PREDICTIVE INFORMATION ATTRIBUTION WITH REGRET ACCOUNTING IN SEQUENTIAL NEURAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural decoding performance can reflect task structure and behavioral history as well as the current neural observation. We ask how much the observation improves prediction beyond the information available to the decoder. We compare a context predictor q_0 with a joint predictor q_1 using their log score difference on held-out events. The expected gain satisfies $G = I(Y; X \mid C) + R_0 - R_1$, relating conditional information to the difference in predictive regret between the fitted models. Controlled experiments recover three regimes with known ground truth. Persistence produces strong marginal decoding with zero conditional gain. Noisy XOR yields 0.531 conditional bits per sample despite zero marginal information. An underfit context model yields 0.363 fitted bits per sample under a conditional null; capacity matching removes the gain. In auditory attention EEG from 18 participants, supervised previous-trial label feedback reduces gain relative to a prevalence baseline, while a positive versus negative lag contrast remains under interleaved and contiguous splits. For spectral decoding of motor imagery EEG from 40 unseen participants, cue history reduces gain over schedule context by 0.0098 bits per trial. These results show how context and predictor capacity change the interpretation of neural decoding performance.

1 INTRODUCTION

Sequential neural decoders operate in structured prediction environments. Targets may persist, cues may follow a schedule, and recent behavior may predict the next outcome. A decoder can therefore perform well even when the current neural observation contributes little. In our motor imagery study, cue history alone reaches 0.761 balanced accuracy, compared with 0.565 for spectral EEG decoding. Joint accuracy measures the complete system but does not identify the evidence supplied by EEG.

Attribution begins by specifying the information available when the prediction is made. Evaluation must also respect the grouping and temporal structure of the observations. Recent neural decoding audits test whether performance depends on structural context or correct trial pairing (Snoek et al., 2019; Xu et al., 2026; Wong et al., 2026; Zhang et al., 2026; Du et al., 2026a; Ding, 2026). These developments lead to our question: given the non-neural information available at prediction time, how much held-out predictive evidence does the neural observation add?

Let Y denote the target, X the neural observation, and C the context. We compare a context predictor $q_0(y \mid c)$ with a joint predictor $q_1(y \mid c, x)$ on the same held-out events (Figure 1). Their expected log score difference measures gain in bits per event, with each estimate indexed by its target and context (Xu et al., 2020; Hewitt et al., 2021; Watson & Wright, 2021). The exact identity $G = I(Y; X \mid C) + R_0 - R_1$ relates conditional information to the difference in predictive regret. This motivates the capacity and calibration comparisons.

We study these effects in controlled systems and two EEG datasets. In auditory attention EEG ($n = 18$), access to the preceding trial label reduces the positive minus negative lag contrast from 0.179 to 0.042 $[0.024, 0.058]$ bits per trial, with a positive contrast across model families and split schemes. In motor imagery EEG ($n = 40$ unseen participants), adding cue history to schedule context absorbs 0.0098 $[0.0009, 0.0178]$ bits per trial. The appendix extends the estimator to continuous ECoG forecasting.

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

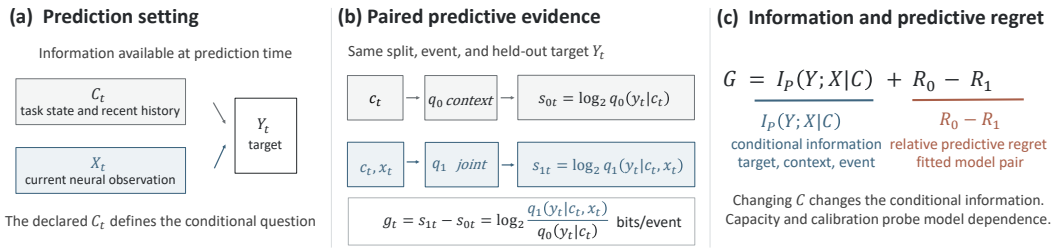

Figure 1: **Predictive information attribution.** (a) Context specifies the information available to the decoder. (b) Both predictors are scored on the same events, giving g_t and its empirical mean $\widehat{G}$. (c) Expected gain $G = \mathbb{E}[g_t]$ equals conditional information plus the difference in predictive regret.

Our contribution is a set of controlled tests that use regret accounting to explain gains under redundancy, synergy, and context underfitting. The EEG analyses then quantify how label feedback and cue history change fitted gain, using temporal and replacement controls to assess the remaining evidence.

2 Related Work

Conditional predictive information. Strictly proper scores evaluate probabilistic predictions against the outcome distribution. Expected log score improvement has a direct information theoretic interpretation (Gneiting & Raftery, 2007). Predictive $\mathcal{V}$-information and conditional probing measure usable information relative to a model family or baseline representation. Conditional predictive importance evaluates the contribution of a feature block (Xu et al., 2020; Hewitt et al., 2021; Watson & Wright, 2021); related decoder gaps distinguish probabilistic neural coding hypotheses (Kuo & Walker, 2026). We apply this foundation to sequential neural decoding. The analysis separates conditional information in the data distribution from the gain of learned predictors and from its finite-sample estimate. The regret identity states the gap between the first two quantities exactly.

Dynamics and neural evaluation. Sequential decoders use behavioral dynamics to stabilize prediction (Glaser et al., 2019). Recent systems learn attention transitions or movement dynamics jointly with a neural decoder (Yashengjiang et al., 2026; Du et al., 2026b). The same temporal structure can inflate an evaluation when autocorrelation or trial identity crosses a split (Xu et al., 2026; Wong et al., 2026). Confound analyses have identified gaze and trial-specific shortcuts in EEG decoding (Snoek et al., 2019; Rotaru et al., 2024; Shin et al., 2026). Misalignment tests give a complementary null for neural tracking (Geirnaert et al., 2026). Our score asks what the current observation adds after the predictable context is included.

Source attribution in neural decoding. Recent work makes the source of decoding performance an explicit object of study. Zhang et al. (2026) distinguish structural shortcuts, local evidence, and contextual aggregation in brain-to-language retrieval. Du et al. (2026a) compare context and neural motor cortex decoders under different information boundaries and trial replacement. Capacity-matched substitutions have also been used to test auditory attention evidence (Ding, 2026), and text perturbations have been used to audit multimodal forecasting (Sridhar et al., 2026). We examine how changes in context and model fit affect these comparisons. Controlled systems provide known information targets, and the two EEG studies measure the sensitivity of fitted gain to context and predictor capacity.

3 Predictive Information Attribution and Regret Accounting

3.1 Information Sets at Prediction Time

Let Y be the prediction target, X the focal observation (neural activity in our experiments), and C the background information available at the declared prediction time. In a sequential problem,

C may contain target or decoder history, candidate stimuli, or protocol variables. The conditional information $I_P(Y; X \mid C)$ changes with this choice, so each analysis specifies C before model fitting.

The deployment information set contains variables available to the intended system when it predicts. An oracle information set may add variables for a separate mechanistic analysis. Future labels and evaluation-fold statistics are excluded. Results from an oracle comparison belong to that enlarged information set and are reported separately.

Table 1: Primary information sets. The appendix gives the full registry, including oracle and excluded variables.

Dataset	Target	Prediction context C	Evaluation split
Synthetic	continuous or binary Y	history, gate, or continuous context declared by the generator	independent train/validation/test samples or trials
DTU	attended candidate	supervised previous-trial label feedback; persistence fitted on training transitions; first-trial fallback to the training-fold prior; candidates define symmetric EEG evidence	folds of complete trials; participant uncertainty
EEGMMI	motor cue	cue history; schedule as a weaker comparison	five outer folds with unseen subjects

3.2 Scientific target, model target, and estimator

At the distribution level, the incremental predictive information supplied by X is the Bayes log score contrast $I_P(Y; X \mid C) = \mathbb{E}_P[\log_2 \frac{p(Y|C,X)}{p(Y|C)}]$. Because these Bayes conditionals are unknown, we fit a context predictor $q_0(y \mid c)$ and a joint predictor $q_1(y \mid c, x)$ on training and validation data and define their population conditional log score gain

$$G(q_0, q_1; P) = \mathbb{E}_P \left[\log_2 \frac{q_1(Y \mid C, X)}{q_0(Y \mid C)} \right]. \tag{1}$$

When X is neural, we call this quantity neural evidence gain (NEGain). The value 2^G is the geometric mean factor by which the joint model changes the probability or density assigned to the observed targets.

For fold map $f(i)$, the held-out estimator is

$$\widehat{G} = \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{q_1^{(-f(i))}(y_i \mid c_i, x_i)}{q_0^{(-f(i))}(y_i \mid c_i)}. \tag{2}$$

Thus I_P is the scientific target, G is the population gain of the fitted models, and $\widehat{G}$ estimates G from paired evaluation predictions. The joint model contains the context model when its neural coefficient or evidence weight is zero. Participant summaries use $\widehat{G}_{\text{pop}} = S^{-1} \sum_{s=1}^{S} n_s^{-1} \sum_{i=1}^{n_s} g_{si}$, where g_{si} is the log score difference for event i of participant s. We report bits per trial for classification and bits per sample for continuous forecasting; ECoG scores average four scalar finger targets within participant and use bits/sample/finger. Values are directly comparable when the target, event, and prediction horizon agree. Appendix A gives the corresponding Gaussian log density form. Accuracy and correlation are secondary summaries.

3.3 Exact predictive regret accounting

Let R_0 and R_1 be the expected KL divergences in bits from the Bayes conditionals to q_0 and q_1, respectively. Adding and subtracting the two Bayes log scores yields the exact identity

$$\boxed{G(q_0, q_1; P) = I_P(Y; X \mid C) + R_0 - R_1.} \tag{3}$$

The proof is in Appendix A. Predictive regret records the loss incurred because a fitted predictor differs from its Bayes conditional. Bayes predictors have $R_0 = R_1 = 0$. Under the conditional

null $Y \perp X \mid C$, fitted gain equals $R_0 - R_1$. A joint model can therefore obtain positive gain by recovering structure missed by the context model. Negative gain occurs exactly when $R_1 - R_0 > I_P(Y; X \mid C)$: the joint model's additional regret exceeds the conditional information in X.

