responsible inference

11.1 Refusal is a feature

The current software's most important conceptual improvement is that it can decline to lock a prediction. Earlier papers concentrated on accuracy and speed. Prediction Stack V3 makes suitability explicit: low signal or packet dropout can block inference; motion, line noise, camera reliability and bursts can alter policy; under-trained labels remain unready; Unknown remains a legitimate output.

11.2 State and intent separation

A broad state such as rest or focus is not equivalent to an intentional command. The current model separates state and intent heads so that an apparent cognitive condition does not automatically trigger an action. Any future control integration should add temporal confirmation, cooldown and an explicit undo path.

11.3 Safety boundary

- No safety-critical device should act on a single unconfirmed CORTEX output.
- Predictions should expose quality, readiness and uncertainty beside the label.
- The application should default to local processing and retain an audit trail.
- Experimental physiology must remain visually and semantically distinct from validated measures.

- Research participants must be able to stop collection and delete their local session.

12. Privacy, ethics and accessibility

12.1 Brain data is sensitive data

Even when a four-channel headset cannot reveal arbitrary private thoughts, raw neurophysiological recordings deserve strong protection. They can encode identity, artefact patterns and task-related information. The design direction is therefore local-first: no automatic cloud upload, explicit export, clear session ownership and minimal collection.

12.2 Consent and public evidence

Before any participant-derived data or stakeholder quotation is published, the release must confirm that the participant was real, consent covered the intended use and the artefact is adequately de-identified. Historical reports contain personas and stakeholder-style quotations that should not be presented as empirical testimony unless their provenance is confirmed.

12.3 Accessibility goal

CORTEX is motivated by reducing friction in human–computer interaction, especially where conventional input is difficult. Accessibility is not achieved by optimistic model output. It requires setup that a non-specialist can understand, visible connection and quality state, alternatives to colour-only status, readable typography, retryable calibration and a predictable failure mode.

13. What I am testing next

The next stage is to turn the current application into a repeatable experiment runner and test it across complete sessions.

Table 13. Consolidation gaps and publication actions.

Next question	Why it matters	What I will do
Which result belongs to which version?	The 92.6%, 94.6% and 95.0% figures came from different tests.	Give every new run an experiment ID, configuration and code commit.
Does it generalise across sessions?	Nearby windows from one recording can be too similar.	Train and test on complete, separate sessions.
How should waveform forecasts be scored?	A single word such as “accuracy” hides the size and shape of the error.	Report the horizon, units, tolerance curve, MAE, MSE, RMSE and R^2.

Next question	Why it matters	What I will do
How stable is a full-hour capture?	The old report contains a factor-of-ten event-count error.	Run a new timestamped soak test and publish the log.
How well does Prediction Stack V3 perform?	The current software is substantially different from the 2025 versions.	Freeze a release candidate and run the full protocol.
Does it work for anyone else?	The work so far is centred on personal calibration.	Start with repeated personal sessions, then design a consented multi-participant study.
Which parts of the method matter most?	A simpler model may explain the same result.	Run ablations and identical causal baselines.
How does it sit beside current BCI work?	This paper uses a focused methodological bibliography.	Expand the review for a specialist-journal submission.

14. Reproducibility and release plan

14.1 Proposed release bundle

27. Freeze a CORTEX v1 research tag and record the exact Swift toolchain and macOS version.
28. Export a clean, minimal reference implementation of the canonical predictive method.
29. Create a de-identified dataset manifest and separate training, validation and test sessions.
30. Provide scripts that regenerate every table and graph from raw or derived data.
31. Run leakage checks, grouped splits and simple causal baselines.
32. Publish a machine-readable metrics file alongside the narrative paper.
33. Run the full MuseScope XCTest target and retain the result bundle.
34. Render a release PDF, proof every figure and remove unverified participant claims.

14.2 Minimum evaluation protocol

Table 14. Recommended minimum protocol.

Element	Minimum requirement
Unit of independence	Whole trial or session, never adjacent windows split at random across train/test
Horizons	State the exact sample and time horizon for every result
Baselines	Chance, majority, persistence, local linear, AR and regularised classifier as appropriate
Metrics	Macro-F1 and confusion matrix for classes; MAE/MSE/RMSE/R^2 and tolerance curve for signals
Uncertainty	Confidence intervals or repeated-session distribution, not only one mean

Element	Minimum requirement
Ablation	Remove key Colyer features to show which elements earn their complexity
Quality policy	Report results before and after blockers, plus rejection coverage
Reproduction	One command regenerates the published metrics and plots

15. Next research programme

15.1 Immediate engineering work

- Split the 8,000-line application file into acquisition, storage, prediction and presentation modules.
- Add an experiment runner that bypasses UI state and operates on recorded sessions.
- Write a versioned feature schema so older recordings can be replayed under newer models.
- Add export of prediction probabilities, blockers, readiness and ground truth per window.
- Create deterministic tests for split construction, metric calculation and time alignment.

