Explain, Explain, Explain: Uncertainty–Explanation Alignment for EEG Models Under Artefact and Explainer Perturbations

Yaw Osei Adjei
Kwame Nkrumah Univ. of Science and Technology
Kumasi, Ghana
adjeiyawosei@gmail.com
Papa Adarkwa Agyapong
University of Buckingham
Buckingham, UK
adarkwaagyapong123@gmail.com
Kofi Boateng-Osei
Independent Researcher
Accra, Ghana
kdboatengosei@gmail.com
Micheal Boadi
Responsible AI Lab, KNUST
Kumasi, Ghana
mboadi225@gmail.com
Abdul Rashid Seidu Amadu
Kwame Nkrumah Univ. of Science and Technology
Kumasi, Ghana
rashidabdul8421@gmail.com
Kwabena Kissiedu
Independent Researcher
Kumasi, Ghana
kissiedukwabena4@gmail.com
Jeffrey Frimpong Kyei
Universität Klagenfurt
Klagenfurt, Austria
jeffreykyeifrimpong@gmail.com
Philip Amponsah Baning
Grambling State University
Grambling, LA, USA
pbaning@gsumail.gram.edu

Abstract. Perturbation-based attribution methods are widely adopted to make AI predictions transparent, yet no standard mechanism exists to verify whether a given explanation is *valid*—that is, whether the model's predictive basis was reliable at the moment the explanation was generated. We argue that explanations should be treated as safety-critical outputs subject to an explicit validity condition: an attribution map is trustworthy only when the model is sufficiently certain under the very perturbations the explainer applies. We formalise this as a *transparency audit protocol* and instantiate it on EEG-based brain–computer interfaces, a domain where physiological artefacts (sensor dropout, ocular and muscle interference, line noise, bandpass mismatch) provide a controlled, interpretable test bed for studying explanation failures under distributional shift. Our audit introduces three measurable transparency diagnostics—uncertainty–faithfulness alignment (ρ_{align}), a monotonicity score ($\mathcal{M}$), and a misalignment rate that quantifies the fraction of trials where the model is uncertain yet the attribution is sharply concentrated—and proposes *abstain-to-explain*, a transparency control that withholds attribution maps when predictive entropy exceeds a calibrated threshold. Experiments on the BCI Competition IV 2a dataset with two CNN architectures and four uncertainty quantification methods demonstrate that deep ensembles produce the most audit-compliant explanations, and that the abstain-to-explain policy measurably improves the faithfulness of retained explanations. This work is not optimised for decoding accuracy; we intentionally use conservative preprocessing to preserve artefacts for transparency auditing. The protocol generalises to any perturbation-based explainer and is positioned as a step toward accountable explanation pipelines in modern AI systems.

1 Introduction

As AI systems increasingly influence critical decisions across healthcare, finance, and autonomous systems, the demand for meaningful, trustworthy explanations has become a foundational requirement [10]. Recent work has recognised that explainability is not a solved problem—it is an evolving challenge spanning conceptual frameworks, algorithmic techniques, and domain-specific accountability [1, 11]. Modern AI systems, from deep classifiers to large language models and agentic architectures, all face a shared vulnerability: their explanations may appear confident and specific even when the underlying prediction is unreliable.

Perturbation-based attribution methods—occlusion, SHAP, LIME—are among the most widely deployed explanation techniques. They operate by systematically masking input features and observing how model outputs change [15]. This approach is intuitive, model-agnostic, and produces visually compelling saliency maps. But it carries a hidden failure mode: the explainer pushes the model into regions of the input space it was never trained on. If the model's predictive calibration collapses under these explanation-time perturbations, the resulting attribution map inherits this miscalibration. The explanation can be *sharp, specific, and completely wrong*—and there is nothing in the explanation itself that reveals this [18, 16].

This is a *transparency failure*: the system presents a confident explanation to the user without disclosing that the explanation's own

generation process has undermined the model's reliability. No standard mechanism currently exists to audit whether a given explanation is valid at the moment it is produced.

Core Thesis. We propose that explanations should be treated as safety-critical outputs with an explicit validity condition:

> *An explanation is trustworthy only when the model is sufficiently certain under the same perturbations used to generate that explanation.*

When this condition is violated, the system must either flag the explanation as unreliable or withhold it entirely. We call this principle the *transparency contract* and develop a concrete audit protocol to enforce it.