For predictor classes $\mathcal{Q}_0$ and $\mathcal{Q}_1$, let $H_{\mathcal{Q}}$ denote the infimum population log loss. Their optimal population contrast is

$$G^*_{\mathcal{Q}_0,\mathcal{Q}_1} = H_{\mathcal{Q}_0}(Y \mid C) - H_{\mathcal{Q}_1}(Y \mid C, X) = I_P(Y; X \mid C) + R^*_0 - R^*_1. \tag{4}$$

This is conditional predictive information relative to the chosen model classes. When $\mathcal{Q}_1$ contains the members of $\mathcal{Q}_0$ that ignore X, the optimal population contrast is nonnegative. Capacity matching and calibration analyses show how strongly the fitted result depends on relative regret.

3.4 Pairing, timing, and population uncertainty

Let $\widetilde{X}$ be drawn from another trial or temporal block within specified matching strata. The contrast $D_{\text{pair}} = \mathbb{E}[\log_2 \frac{q_1(Y|C,X)}{q_1(Y|C,\widetilde{X})}]$ measures the fitted model's reliance on pairing within those strata. Appendix D specifies the matching variables for each analysis. Matched temporal windows test when the evidence appears.

All learned transformations and model choices use training or validation data. The same held-out observations enter both terms of every contrast. Real-data intervals bootstrap participant means conditional on the fitted predictions; simulation intervals use independent generator replicates. A DTU sensitivity analysis also resamples the calibration cohort and reselects its scalar evidence weight.

4 Controlled Attribution Regimes

Equation 3 separates two axes: the distributional information in X given C, and the relative regret of the predictors used to reveal it. We construct three regimes that vary these axes independently.

Redundancy and innovation in a sequential system. Consider

$$Y_t = aY_{t-1} + \epsilon_t, \qquad X_t = bY_{t-1} + \gamma\epsilon_t + \eta_t, \tag{5}$$

where ϵ_t is independent standard Gaussian innovation and η_t is a stationary Gaussian process of unit variance, independent of the target process. Our generator uses $\eta_t = 0.6\eta_{t-1} + 0.8z_t$ with independent standard Gaussian z_t. With $C_t = Y_{t-1}$ and $\gamma = 0$, $Y_t \perp X_t \mid C_t$ even though marginal tracking can become strong as persistence a grows. For $\gamma > 0$, X_t measures the unpredictable innovation ϵ_t, and the normalized generator gives

$$I(Y_t; X_t \mid Y_{t-1}) = \frac{1}{2}\log_2(1 + \gamma^2). \tag{6}$$

The analytic target depends on innovation strength and is invariant to target persistence, separating conditional evidence from marginal alignment.

Synergy hidden from marginal decoding. Let C and X be independent fair bits and let $Y = C \oplus X \oplus N$, where $N \sim \text{Bernoulli}(\delta)$. Then $I(Y; X) = 0$, but

$$I(Y; X \mid C) = 1 - h_2(\delta). \tag{7}$$

Here the context reveals evidence that is unavailable to either marginal predictor, establishing the opposite direction from the redundancy regime.

Unequal regret under a conditional null. We draw $C \sim \text{Uniform}[-2, 2]$, set $X = C^2 + 0.6\eta$, and sample $Y \sim \text{Bernoulli}(\sigma(3(C^2 - 0.8)))$. Because the noises are independent, $I(Y; X \mid C) = 0$. A linear context predictor misses the quadratic response, whereas the joint predictor can use X as a redundant proxy. Hence Equation 3 predicts $G = R_0 - R_1 > 0$. Supplying C^2 to a capacity-matched context predictor removes the representational asymmetry; output calibration alone leaves it intact.

Context refinement. Let a richer deployment information set be $C_H = (C_L, Z)$, and let G_L and G_H denote gains from separately fitted predictor pairs. The refinement contrast

$$A_{L\to H} = G_L - G_H \tag{8}$$

has the exact decomposition

$$\begin{aligned} A_{L\to H} &= I(Y; X \mid C_L) - I(Y; X \mid C_L, Z) + \Delta_R \\ &= I(Y; Z \mid C_L) - I(Y; Z \mid C_L, X) + \Delta_R, \end{aligned} \tag{9}$$

where $\Delta_R = (R_{0L} - R_{1L}) - (R_{0H} - R_{1H})$. A context refinement can change both conditional information and relative regret. The information term is positive when Z is redundant with X for predicting Y, and negative when Z reveals a synergistic interaction. Each level of a context ladder therefore defines a distinct prediction question.

5 Experimental Protocol

The experiments address three questions. How does gain behave when conditional information and model regret are known? Does auditory EEG add evidence in the expected temporal direction? How does motor EEG attribution change when protocol history becomes available? Each q_0 and q_1 pair is scored on the same held-out observations. Appendix Table B.1 records each target, prediction unit, context, neural observation, and split. Population inference is specified in Appendix D.

Controlled ground truth: how do information and model regret affect gain? In the linear-Gaussian regime, we independently vary target persistence $a \in \{0, 0.5, 0.8, 0.95\}$ and neural innovation $\gamma \in \{0, 0.15, 0.30, 0.60, 1.00\}$ over 64 replicates per cell, with complete 256-sample trials split 32/8/16. Linear Gaussian predictors are compared with analytic conditional information and marginal correlation. We also evaluate context-matched replacement. Noisy XOR uses six flip probabilities and 128 replicates with 4,000 training and 8,000 test samples. A 128-replicate conditional-null system with $X = C^2 + \epsilon$ compares an underfit context model and a joint model using the proxy. An independent dummy, quadratic capacity matching, and validation calibration separate representation from probability scale. Appendix C gives complete generators and splits.

Auditory attention: does EEG at positive lags add evidence about the attended speaker? The DTU corpus contains 18 participants, each with 60 fifty-second trials of 64-channel EEG recorded with competing male and female speakers (Fuglsang et al., 2017). After restoring fixed male/female candidate order, the target is the identity of the attended speaker. The primary comparison is a supervised label-feedback evaluation: the preceding trial's true label is supplied before each prediction. The context predictor uses this label and a persistence probability estimated from training transitions; the first trial falls back to the training-fold attention prior. A weaker comparison uses only that prior. The two candidate envelopes enter symmetrically when EEG reconstruction is converted to the difference between male and female tracking evidence. A ridge decoder reconstructs the envelope from nine positive EEG lags (125–375 ms relative to envelope time), and candidate correlations define trial evidence. Setting the evidence weight to zero recovers the context-only predictor exactly. Five folds contain 36/12/12 train/validation/test trials; the fusion weight is calibrated from the other 17 participants' out-of-fold predictions. The original folds are interleaved; a ridge sensitivity refits the complete pipeline using contiguous 12-trial test blocks. A matched EEG-leading window at negative lags (-375 to -125 ms) tests directionality. EEGNet-lite and a causal TCN use the same observations and interleaved splits.

Motor imagery: how much fitted EEG gain is absorbed by protocol history? We study imagined left- versus right fist trials from 40 participants in the PhysioNet EEG Motor Movement/Imagery dataset (Schalk et al., 2004). Five outer folds assign 24/8/8 participants to train/validation/test. Unlabeled R04 standardizes participant features; R08 and R12 supply prediction trials. Neural features average 8–30 Hz motor channel log power over three one-second post-cue windows; symmetric pre-cue and frontal features provide controls. The context ladder adds two preceding cues to run and trial position. Regularized logistic predictors are selected on validation participants and refit on train plus validation. A protocol lookup tests context capacity, and affine logit calibration tests probability scale.

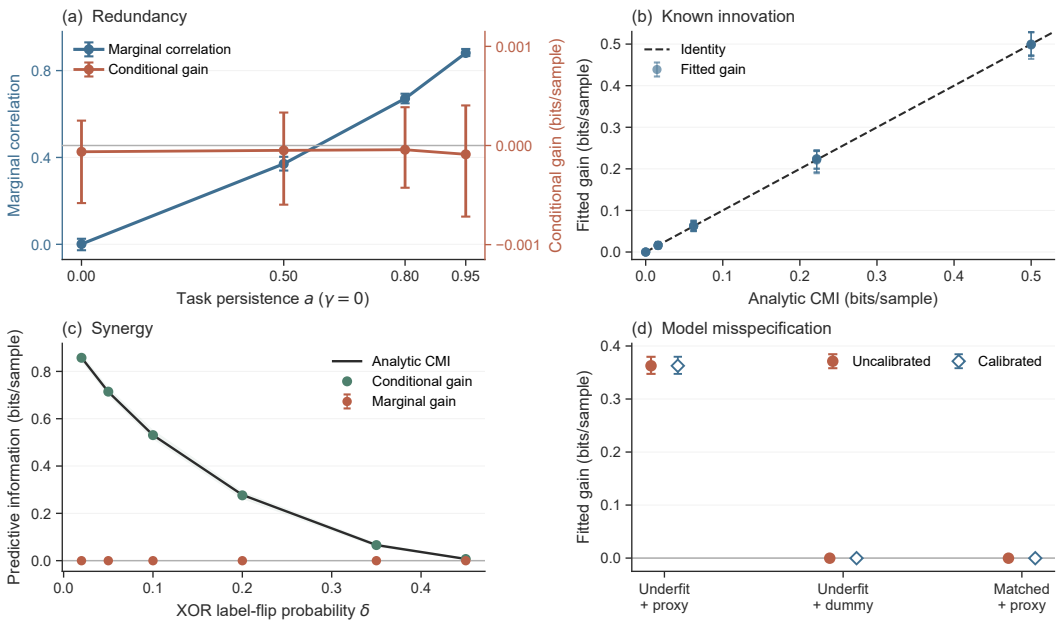

Figure 2: **Controlled regimes and predictive regret.** (a) Persistence raises marginal correlation at zero conditional gain. (b) Fitted gain follows analytic CMI. (c) XOR reveals conditional information absent marginally. (d) Under a conditional null, capacity matching removes the proxy gain; calibration leaves it unchanged. Panels (a)–(b) use 64 replicates per cell and (c)–(d) use 128.