15.2 Scientific work

- Repeat the four-task study across multiple sessions from the same user.
- Use session-held-out validation and compare it with the historical random-fold result.
- Separate waveform forecasting from downstream task classification.
- Measure coverage: how often the quality-aware system chooses to abstain.
- Evaluate whether IMU/camera context improves robustness without leaking task cues.
- Only after personal repeatability is established, design a consented multi-participant study.

15.3 Product work

The most credible near-term product is a transparent personal BCI research environment rather than a universal thought reader. It can help a user capture high-quality sessions, build bounded personal labels, inspect what the headset actually produced and test conservative control primitives. That product direction aligns with the strongest current software and leaves room for stronger predictive research without making the interface depend on an unverified headline.

Conclusion

CORTEX is valuable because it has survived contact with implementation.

The project began with an unusually ambitious idea and gradually accumulated the machinery needed to test it: direct data acquisition, signal transformations, baselines, controlled tasks, latency trials, native software, session storage, replay and quality gates. The older results show how the idea developed; the current application gives me a much better base for testing it properly across whole sessions.

For the next CORTEX release I care about three things: whether I can follow the data from the headset to the result, whether the evaluation keeps training and testing genuinely separate, and whether the system knows when not to predict. Those tests matter more to me than forcing every old percentage to survive unchanged.

Publication statements

Data and code availability

This preprint was built from the current CORTEX source, a direct four-channel EEG capture and the project's earlier papers and source generations. I plan to publish the current code and new experiment outputs together. The older 2025 percentages are reported from their original versioned documents, and I still need to gather some of those original run files into the public release. No participant-derived raw EEG is distributed with this preprint.

Ethics and consent

This paper brings together existing project work; it did not introduce a new participant study and I am not claiming institutional ethics approval. I have left out historical personas, stakeholder-style quotations and anything whose consent history I could not confirm. Any future multi-participant study or public release of participant recordings will use documented consent, de-identification, withdrawal and deletion procedures.

Funding and competing interests

This project received no external funding. I have no competing interests to declare.

Author contribution

Connor Colyer: conceptualisation, research design, software development, investigation, visualisation, project administration and writing.

References

[1] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G. & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. Clinical Neurophysiology, 113(6), 767–791. https://doi.org/10.1016/S1388-2457(02)00057-3

[2] Pfurtscheller, G. & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. Clinical Neurophysiology, 110(11), 1842–1857. https://doi.org/10.1016/S1388-2457(99)00141-8

[3] Krigolson, O. E., Williams, C. C., Norton, A., Hassall, C. D. & Colino, F. L. (2017). Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research. Frontiers in Neuroscience, 11, 109. https://doi.org/10.3389/fnins.2017.00109

[4] Ratti, E., Waninger, S., Berka, C., Ruffini, G. & Verma, A. (2017). Comparison of medical and consumer wireless EEG systems for use in clinical trials. Frontiers in Human Neuroscience, 11, 398. https://doi.org/10.3389/fnhum.2017.00398

[5] Combrisson, E. & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. Journal of Neuroscience Methods, 250, 126–136. https://doi.org/10.1016/j.jneumeth.2015.01.010

[6] Varma, S. & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

[7] Debener, S., Minow, F., Emkes, R., Gandras, K. & de Vos, M. (2012). How about taking a low-cost, small, and wireless EEG for a walk? Psychophysiology, 49(11), 1617–1621. https://doi.org/10.1111/j.1469-8986.2012.01471.x

Appendix A — Project evidence map

Table A1. The main project artefacts used to reconstruct the three-year development history.

ID	Artefact	Date	Role	Source type
E01	CORTEX Summary.pages	13 Mar 2025	Early equation and 94.6% summary	Research draft
E02	CORTEX FuLL copy.pages	22 Apr 2025	Long paper, comparisons and waveform chart	Research report
E03	Blank 4 copy.pages	13 May 2025	Engineering report and acceptance-test appendices	Engineering report
E04	CTX.pages	2025 revisions	Expanded architecture, applications and ethics	Design dossier
E05	Cortex v0.1–v0.4	Dec 2024–2025	Historical Swift prototype lineage	Source code
E06	CTX Swift prototype	25 Jul 2025	Synthetic two-second alpha-band simulation	Source code
E07	MuseScope/MuseWire	Current	Direct BLE research platform	Source and tests
E08	muse-1fdd-eeg.csv	Current corpus	7,644-sample direct EEG capture	Captured data
E09	Colyer Thesis Workspace	2026	Broader formalisation and current theory work	LaTeX source

Appendix B — Metric registry

Table B1. Master metric registry.