6 Results

We compare fitted gains with known information targets, then examine their dependence on context in DTU and EEGMMI. Table 2 summarizes the results.

6.1 Controlled cases separate redundancy, synergy, and model regret

Redundancy does not become incremental evidence. Figure 2a increases target persistence while keeping neural innovation at zero. At $a = 0.95$, correlation from the neural input alone reaches 0.881 [0.867, 0.900], while NEGain on the same observations is -0.0001 [-0.0007, 0.0004] bits/sample, and the conditional replacement gap is 0.0000 [-0.0006, 0.0008]. Persistent dynamics therefore yield strong marginal alignment without evidence beyond the declared history.

Known innovation is recovered on the information scale. When X_t measures unpredictable innovation, fitted gain follows analytic CMI and changes little with persistence (Figure 2b). At $\gamma = 0.60$, the analytic value is 0.2218 bits/sample and the estimate is 0.2210 [0.1921, 0.2449] at $a = 0.95$ and 0.2236 [0.2004, 0.2428] at $a = 0$.

Conditioning can also reveal synergy. In noisy XOR at $\delta = 0.10$, marginal neural gain is -0.0002 [-0.0011, 0.0003] bits/sample, whereas conditional NEGain is 0.5306 [0.5130, 0.5539], matching the analytic 0.5310 (Figure 2c).

Relative regret can create fitted gain under a conditional null. In Figure 2d, true CMI is zero, yet a linear context model compared with a joint model containing a redundant quadratic proxy yields 0.3627 [0.3474, 0.3796] bits/sample in 128/128 replicates. Context and joint regrets are 0.5164 and 0.1537 bits/sample, exactly accounting for the gain. Calibration leaves it unchanged; an independent dummy and a capacity-matched context model both give -0.0002 bits/sample.

6.2 PREVIOUS-TRIAL LABELS REDUCE AUDITORY GAIN

Gain decreases when label feedback is available. The attended target changes only one to three times across the 60 trials of each DTU participant. Under supervised label feedback, a context model reaches 0.947 balanced accuracy and 0.285 bits/trial NLL. With this context supplied to both predictors, ridge post-lag NEGain is 0.046 [0.025, 0.065], compared with 0.004 [−0.005, 0.011] for the EEG-leading control. Their paired contrast is 0.042 [0.024, 0.058] bits/trial (Figure 3a). The corresponding prevalence-only contrast is 0.179 [0.123, 0.238]. Trial history therefore reduces fitted gain while preserving a positive versus negative lag contrast.

The contrast remains positive with contiguous folds. Refitting ridge with five contiguous 12-trial test blocks yields a post minus pre contrast of 0.032 [0.019, 0.046] bits/trial under supervised label feedback, positive for 16 of 18 participants (Figure 3b). Both split schemes therefore support the temporal direction.

The direction persists across model families. Under the same previous-trial context and interleaved folds, post minus pre NEGain is 0.020 [0.010, 0.030] for EEGNet and 0.016 [0.002, 0.027] for the causal TCN, compared with 0.042 [0.024, 0.058] for ridge (Figure 3c). Model family changes the magnitude but preserves the positive temporal direction. Averaging five seeds within participant gives contrasts of 0.0145 [0.0052, 0.0234] for EEGNet and 0.0149 [0.0019, 0.0253] for TCN under the same feedback context (Appendix E).

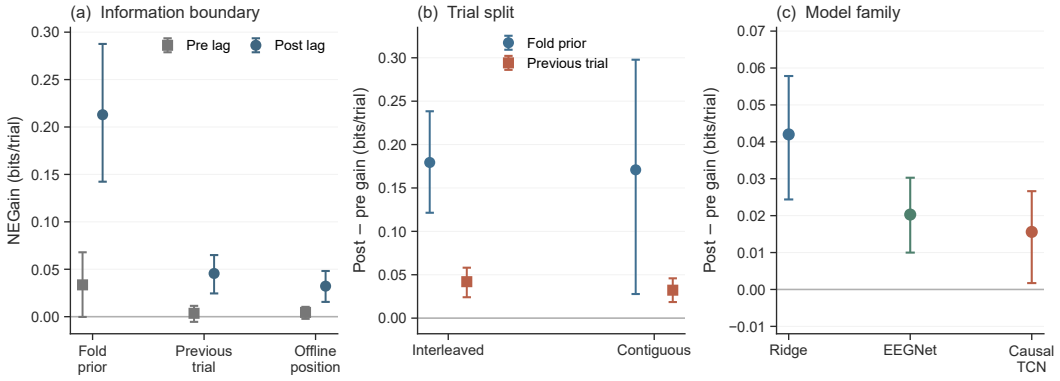

Figure 3: **Auditory EEG gain under different history conditions.** (a) Ridge pre- and post-lag NEGain under a training-fold prior, previous-trial label feedback, and an offline trial-position interpolation diagnostic. (b) Ridge post minus pre contrasts under interleaved and contiguous folds. (c) Post minus pre contrasts across model families with previous-trial label feedback. Lags are relative to the speech envelope. Intervals bootstrap participants ($n = 18$).

Longer observations resolve evidence beyond label feedback. The ridge post minus pre contrast is −0.0001 [−0.0035, 0.0033] bits/trial for the first second and 0.0138 [0.0071, 0.0200] for the first 20 seconds (Table E.5). The decoder is fixed; each prefix uses candidate correlations and weights selected from the other participants. The one-second effect found with a prevalence baseline is unresolved after previous-label feedback is supplied.

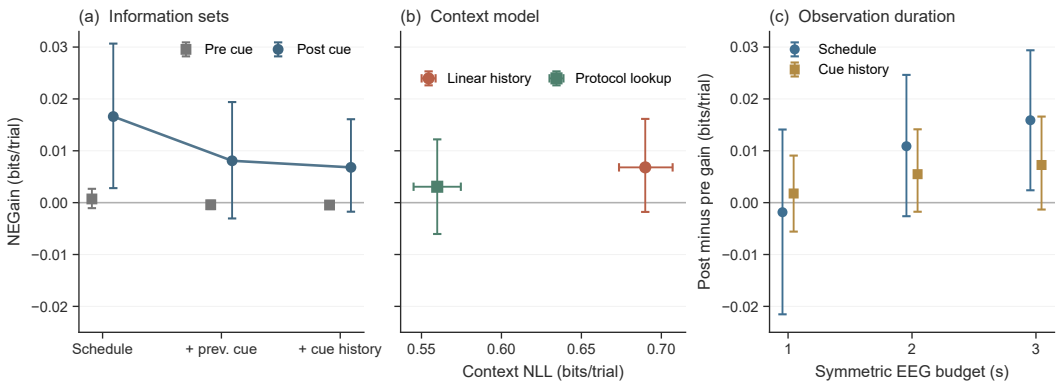

Figure 4: **Motor EEG gain across context and observation conditions.** (a) Three-second pre- and post-cue NEGain across increasingly informative legal contexts. The immediately previous cue absorbs most of the gain conditioned on schedule. (b) Context negative log likelihood and uncalibrated post NEGain for linear and lookup representations of cue history. (c) Post minus pre NEGain for symmetric observation durations of one to three seconds. Results use folds with unseen subjects and R04-only unlabeled normalization, and R08/R12 evaluation ($n = 40$).

6.3 Cue history reduces fitted motor EEG gain

Preceding cues reduce gain over schedule context. With schedule and trial position as context, spectral decoding of three seconds of post-cue motor EEG yields 0.0166 [0.0028, 0.0305] bits/trial and a post minus pre contrast of 0.0159 [0.0025, 0.0296] (Figure 4a). One preceding cue absorbs 0.0085 [0.0025, 0.0141]; history with two cues yields 0.0068 [−0.0017, 0.0162] and total absorption of 0.0098 [0.0009, 0.0178]. Under full history, balanced accuracy is 0.761 for context, 0.763 jointly, and 0.565 for independently tuned EEG-only prediction. The preceding cue explains most of the ladder change.

A richer representation improves context prediction. A protocol lookup selected by validation lowers raw context NLL by 0.1301 [0.1167, 0.1437] bits/trial, from 0.6899 to 0.5598, relative to the linear model based on cue history (Figure 4b). Post NEGain is 0.0031 [−0.0060, 0.0122] for the lookup and 0.0068 [−0.0018, 0.0161] for the linear model. Calibration changes either estimate by at most 0.0005 bits/trial. The paired linear-minus-lookup gain difference is 0.0037 [−0.0046, 0.0122]. Context prediction improves clearly; the reduction in residual neural gain remains uncertain.

The context ladder and lookup comparison use the same outer folds, evaluation trials, motor features, and R04 normalization. Adding preceding cues expands the information set; replacing the linear context predictor with a lookup changes how the same information is used.

Gain estimates vary with observation duration. The post minus pre point estimates conditioned on schedule are positive in both evaluation runs: 0.0133 [−0.0040, 0.0299] bits/trial in R08 and 0.0185 [−0.0010, 0.0390] in R12. No isolated one-second window forms a post-cue peak; under symmetric budgets the contrast progresses from −0.0018 at one second to 0.0109 at two seconds and 0.0159 [0.0024, 0.0294] at three seconds (Figure 4c). The three-second aggregate gives a clearer temporal contrast than the individual one-second windows.

Under schedule context, frontal features yield post-cue NEGain of 0.0007 [−0.0051, 0.0067] bits/trial, and left–right motor differences yield 0.0141 [−0.0068, 0.0327] (Table F.1). These controls retain the same observation windows and evaluation folds.

Neural-model sensitivity analyses further show that fitted gain depends on stable probabilistic prediction, reinforcing that NEGain should not be interpreted as a model-free information estimate (Appendices F.5–F.6).

Table 2: Principal results in bits per sample or trial. DTU rows report the positive minus negative lag contrast for ridge with previous-trial label feedback; EEGMMI rows use spectral logistic decoding. Real-data intervals bootstrap participants; synthetic intervals are replicate percentiles.

Dataset	Contrast	Estimate	95% interval	Unit
Synthetic	NEGain, $a = .95, \gamma = 0$	-0.0001	$[-0.0007, .0004]$	bits/sample
Synthetic	XOR conditional, $\delta = .10$	0.5306	$[0.5130, 0.5539]$	bits/sample
Synthetic	zero CMI, underfit proxy	0.3627	$[0.3474, 0.3796]$	bits/sample
Synthetic	zero CMI, matched proxy	-0.0002	$[-0.0013, 0.0001]$	bits/sample
DTU	post minus pre, interleaved	0.042	$[0.024, 0.058]$	bits/trial
DTU	post minus pre, contiguous	0.032	$[0.019, 0.046]$	bits/trial
EEGMMI	post, schedule context	0.0166	$[0.0028, 0.0305]$	bits/trial
EEGMMI	post, cue history context	0.0068	$[-0.0017, 0.0162]$	bits/trial
EEGMMI	post reduction with cue history	0.0098	$[0.0009, 0.0178]$	bits/trial

7 Discussion and Limitations

The two EEG studies show why context belongs in the definition of a decoding comparison. In DTU, supervised previous-trial label feedback substantially reduces gain, but the positive versus negative lag contrast remains positive across the tested models and split schemes. In EEGMMI, preceding cues reduce spectral-model gain over schedule context. Lookup improves context prediction, while its effect on the remaining gain is uncertain.

These comparisons assign different roles to the same recording. A prevalence context can credit EEG with predictive structure associated with persistent trial labels. Supplying label feedback changes the question to whether EEG improves prediction after that history is available. Likewise, conditioning on cue history makes the EEGMMI comparison concern variation beyond the protocol sequence. The appropriate context follows the intended prediction setting, so gains under different information sets answer different scientific questions. Richer context can also increase gain, as XOR illustrates, by making otherwise unusable evidence predictive.

The information boundary defines the attribution question, while predictive regret explains why fitted gain can differ from conditional information. Adding context can change the information target (Equation 9); at a fixed context, changing the predictors affects gain through $R_0 - R_1$. The synthetic null illustrates this distinction: supplying the missing quadratic term to the context predictor removes apparent neural gain while conditional information remains zero. On real data, these comparisons measure the sensitivity of fitted gain; identifying its information and regret terms separately requires the Bayes conditionals.

This distinction suggests two separate comparisons in practice. Evaluate context sets on the same held-out events to establish how attribution depends on the prediction setting. Within each set, compare model capacities using a common validation protocol to assess whether the fitted gain depends on missed predictive structure. Calibration examines sensitivity to probability scale. Temporal controls locate the evidence relative to the target, and replacement controls assess reliance on the correct pairing within the specified strata.

8 Conclusion

High decoding performance alone does not establish how much evidence the current neural observation provides. Paired log scores address this question relative to a declared information set. Controlled experiments show how conditional information and predictive regret jointly determine fitted gain; the EEG studies show how trial and cue history change the attribution of decoding performance. The information boundary defines the scientific question, and regret accounting clarifies what the fitted comparison can establish.

AI USE STATEMENT

In this work, we used OpenAI Codex to assist with literature discovery, implementation and debugging of analysis code, construction of the controlled synthetic experiments, figure preparation, and drafting and editing the manuscript. We reviewed the cited sources, tested the analysis code, and verified the reported results against executable analyses and saved outputs. The authors take responsibility for the study design, scientific claims, and final content of the paper.

ETHICS STATEMENT

This study reanalyzes public, previously collected EEG and ECoG datasets. The original dataset reports describe their consent, ethics, and access procedures (Fuglsang et al., 2017; Schalk et al., 2004; Tangermann et al., 2012). We recruited no participants and performed no re-identification.

REPRODUCIBILITY STATEMENT

The main text registers information sets, estimands, split units, controls, and population uncertainty. The appendix provides preprocessing and model details, the complete information set registry, proof expansion, participant summaries, and sensitivity analyses. The anonymized supplementary code package contains frozen configurations, aggregate result tables, validation tests, and scripts for reproducing every reported table and figure.

REFERENCES

Yu Ding. Falsifiable substitution tests reveal task-structured neural evidence for auditory attention, 2026. URL https://www.biorxiv.org/content/10.64898/2026.08.03.742580v2. bioRxiv preprint.

Zonghan Du, Zhong Yuan Lai, Liang Hu, and Tangwei Ye. A context-conditional audit of trial-pairing-dependent neural gain in motor-cortex decoding, 2026a. URL https://www.biorxiv.org/content/10.64898/2026.08.04.742876v1. bioRxiv preprint, posted August 10, 2026.

Zonghan Du, Zhongyuan Lai, and Liang Hu. Decoding by dynamics: Reframing neural decoding as stable control inference with behavior priors, 2026b. URL https://www.biorxiv.org/content/10.64898/2026.07.25.740738v1.

Søren A. Fuglsang, Torsten Dau, and Jens Hjortkjær. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage*, 156:435–444, 2017. doi: 10.1016/j.neuroimage.2017.04.026.

Simon Geirnaert, Alexander Bertrand, Tom Francart, and Jonas Vanthornhout. Modeling and interpreting correlations, null distributions and significance levels in neural tracking of natural stimuli, 2026. URL https://arxiv.org/abs/2608.10887.

Joshua I. Glaser, Ari S. Benjamin, Rafi Farhoodi, and Konrad P. Kording. The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology*, 175:126–137, 2019. doi: 10.1016/j.pneurobio.2019.01.008.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, 2021. doi: 10.18653/v1/2021.emnlp-main.122.

Po-Chen Kuo and Edgar Y. Walker. An information-theoretic framework for optimizing experimental design to distinguish probabilistic neural codes. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf?id=doxBjZ88H3.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional network for EEG-based brain-computer interfaces, 2018. URL https://arxiv.org/abs/1611.08024. Version 4.

Iustina Rotaru, Simon Geirnaert, Nicolas Heintz, Iris Van de Ryck, Alexander Bertrand, and Tom Francart. What are we really decoding? unveiling biases in EEG-based decoding of the spatial focus of auditory attention. *Journal of Neural Engineering*, 21(1):016017, 2024. doi: 10.1088/1741-2552/ad2214.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072.

Wonchul Shin, Inyong Choi, and Kyogu Lee. Breaking shortcut learning for cross-trial EEG-guided target speech extraction via two-stage training, 2026. URL https://arxiv.org/abs/2606.24164. Accepted at Interspeech 2026.

Lukas Snoek, Steven Miletić, and H. Steven Scholte. How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, 184:741–760, 2019. doi: 10.1016/j.neuroimage.2018.09.074.

Karthik Sridhar, Atharva Gupta, Nishant Pradhan, Murari Mandal, Dhruv Kumar, and Saurabh Deshpande. Semantics or structure? auditing text sensitivity in multimodal time-series forecasting, 2026. URL https://arxiv.org/abs/2608.22321. ICML 2026 Workshop on Foundation Models for Structured Data.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

David S. Watson and Marvin N. Wright. Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110:2107–2129, 2021. doi: 10.1007/s10994-021-06030-6.

R. Wong, S. I. Zhu, M. H. McCullough, and G. J. Goodhill. Preventing data leakage in neural decoding, 2026. URL https://www.biorxiv.org/content/10.64898/2026.01.26.701583v1.

Xiran Xu, Bo Wang, Boda Xiao, Yadong Niu, Yiwen Wang, Xihong Wu, Heping Cheng, and Jing Chen. The impacts of temporal autocorrelations on EEG decoding. *Biomedical Signal Processing and Control*, 113:108783, 2026. doi: 10.1016/j.bspc.2025.108783.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

Yushan Yashengjiang, Jie Zhang, Miao Sun, Huadong Liang, Xin Li, and Zhen-hua Ling. End-to-end markov state sequence learning for auditory attention decoding, 2026. URL https://arxiv.org/abs/2607.18614.

Xinyu Zhang, Sichao Liu, Runhao Lu, Alexandra Woolgar, and Lihui Wang. What are we actually decoding? source attribution for non-invasive brain-to-language retrieval, 2026. URL https://arxiv.org/abs/2605.24524.

A MATHEMATICAL DETAILS

A.1 PROPER SCORE DECOMPOSITION

Define conditional regrets using base 2 logarithms

$$R_0 = \mathbb{E}_C D_{\mathrm{KL},2}(P(Y \mid C) \,\|\, q_0(Y \mid C)), \tag{A.1}$$

$$R_1 = \mathbb{E}_{C,X} D_{\mathrm{KL},2}(P(Y \mid C, X) \,\|\, q_1(Y \mid C, X)). \tag{A.2}$$

Adding and subtracting the true conditional log densities in Equation 1 gives

$$\begin{aligned} G &= \mathbb{E}\left[\log_2 \frac{P(Y \mid C, X)}{P(Y \mid C)}\right] \\ &\quad + \mathbb{E}\left[\log_2 \frac{q_1(Y \mid C, X)}{P(Y \mid C, X)}\right] - \mathbb{E}\left[\log_2 \frac{q_0(Y \mid C)}{P(Y \mid C)}\right] \\ &= I(Y; X \mid C) - R_1 + R_0. \end{aligned} \tag{A.3}$$

This is Equation 3. In the Bayes-optimal case both regrets vanish. With learned predictors, the identity assigns every departure from conditional mutual information to the difference in predictive regret.