Metric	Version/test	Value	How it is used here
Forecast accuracy	March 2025 summary	94.6% at 2 s	Early forecasting result
Held-out accuracy	April 2025 Table 5.1	92.6% ± 3.5	Main April comparison
Later classification	April discussion/conclusion	95.0% at 2 s	Separate later result
Per-task F1	April Table 5.2	0.93–0.96	Task-level April results
Pairwise AUC	April Table 5.3	0.92–0.99	Pairwise April results
Signal error	April Table 5.5	MSE 0.015; RMSE 0.122	Table value used in Section 7
AR signal error	April Table 5.5	MSE 0.021; RMSE 0.145	Table baseline used in Section 7
Cross-validated macro-F1	May AT-02	Mean 0.907	Ten recorded fold scores
Latency	May AT-01	Mean 1,475.8 ms; max 1,599 ms	Complete 30-value series
Setup	May AT-03	Mean 278.6 s	Internally consistent
Stability	May AT-05	99.64% reported	Excluded from headline results; rerun planned
Current Prediction Stack V3	Current	No held-out metric yet	Next benchmark

Appendix C — Software inventory

Table C1. Principal local CORTEX software lineage.

Path/project	Technology	Function
Software/EEGVisualization	Swift/Xcode	Early signal visualisation project
Software/Cortex/v0.1	SwiftUI/CoreBluetooth	First named CORTEX app and headset scan
Software/Cortex/v0.2–v0.3	Swift/Muse SDK	Vendor-SDK integration experiments
Software/Cortex/v0.4	SwiftUI	Graph/prediction-era prototype
Software/Cortex/CTX	SwiftUI Charts	Synthetic alpha-band two-second predictor simulation
Software/Muse/musewire.swift	Swift/CoreBluetooth	Direct Muse capture utility
Software/Muse/musescope.swift	SwiftUI/CoreBluetooth/ AVFoundation/SQLite	Current research application
Software/Muse/PredictionCore.swift	Swift	Prediction Stack V3
Research & Data/Colyer Equation	LaTeX	Current formal thesis workspace

Appendix D — Glossary

Table D1. Working glossary.

Term	Meaning in this preprint
BCI	Brain–computer interface.
EEG	Electroencephalography: scalp measurement of electrical potential changes.
Muse 2	Consumer headband used as the current sensor platform.
TP9/AF7/AF8/TP10	The four principal Muse EEG positions used by CORTEX.
Epoch/window	A bounded segment of samples processed as one model input.
Hop	The number of samples between consecutive windows.
Macro-F1	Unweighted mean of per-class F1 scores.
AUC-ROC	Area under the receiver-operating-characteristic curve.
MSE/RMSE	Mean squared error and its square root.
Readiness	Whether a label has sufficient accepted training evidence for use.
Unknown	Explicit fallback when no reliable label is available.
Hard blocker	A quality condition that prevents a lock-eligible result.
Soft blocker	A condition that reduces or conditionally prevents inference.
Causal split	An evaluation split that prevents future or dependent-session information leaking into training.

Appendix E — Project source corpus

The following project artefacts were used to build this preprint. They are listed as primary project provenance and are distinct from the external research bibliography.

- Connor Colyer, CORTEX Summary, Pages research draft, research begun 8 November 2023; file revised March 2025.
- Connor Colyer, The Colyer Predictive Thought Algorithm / CORTEX FuLL copy, long-form research draft, April 2025.
- Connor Colyer, CORTEX engineering report (archived as Blank 4 copy), May 2025.
- Connor Colyer, CTX research and design dossier, 2025 revisions.
- CORTEX Swift prototype directories v0.1, v0.2, v0.3, v0.4 and CTX.
- MuseScope and MuseWire source, README, tests and direct EEG CSV.
- Colyer Thesis Workspace: LaTeX chapters, system specification and compiled working PDF, 2026.

NEXT RELEASE

For new experiments I will publish the experiment ID, code commit, dataset manifest and machine-readable output together, so each result can be traced directly back to the run that produced it.