A.2 OPTIMAL POPULATION MODELS WITHIN FIXED CLASSES

Let log loss be measured in bits and define

$$H_{\mathcal{Q}_0}(Y \mid C) = \inf_{q \in \mathcal{Q}_0} \mathbb{E}[-\log_2 q(Y \mid C)], \tag{A.4}$$

$$H_{\mathcal{Q}_1}(Y \mid C, X) = \inf_{q \in \mathcal{Q}_1} \mathbb{E}[-\log_2 q(Y \mid C, X)]. \tag{A.5}$$

The cross-entropy identity gives

$$H_{\mathcal{Q}_0}(Y \mid C) = H(Y \mid C) + R_0^*, \tag{A.6}$$

$$H_{\mathcal{Q}_1}(Y \mid C, X) = H(Y \mid C, X) + R_1^*. \tag{A.7}$$

Subtracting the second line from the first proves Equation 4. If every $q_0 \in \mathcal{Q}_0$ has a counterpart in $\mathcal{Q}_1$ that ignores X, then $H_{\mathcal{Q}_1}(Y \mid C, X) \leq H_{\mathcal{Q}_0}(Y \mid C)$ and the optimal population contrast is nonnegative. This property concerns the model classes at their population optima; finite sample held-out estimates can be negative.

A.3 CONTEXT REFINEMENT IDENTITY

For $C_H = (C_L, Z)$, the chain rule gives two expansions of the same conditional mutual information:

$$I(Y; X, Z \mid C_L) = I(Y; X \mid C_L) + I(Y; Z \mid C_L, X), \tag{A.8}$$

$$= I(Y; Z \mid C_L) + I(Y; X \mid C_L, Z). \tag{A.9}$$

Equating them yields

$$I(Y; X \mid C_L) - I(Y; X \mid C_L, Z) = I(Y; Z \mid C_L) - I(Y; Z \mid C_L, X). \tag{A.10}$$

Applying Equation 3 to the predictor pairs with low and high context and subtracting proves Equation 9. The sign of the information term depends on the distribution: positive values indicate net redundancy between X and Z for predicting Y, whereas negative values indicate net synergy.

A.4 GAUSSIAN SCORES AND EMPIRICAL AGGREGATION

For Gaussian predictive distributions with means μ_k and residual variances σ_k^2, the gain for each observation is the log-density ratio

$$g_i = \log_2 \frac{\mathcal{N}(y_i; \mu_1, \sigma_1^2)}{\mathcal{N}(y_i; \mu_0, \sigma_0^2)} \tag{A.11}$$

$$= \frac{1}{2 \ln 2}\left[\ln \frac{\sigma_0^2}{\sigma_1^2} + \frac{(y_i - \mu_0)^2}{\sigma_0^2} - \frac{(y_i - \mu_1)^2}{\sigma_1^2}\right]. \tag{A.12}$$

Residual variances are estimated from validation residuals and then fixed for evaluation. For classification, the score is the log probability of the observed label. We first average paired differences between observations within the declared prediction unit, then within participant for population inference. Thus long trials and participants with more usable windows do not receive additional population weight.

B Complete information set registry

Table B.1: Registered information sets used by every reported analysis. Legal context is available at the declared prediction time; oracle variables support sensitivity analyses for the information set.

Dataset	Field	Registered specification
Synthetic	Target/unit	Continuous AR target or noisy XOR binary label per sample
	Declared context	Past target, binary XOR gate, or continuous generator variable C
	Neural observation	AR mixture, independent XOR bit, or redundant noisy C^2 proxy
	Known truth	Analytic CMI for AR/XOR; zero CMI for misspecification null
	Model stress	Underfit, dummy input, calibrated, and matched capacity comparisons
	Split/population	Independent train/validation/test samples or trials; generator replicate
	Excluded	Evaluation outcomes during fitting, calibration, or model selection
DTU	Target/unit	Attended candidate identity per complete trial
	Supervised feedback	Previous-trial true label supplied before prediction; persistence fitted on training transitions; first trial uses the training-fold prior. Candidates enter the symmetric EEG evidence operator
	Neural observation	EEG at positive lags of 125–375 ms; matched EEG-leading control at negative lags
	Split/population	Folds of complete trials; participant bootstrap
	Excluded/diagnostic	Current trial identity is excluded from context. Trial-position interpolation is an offline diagnostic; the prior-only comparison omits label feedback
EEGMMI	Target/unit	Motor imagery cue per trial
	Legal context	Cue history; schedule and protocol lookup comparisons
	Neural observation	Three-second motor mu/beta summaries or band-limited EEGNet input; pre-cue and frontal controls
	Split/population	Five outer folds with unseen subjects; participant bootstrap
	Excluded	Current cue and labeled R04 calibration
BCIC-IV-4	Target/unit	Future finger flexion per 40 ms sample per finger
	Legal context	Cross-fitted past neural state estimates
	Neural observation	Causal 240 ms broadband, high-gamma, or low-frequency ECoG
	Split/summary	Official chronological test; three individual participants and their mean after finger averaging
	Oracle/excluded	Measured past kinematics is oracle only; test kinematics are excluded

C Controlled attribution details

C.1 Persistent Gaussian system

Each independent trial begins from the stationary distribution of $Y_t = aY_{t-1}+\epsilon_t$, with $\epsilon_t \sim \mathcal{N}(0,1)$. The neural noise follows $\eta_t = 0.6\eta_{t-1} + \sqrt{1-0.6^2}\,u_t$ with $u_t \sim \mathcal{N}(0,1)$, and the observation is $X_t = Y_{t-1}+\gamma\epsilon_t+\eta_t$. We discard 64 burn-in samples and keep 256 samples per trial. Each replicate contains 32 training, 8 validation, and 16 test trials. Ridge coefficient 10^{-4} is used for the linear context, neural input only, joint, and X given C regressions. Separate Gaussian variances for the context and joint models are estimated from validation residuals.

The conditional replacement preserves the fitted component $\widehat{\mathbb{E}}[X \mid C]$ and circularly shifts complete residual trials by a nonzero offset. Null-normalized temporal tracking uses each test trial as an aligned window and the other test trials as its misalignment distribution. Table C.1 shows the slice with high persistence. The fitted gain follows the analytic target across innovation strengths even as marginal correlation and tracking remain large at $\gamma = 0$.

Table C.1: Persistent Gaussian system at $a = 0.95$. Intervals are empirical 2.5th–97.5th percentiles over 64 generator replicates.

γ	Analytic CMI	NEGain [interval]	Neural corr.	Null-normalized tracking
0.00	0.0000	-0.0001 [-0.0007, 0.0004]	0.881	6.52
0.15	0.0161	0.0154 [0.0088, 0.0223]	0.895	6.74
0.30	0.0622	0.0626 [0.0513, 0.0740]	0.913	6.99
0.60	0.2218	0.2210 [0.1921, 0.2449]	0.934	7.78
1.00	0.5000	0.4993 [0.4644, 0.5339]	0.949	8.36

C.2 Noisy XOR synergy

For each flip probability, independent fair bits C and X generate $Y = C \oplus X \oplus N$. Jeffreys-smoothed categorical probabilities are fit on 4,000 samples and evaluated on 8,000 samples. The context-only predictor uses the two cells indexed by C, and the joint predictor uses the four cells indexed by (C, X). Across 128 replicates, marginal neural gain stays near zero while conditional gain follows $1 - h_2(\delta)$ throughout the noise curve (Table C.2). This experiment establishes that conditioning can reveal evidence that neither marginal predictor contains.

Table C.2: Noisy XOR synergy curve. Values are bits/sample; intervals summarize 128 independent replicates. Analytic marginal information is zero in every row.

δ	Analytic conditional information	Conditional NEGain [interval]	Marginal neural gain
0.02	0.8586	0.8575 [0.8411, 0.8748]	-0.0002
0.05	0.7136	0.7145 [0.6964, 0.7326]	-0.0002
0.10	0.5310	0.5306 [0.5130, 0.5539]	-0.0002
0.20	0.2781	0.2761 [0.2603, 0.2932]	-0.0002
0.35	0.0659	0.0658 [0.0569, 0.0745]	-0.0002
0.45	0.0072	0.0069 [0.0042, 0.0101]	-0.0002

C.3 Conditional null with unequal model regret

We sample 4,000/2,000/8,000 train/validation/test observations per replicate. Logistic regressions use $C = 100$ and Platt maps fitted only on validation data use an affine transformation of the logit. The oracle Bernoulli probability supplies the population log score needed to measure R_0 and R_1. Table C.3 closes the decomposition numerically: the underfit pair's 0.3627-bit gain equals its regret reduction, calibration leaves the structural gap intact, and giving both predictors the quadratic representation returns gain to zero.

Table C.3: Model misspecification under true $I(Y; X \mid C) = 0$. Values are bits/sample with percentile intervals over replicates.

Context and joint models	Calibration	NEGain [interval]	R_0	R_1
linear C; linear (C, X)	raw	0.3627 [0.3474, 0.3796]	0.5164	0.1537
linear C; linear (C, X)	Platt	0.3627 [0.3474, 0.3797]	0.5167	0.1540
linear C; linear (C, U)	raw	-0.0002 [-0.0012, 0.0002]	0.5164	0.5166
quadratic C; quadratic (C, X)	raw	-0.0002 [-0.0013, 0.0001]	0.0006	0.0008

D Shared estimation and model specifications

Population inference. Synthetic intervals are empirical 2.5th to 97.5th replicate percentiles. DTU and EEGMMI bootstrap participant means for gains and paired contrasts, conditional on fitted predictions: 20,000 draws in the original analyses and 10,000 in the added neural-model and label-feedback sensitivity analyses. Replacement strata are described below.

Every comparison fits q_0 and q_1 on the same development observations and scores them on the same held-out targets. Standardization parameters, regularization, early stopping, predictive variances, and probability calibration are estimated without evaluation targets. Table D.1 records the selected model families and search spaces.

Table D.1: Model and selection specifications. All grids are selected by validation loss within the declared split.

Analysis	Component	Specification
Synthetic AR	linear Gaussian	Ridge 10^{-4}; validation residual variance fitted for each model
Synthetic XOR	categorical	Jeffreys-smoothed Bernoulli cell probabilities
Synthetic null	logistic	$C = 100$; optional affine logit calibration using validation data
DTU	ridge	$\lambda \in \{1, 10, 10^2, \ldots, 10^6\}$; 577 parameters
DTU	EEGNet-lite	temporal convolution, depthwise spatial convolution, separable temporal convolution, ELU, dropout 0.25; 1,465 parameters
DTU	causal TCN	width 32; two residual blocks with dilations 1 and 2; kernel 3; dropout 0.20; 8,449 parameters
DTU	deep optimization	AdamW, learning rate and weight decay 10^{-3}; batch 2,048; at most 12 epochs; patience 3 on envelope MSE
DTU	candidate fusion	$w \in \{0, 0.5, 1, 2, 4, 8, 16, 32, 64, 128\}$ selected by leaving one participant out
EEGMMI	context logistic	$\ell_2 \in \{0.01, 0.1, 1, 10, 100, 1000\}$
EEGMMI	EEG innovation	$\ell_2 \in \{1, 10, 100, 1000, 10000\}$ with context logit as fixed offset
EEGMMI	EEG-only comparator	the same ℓ_2 grid selected independently on EEG-only validation loss
EEGMMI	protocol lookup	smoothing $\in \{0, 0.1, 0.5, 1, 2, 5, 10, 20, 50\}$; selected by validation
ECoG	ridge forecasting	$\lambda \in \{0.1, 1, 10, \ldots, 10^6\}$; validation normalized MSE

The reference environment uses Python 3.10 or newer, NumPy 1.26 or newer, SciPy 1.11 or newer, scikit-learn 1.4 or newer, pandas 2.1 or newer, Matplotlib 3.8 or newer, and PyTorch 2.3 or newer. Deep model runs use no pretraining and were completed on one 16 GB RTX 5080 GPU.

For DTU and EEGMMI, we compute 20,000 nonparametric bootstrap draws from means by participant, conditional on the realized cross-fitted predictions. Synthetic intervals are percentile ranges over complete generator replicates. ECoG intervals use 10,000 participant bootstrap draws after averaging the four scored fingers within participant. All paired contrasts resample paired differences within participant.

Replacement diagnostics use frozen predictors. DTU matches participant, test fold, and the fitted context probability p_t. EEGMMI matches participant and run, then coarse trial-position bins for schedule context or the preceding one or two cues for history context. ECoG exchanges one-second blocks matched on level and movement predicted from context. These strata preserve selected context features; they do not define exact draws from $P(X \mid C)$.

E DTU AUDITORY ATTENTION DETAILS AND ADDITIONAL RESULTS

E.1 DATA, PREPROCESSING, AND CANDIDATE FUSION

The DTU corpus provides 60 fifty-second trials for each of 18 participants (Fuglsang et al., 2017). We use the released preprocessed 64-channel EEG and speech envelopes, average adjacent samples to 32 Hz, and retain complete trials. Official attended/unattended storage is restored to a fixed male/female order using the released event code. This operation defines two exchangeable candidate streams. At evaluation, the current trial's attended identity is used for scoring; the primary context receives the previous trial's true label as feedback, including when that trial belongs to a held-out fold. This is a supervised evaluation setting that requires external label feedback; an autonomous decoder would need an available substitute for that information. The released preprocessing func-

tions retain event order when selecting, splitting, and appending trials. The attended label changes one to three times per participant over the retained sequence, with a mean of 2.33 changes.

For envelope time t, the decoder at positive lags receives nine EEG samples at lags 125–375 ms and the temporal control receives the matched EEG-leading -375 to -125 ms window. Training features and targets are standardized within fold. Five deterministic trial folds assign 36/12/12 trials to train/validation/test. The decoder reconstructs the attended speech envelope. For a held-out trial, candidate evidence is

$$e = \operatorname{atanh}(r_{\text{male}}) - \operatorname{atanh}(r_{\text{female}}), \tag{E.1}$$

where each r is the correlation between the full trial reconstruction and the corresponding envelope. Let ρ be the training-fold frequency with which consecutive attended labels agree. Under the previous-trial information set, the context probability for trial t is

$$p_t = \begin{cases} \pi_{\text{train}}, & t = 1, \\ \rho, & Y_{t-1} = \text{male}, \\ 1 - \rho, & Y_{t-1} = \text{female}. \end{cases} \tag{E.2}$$

Only consecutive pairs whose two members belong to the training fold estimate ρ. The joint probability is

$$q_1(Y = \text{male} \mid C, X) = \sigma(\operatorname{logit}(p_t) + we)\,, \tag{E.3}$$

and $q_0(Y = \text{male} \mid C) = p_t$. For one evaluation participant, the evidence weight w is selected from out-of-fold predictions of the other 17 participants and is then fixed on the target participant. This construction gives the male and female candidates symmetric roles while using the preceding attended label recorded in the task history. Candidate envelopes enter q_1 through e, and $w = 0$ recovers q_0 exactly. The weaker information set replaces p_t by π_{train}.

The original split assigns trials by index modulo five, with 36 training, 12 validation, and 12 test trials in every fold. The split sensitivity instead holds out each contiguous 12-trial block once and uses the following block for validation. A separate offline diagnostic predicts each test label from the three nearest labeled training positions. This diagnostic can use positions on both sides of a test trial and is reported only as a measure of ordering sensitivity.

For the prevalence-only analysis, the primary participant bootstrap holds selected weights fixed. A sensitivity redraws 17 calibration participants per target, reselects post/pre weights, and jointly resamples evaluation participants. The post minus pre interval changes from [0.123, 0.238] to [0.120, 0.246] bits/trial; the post weight is 8 in 98.2% of draws, so calibration sampling adds little uncertainty relative to variation between participants.

E.2 Trial history and split sensitivity

Table E.1 refits ridge under the two trial splits. Previous-trial context yields nearly identical context accuracy and similar residual gains. The paired temporal contrast remains positive under both splits.

Table E.1: DTU ridge sensitivity with previous-trial context. Intervals bootstrap 18 participants.

Split	Context BAcc	Post NEGain	Pre NEGain	Post−pre [interval]
Interleaved	0.947	0.046	0.004	0.042 [0.024, 0.058]
Contiguous	0.945	0.041	0.009	0.032 [0.019, 0.046]

E.3 Model family and seed robustness

Table E.2 asks whether the positive temporal contrast is tied to one decoder family. Ridge, EEGNet-lite, and the causal TCN all give positive post minus pre NEGain and positive mean replacement gaps. Ridge has the largest mean temporal contrast among the three models. Replacement averages all nonzero cyclic shifts of evidence within each fold and context-probability group. Singleton groups retain their original evidence and contribute zero to the gap; 98.3% of trials have an eligible replacement.

Table E.2: DTU model family comparison with supervised label feedback, identical trials, candidate information, and EEG windows. Deep models use seed 20260821; intervals bootstrap 18 participants.

Model	View	NEGain [interval]	Balanced acc.	Swap gap	Positive participants
Ridge	post	0.046 [0.025, 0.065]	0.954	0.070	14/18
Ridge	pre	0.004 [−0.005, 0.011]	0.949	0.010	12/18
EEGNet-lite	post	0.020 [0.010, 0.031]	0.948	0.028	14/18
EEGNet-lite	pre	0.000 [−0.002, 0.002]	0.947	0.001	9/18
Causal TCN	post	0.018 [0.002, 0.033]	0.949	0.037	13/18
Causal TCN	pre	0.003 [−0.007, 0.011]	0.949	0.003	13/18
Ridge	post−pre	0.042 [0.024, 0.058]	–	–	–
EEGNet-lite	post−pre	0.020 [0.010, 0.030]	–	–	–
Causal TCN	post−pre	0.016 [0.002, 0.027]	–	–	–

The five complete training reruns vary initialization, minibatch order, and early stopping. We reuse their frozen predictions and select evidence weights separately for supervised label feedback. Table E.3 reports the primary context; the corresponding prevalence-only averages are 0.067 and 0.117 bits/trial. Seed averages are formed within each participant before resampling the 18 participants. Across 900 fits per family, the median selected epoch is two for EEGNet-lite and three for TCN. The selected checkpoint reaches the 12-epoch limit in 0/900 and 5/900 fits, respectively.

Table E.3: DTU post minus pre NEGain (bits/trial) with supervised previous-label feedback. Intervals use 10,000 participant resamples.

Training seed	EEGNet-lite	Causal TCN
20260821	0.0201 [0.0098, 0.0302]	0.0171 [0.0030, 0.0285]
20260831	0.0163 [0.0035, 0.0288]	0.0132 [-0.0038, 0.0291]
20260841	0.0103 [-0.0007, 0.0215]	0.0168 [0.0090, 0.0247]
20260851	0.0138 [0.0040, 0.0230]	0.0138 [-0.0011, 0.0274]
20260861	0.0119 [0.0006, 0.0224]	0.0137 [-0.0058, 0.0283]
Average across seeds	0.0145 [0.0052, 0.0234]	0.0149 [0.0019, 0.0253]

E.4 Evidence accumulation

The ridge decoder is fit once per fold; candidate correlations are then recomputed from the first 1, 2, 5, 10, and 20 seconds or the complete held-out reconstruction. Table E.4 asks whether log evidence and hard decisions evolve together under the prevalence-only context. The contrast is positive at one second, while balanced accuracy first exceeds 0.60 at 20 seconds. Thus weak probabilistic evidence appears before it is large enough to change most trial decisions.

Table E.4: DTU evidence accumulation with the prevalence-only context using one fitted ridge decoder.

Prefix	Post NEGain	Pre NEGain	Post−pre [interval]	Post balanced acc.	Swap gap
1 s	0.011	−0.000	0.011 [0.001, 0.021]	0.535	0.028
2 s	0.016	−0.000	0.016 [0.002, 0.029]	0.535	0.028
5 s	0.022	−0.000	0.022 [0.004, 0.039]	0.532	0.049
10 s	0.037	0.008	0.029 [0.002, 0.055]	0.558	0.099
20 s	0.089	0.015	0.074 [0.031, 0.118]	0.635	0.231
Full	0.213	0.034	0.179 [0.123, 0.238]	0.707	0.478

The supervised feedback context resolves a different duration dependence. The one-second contrast is near zero, and the first four prefixes have intervals crossing zero. The 20-second and full-record contrasts are positive (Table E.5). The full usable reconstruction spans 48.625 seconds after lag

alignment. The context balanced accuracy is 0.947 at every duration, so changes in hard decisions are small.

Table E.5: DTU ridge evidence accumulation with supervised previous-label feedback. Gains are bits/trial; intervals use 10,000 participant resamples.

Prefix	Post gain	Pre gain	Post−pre [interval]	Joint BAcc
1 s	-0.0001	0.0000	-0.0001 [-0.0035, 0.0033]	0.949
2 s	0.0027	-0.0013	0.0041 [-0.0004, 0.0082]	0.948
5 s	0.0024	0.0029	-0.0005 [-0.0066, 0.0049]	0.950
10 s	0.0042	0.0055	-0.0014 [-0.0109, 0.0066]	0.950
20 s	0.0158	0.0020	0.0138 [0.0071, 0.0200]	0.950
Full (48.625 s)	0.0456	0.0036	0.0420 [0.0247, 0.0581]	0.954

F EEGMMI MOTOR IMAGERY DETAILS AND ADDITIONAL RESULTS

F.1 DATA, FEATURES, AND OUTER FOLD PREDICTION

We use participants 1–40 from the PhysioNet EEG Motor Movement/Imagery corpus (Schalk et al., 2004). R04 provides unlabeled normalization for each participant; imagined left and right fist events from R08 and R12 provide the prediction trials. Five deterministic outer folds allocate every participant once to a test fold of eight participants and once to a validation fold of eight participants; the remaining 24 form the training fold.

For each one-second window, a periodogram with a Hann taper yields log power in 4–8, 8–10, 10–13, 13–20, 20–30, and 30–40 Hz bands, together with total log power and log mean absolute first difference. The primary motor view concatenates the 8–30 Hz bands over non-midline FC, C, and CP channels. The homologous laterality view subtracts nine left/right FC, C, and CP channel pairs; the anatomical control uses non-FC frontal channels. Each feature is standardized by the same participant's unlabeled R04 mean and standard deviation.

The three-second representation averages the feature vectors from 0.5–1.5, 1.5–2.5, and 2.5–3.5 s. Its temporal control averages the symmetric -3.5–-2.5, -2.5–-1.5, and -1.5–-0.5 s windows. Context regularization is selected on validation participants, then the context model is refit on training plus validation participants. The EEG innovation model is fit with the context logit as a fixed offset, so its coefficient estimates the incremental neural log odds under the declared information set. The EEG-only comparator selects its regularization independently using EEG-only validation loss before the same development-set refit.

F.2 CONTEXT AND ANATOMICAL CONTROLS

Schedule context contains a run indicator and linear, quadratic, and two Fourier harmonics of trial position within a run. The model with one cue adds the previous cue and an indicator for the first trial. Cue history context adds the previous two cues, their interaction, indicators for the first and second trials, and an interaction between the previous cue and run. Table F.1 asks whether the result conditioned on schedule is temporally and anatomically concentrated. Motor post-cue evidence exceeds its pre-cue and frontal counterparts. Under cue history, the spectral joint model has similar accuracy to the context model, and its remaining gain is small.

Table F.1: EEGMMI neural input controls. Values are mean NEGain by participant in bits/trial; intervals bootstrap predictions for 40 unseen participants.

Context	View	Time	NEGain [interval]	Swap gap	Joint balanced acc.
Schedule	motor μ/β	post	0.0166 [0.0028, 0.0305]	0.0381	0.578
Schedule	motor μ/β	pre	0.0007 [−0.0011, 0.0027]	0.0016	0.524
Schedule	laterality	post	0.0141 [−0.0068, 0.0327]	0.0472	0.571
Schedule	frontal μ/β	post	0.0007 [−0.0051, 0.0067]	0.0101	0.532
Cue history	motor μ/β	post	0.0068 [−0.0017, 0.0162]	0.0104	0.763
Cue history	motor μ/β	pre	−0.0005 [−0.0012, 0.0002]	−0.0004	0.761
Cue history	laterality	post	−0.0063 [−0.0245, 0.0090]	0.0119	0.746
Cue history	frontal μ/β	post	−0.0013 [−0.0042, 0.0013]	0.0003	0.762

F.3 Context capacity and calibration

The protocol lookup uses run, trial position, and two preceding cues; validation selects its smoothing strength. Validation-fitted affine logit calibration is applied separately to context and joint predictions. Table F.2 tests sensitivity to representation and probability scale. Lookup lowers context NLL by 0.1301 [0.1167, 0.1437] bits/trial. The paired linear-minus-lookup gain difference is 0.0037 [−0.0046, 0.0122], leaving the reduction uncertain. Calibration changes the two point estimates by 0.0005 and −0.0001 bits/trial.

Table F.2: EEGMMI context capacity and calibration sensitivity with the three-second motor view.

Context model	Calibration	Context NLL	Pre NEGain	Post NEGain [interval]	Post−pre
Linear cue history	raw	0.6899	−0.0005	0.0068 [−0.0018, 0.0161]	0.0073
Linear cue history	validation	0.6975	−0.0001	0.0073 [−0.0015, 0.0174]	0.0074
Protocol lookup	raw	0.5598	−0.0004	0.0031 [−0.0060, 0.0122]	0.0035
Protocol lookup	validation	0.5458	−0.0001	0.0030 [−0.0014, 0.0077]	0.0031

F.4 History, run, and sensitivity to observation duration

Table F.3 separates three questions. First, most absorption occurs after adding one previous cue: post gain drops by 0.0085 [0.0025, 0.0141] bits/trial, whereas adding the second cue changes it by only 0.0013 [−0.0041, 0.0066]. Second, R08 and R12 yield temporal contrasts in the same direction, with a paired run difference whose interval spans zero. Third, the symmetric observation duration curve grows from one to three seconds; the three-second aggregate gives a clearer contrast than the one-second estimate. The individual one-second windows do not identify a distinct post-cue peak.

Table F.3: EEGMMI context, run, and sensitivity to symmetric observation duration.

Question	Condition	Pre NEGain	Post NEGain	Post−pre [interval]
Context ladder	schedule	0.0007	0.0166	0.0159 [0.0025, 0.0296]
	+ previous cue	−0.0004	0.0081	0.0085 [−0.0027, 0.0197]
	+ history with two cues	−0.0005	0.0068	0.0073 [−0.0015, 0.0167]
Run replication	R08, schedule	0.0018	0.0151	0.0133 [−0.0040, 0.0299]
	R12, schedule	−0.0004	0.0182	0.0185 [−0.0010, 0.0390]
Observation budget	1 s, schedule	0.0015	−0.0003	−0.0018 [−0.0215, 0.0141]
	2 s, schedule	0.0032	0.0141	0.0109 [−0.0026, 0.0246]
	3 s, schedule	0.0007	0.0166	0.0159 [0.0024, 0.0294]

The prespecified one-second temporal curve contains three pre-cue and three post-cue windows. Its mean post minus pre contrast is −0.0011 [−0.0074, 0.0049] bits/trial under schedule context and 0.0024 [−0.0009, 0.0059] under cue history. The symmetric budget analysis is reported separately in Table F.3.

F.5 Initial neural model comparison on matched trials

The initial comparison uses spectral logistic regression and a compact EEGNet-8,2 adaptation (Lawhern et al., 2018) under the linear cue-history and protocol-lookup contexts. The comparison holds the 40 participants, five outer folds, trial labels, 18 non-midline FC/C/CP channels, and observation windows fixed. Pre-cue windows span $[-3.5, -0.5]$ s and post-cue windows $[0.5, 3.5]$ s. For EEGNet, each 480-sample window is projected onto Fourier frequencies $8 \leq f < 30$ Hz. Channel means and scales use that participant's unlabeled R04 epochs only, separately for each temporal view. Thus the comparison varies the neural representation and model family together.

EEGNet uses eight temporal filters of length 80, depth multiplier two, 16 separable filters of length 16, ELU, batch normalization, average pools of four and eight, and dropout 0.5. This initial implementation omitted the published spatial and output max-norm constraints; Appendix F.6 evaluates the constrained implementation. A bias-free linear head adds one neural logit to the frozen context logit. Its zero initialization recovers the context predictor. Context fitting follows the original capacity experiment, including participant cross-fitting for lookup training offsets. An independent EEG-only network uses a zero context logit. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, and batches of 64. Validation participant-mean log loss selects the epoch, with a maximum of 100 and patience 15, allowing epoch zero. We then refit on 32 development participants for that fixed epoch count. Three seeds (20260916, 20260926, 20261006) are averaged within participant; 10,000 paired participant resamples give intervals. Trial predictions, stopping epochs, and validation curves are included in the supplement.

Table F.4: Initial EEGMMI post-cue neural comparison without max-norm constraints. Context NLL and gain use bits/trial; joint BAcc is balanced accuracy. EEGNet results average three seeds within participant.

Neural model	Context model	Context NLL	NEGain [interval]	Joint BAcc
Spectral logistic	Linear history	0.6899	0.0068 [-0.0018, 0.0160]	0.763
Spectral logistic	Lookup	0.5598	0.0031 [-0.0059, 0.0122]	0.743
EEGNet	Linear history	0.6899	-0.3671 [-1.2525, 0.0922]	0.782
EEGNet	Lookup	0.5598	-0.0808 [-0.3700, 0.0787]	0.796

EEGNet improves EEG-only balanced accuracy to 0.654 [0.624, 0.683], compared with 0.565 for the spectral model. Its conditional log scores show substantial variation across participants. Participant 21 has seed-averaged post-cue gain of -17.3000 bits/trial with linear history, accounting for most of the negative mean. This participant remains in all summaries. The paired EEGNet-minus-spectral gain differences are -0.3739 [-1.2613, 0.0853] with linear history and -0.0839 [-0.3727, 0.0757] with lookup context. The EEGNet linear-minus-lookup difference is -0.2863 [-0.8906, 0.0208]. The direction of the context-capacity point estimate therefore does not generalize across these neural models.

Negative fitted gain records prediction loss relative to the context model. In Equation 3, it is compatible with positive conditional information when the joint predictor incurs sufficiently greater regret. The initial comparison motivates the input-scale diagnosis and matched follow-up below.

F.6 Input scale and probability calibration

Diagnosis. Participant 21's C5 channel changes markedly between calibration and prediction runs. After the original R04 normalization, its post-cue root mean square amplitude is 1.0 in R04, 57.9 in R08, and 49.8 in R12. The shift is also present in the raw recording. R04 scaling therefore leaves extreme input values in the subsequent runs. We specified the following exploratory sensitivity analysis after identifying this failure, retaining all 40 participants, trials, electrodes, and three training seeds.

Matched protocol. Both EEGNet variants use the architecture above with spatial-filter norm bounded by 1 and binary-head weight norm bounded by 0.25, following the constraints in the published implementation (Lawhern et al., 2018). One variant receives the original normalized

waveform. The other clips amplitudes symmetrically at the 99.5th percentile of absolute training values. This single threshold is estimated separately for each fold and temporal view from the 24 training participants' R08/R12 windows and applied to every channel and participant. The resulting thresholds range from 3.71 to 4.35. The quantile is fixed before fitting, with no threshold search.

We retain the five outer folds and the original optimizer, stopping rule, windows, and frequency range. Each fold fits on 24 participants, selects the checkpoint on eight validation participants, and tests on eight distinct participants. Context models and the spectral comparator use the same split. For both predictors, we fit separate positive temperatures to validation log loss (Guo et al., 2017), optimizing log temperature over $[-5, 5]$. The selected checkpoint remains fixed during calibration and testing; validation participants are not added to training. Thus the comparison of input variants is matched, while the initial 32-participant refit above is a separate reference. We average three seeds within participant before 10,000 paired participant bootstrap resamples.

Table F.5: EEGMMI exploratory follow-up on matched 24/8/8 participant splits. Post-cue gains and 95% participant intervals use bits/trial ($n = 40$). Temperature scaling preserves balanced accuracy (BAcc). Both EEGNet variants include weight constraints.

Neural predictor	Raw gain [interval]	Calibrated gain [interval]	BAcc
Linear cue history			
Spectral logistic	0.007 [−0.001, 0.016]	0.008 [0.000, 0.018]	0.761
EEGNet, original input	−0.162 [−0.644, 0.092]	−0.193 [−0.738, 0.093]	0.779
EEGNet, bounded input	0.082 [0.040, 0.122]	0.082 [0.037, 0.123]	0.783
Protocol lookup			
Spectral logistic	−0.001 [−0.012, 0.008]	0.001 [−0.003, 0.005]	0.733
EEGNet, original input	−0.044 [−0.251, 0.073]	−0.030 [−0.176, 0.052]	0.790
EEGNet, bounded input	0.062 [0.027, 0.095]	0.048 [0.024, 0.071]	0.809
EEG only (zero context logit)			
Spectral logistic	0.013 [0.001, 0.027]	0.014 [0.002, 0.027]	0.568
EEGNet, original input	−0.189 [−0.781, 0.121]	−0.237 [−0.924, 0.122]	0.655
EEGNet, bounded input	0.099 [0.042, 0.148]	0.097 [0.039, 0.151]	0.672

Results. With the original input, constrained EEGNet still gives negative mean gains; temperature scaling also leaves both conditional means negative (Table F.5). Bounded input gives positive gains with both contexts before and after calibration. Within this matched protocol, participant 21's uncalibrated linear-history gain changes from -9.2741 to -0.4057 bits/trial; its negative result remains in all summaries. The bounded EEGNet exceeds the matched spectral comparator by 0.0750 [0.0304, 0.1136] bits/trial with linear history and 0.0632 [0.0307, 0.0939] with lookup, before calibration. EEG-only balanced accuracy is 0.672 [0.641, 0.705], versus 0.568 [0.544, 0.595] for spectral decoding.

Pre-cue raw gains with bounded input are -0.0022 [-0.0052, 0.0005] with linear history and -0.0009 [-0.0023, 0.0004] with lookup. The paired post minus pre contrasts are 0.0842 [0.0424, 0.1224] and 0.0627 [0.0282, 0.0950], respectively. After calibration they remain positive at 0.0854 [0.0420, 0.1257] and 0.0479 [0.0243, 0.0703]. These temporal comparisons support a post-cue contribution under both context representations.

The linear-minus-lookup gain contrast depends on calibration: it is 0.0201 [-0.0056, 0.0428] before and 0.0341 [0.0018, 0.0606] after calibration. Because calibration changes both context and joint losses, a smaller gain need not imply a worse joint predictor. The follow-up establishes sensitivity to input handling and model fit; its preprocessing choice requires evaluation in a new cohort for confirmatory evidence. Saved outputs include every input variant, calibration condition, participant, and seed.

G CONTINUOUS ECOG SENSITIVITY TO THE INFORMATION SET

BCI Competition IV-4 provides a continuous target case with three participants (Tangermann et al., 2012). We forecast four scored finger flexion targets 80, 200, and 400 ms into the future. The deployment context uses past cross-fitted broadband state estimates at 40–1280 ms lags. At the shortest lag, the past decoder window overlaps the current 240 ms window by 200 ms. The comparison therefore measures information beyond a compressed past neural state, and does not isolate activity in the final 40 ms. Cross-fitting uses a temporal gap equal to the longest history. The oracle comparison replaces estimated states with measured past kinematics at the same lags.

Raw ECoG is converted into causal 240 ms windows at 25 Hz. A fixed random orthogonal projection drawn once per participant maps every feature block to at most 24 spatial dimensions. Low-frequency features concatenate 1–30 and 30–60 Hz log power; high-gamma features concatenate 60–100 and 100–200 Hz log power with log first-difference power; broadband combines all five blocks. Training standardization is applied before ridge regression. We score the four competition fingers and use the official chronological test segment (Tangermann et al., 2012); the ring finger channel excluded by the competition is not scored.

The legal state decoder is selected on the first 80% of the official training segment. Five contiguous cross-fitting folds with a 1,280 ms gap produce out-of-fold training states; the same decoder predicts test states without test kinematics. Context and neural innovation ridge penalties are selected on the remaining training segment. Conditional replacements exchange one-second neural blocks matched on level and movement predicted from context.

We score each finger with a scalar Gaussian density, then average the four finger gains within participant. The unit is bits/sample/finger. Broadband mean gains under decoded-state history are 0.0153, 0.0152, and 0.0110 at 80, 200, and 400 ms. Measured kinematic history gives means of 0.0039, 0.0044, and 0.0031. Tables G.1–G.2 show the three participants individually and their means. These are descriptive results for the observed participants.

Table G.1: ECoG NEGain with decoded-state history (bits/sample/finger). S1–S3 are participant means over four fingers; the last column averages them.

Horizon	Neural view	S1	S2	S3	Mean
80 ms	Broadband	0.0129	0.0048	0.0282	0.0153
80 ms	High gamma	0.0124	0.0054	0.0194	0.0124
80 ms	Low frequency	0.0026	0.0004	0.0042	0.0024
200 ms	Broadband	0.0060	0.0078	0.0319	0.0152
200 ms	High gamma	0.0180	0.0074	0.0194	0.0149
200 ms	Low frequency	0.0001	0.0020	0.0043	0.0021
400 ms	Broadband	0.0024	0.0088	0.0217	0.0110
400 ms	High gamma	0.0036	0.0080	0.0222	0.0113
400 ms	Low frequency	0.0008	0.0026	0.0020	0.0018

Table G.2: Broadband ECoG NEGain with measured kinematic history, an oracle context (bits/sample/finger). The lags match the decoded-state history.

Horizon	S1	S2	S3	Mean
80 ms	0.0053	0.0038	0.0025	0.0039
200 ms	0.0048	0.0053	0.0033	0.0044
400 ms	0.0009	0.0068	0.0017	0.0031

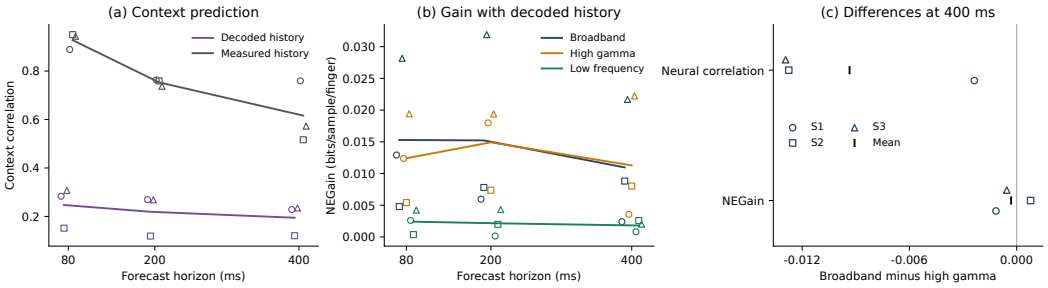

Figure G.1: Individual participants in continuous ECoG forecasting. (a) Context correlation; (b) gain by band with decoded-state history; (c) broadband minus high gamma at 400 ms. Marker shapes identify participants. Lines in (a)–(b) and black marks in (c) indicate means. Gains use bits/sample/finger; correlation is dimensionless.