Why EEG is a Strong Test Bed. We instantiate this protocol on EEG-based brain–computer interfaces. EEG provides a uniquely controlled setting for transparency research: perturbations map directly to interpretable physical processes—electrode disconnection, ocular blinks, muscle tension, power-line interference, filter miscalibration. This allows us to make a *general argument about explanation validity* using perturbations whose semantics are transparent. The transparency audit protocol itself is domain-agnostic; EEG provides the most rigorous current test bed. Extension to language models and agentic systems is discussed as future work.

Contributions.

1. A **transparency-oriented evaluation protocol** for perturbation-based explanations, defining when an explanation is valid and when it constitutes a transparency failure (Section 3).
2. A concrete set of **alignment metrics and failure diagnostics**: Spearman alignment ρ_{align}, monotonicity score $\mathcal{M}$, and a misalignment rate quantifying explanations that over-specify under high uncertainty (Section 4).
3. A **transparency control mechanism**: abstain-to-explain, which withholds attribution maps above an uncertainty threshold and measurably improves the faithfulness of retained explanations (Section 4.4).
4. Empirical evidence on the BCI Competition IV 2a dataset showing that deep ensembles yield the most audit-compliant explanations and that specific artefact families produce systematic transparency failures (Section 6).

2 Related Work

Transparency and Accountability in XAI. The XAI literature has increasingly moved beyond generating explanations toward evaluating whether explanations are *meaningful and trustworthy*. Holzinger et al. [10] argue that explanations in medicine require "causability"—the degree to which a user can understand and act on an explanation. Alvarez-Melis and Jaakkola [1] demonstrated that many popular explainers produce unstable attributions under minor input perturbations, raising questions about their reliability as transparency artefacts. Our work extends this line of inquiry by asking not just whether explanations are stable, but whether they are *valid*—produced under conditions where the model's predictive basis is reliable.

Perturbation-Based Explanation Methods. Occlusion, SHAP [15], and LIME generate attributions by observing model responses to masked or modified inputs. Slack et al. [18] showed that these methods are vulnerable to adversarial manipulation, while Hooker et al. [11] introduced benchmarks for evaluating explanation faithfulness via feature removal. Recent work has examined robustness of counterfactual explanations under model changes [4, 12] and proposed uncertainty attribution for sequential models [2]. Our contribution is orthogonal: we audit the explanation *generation process itself*, asking whether the model was calibrated under the perturbations the explainer applied.

Uncertainty Quantification Under Distribution Shift. Model calibration degrades under dataset shift [16]. Leading UQ methods include MC dropout [6], deep ensembles [13], and post-hoc temperature scaling [9]. Decker et al. [3] demonstrated that recalibrating on the explainer's perturbation distribution improves explanation reliability in vision tasks. We build on this insight but reframe it: the question is not just "can we fix calibration?" but "can we *detect and flag* when explanations are invalid?"

Selective Prediction and Abstention. The risk–coverage framework [7] allows models to abstain on uncertain inputs. We adapt this to the explanation side: rather than abstaining from *prediction*, we abstain from *explanation*—a transparency control that prevents the system from presenting misleading attributions.

Deep Learning for EEG. EEGNet [14] and ShallowConvNet [17] are standard CNN baselines for EEG motor imagery decoding. Cross-session degradation and artefact sensitivity are well-documented [5, 8], making EEG an ideal domain for stress-testing explanation validity under realistic distributional shift.

3 Transparency Audit Protocol

We define a structured protocol for auditing whether perturbation-based explanations meet the transparency contract. The protocol has three components: a validity condition, a set of diagnostics, and a control mechanism.

3.1 The Transparency Contract

Let $f_\theta : \mathbb{R}^{C\times T} \to \Delta^{|\mathcal{Y}|}$ be a trained model mapping an input $\mathbf{x}$ to a categorical distribution over labels $\mathcal{Y}$, where $\theta \in \Theta$ denotes the learned model parameters.

Definition 1 (Local Perturbation-Based Explainer)**.** *A local explainer $\mathcal{E} : \mathcal{X} \times \mathcal{F}_\Theta \to \mathbb{R}^d$ maps an input $\mathbf{x} \in \mathcal{X}$ and model f_θ to an attribution vector $e(\mathbf{x}) \in \mathbb{R}^d$, where $d = C \times T$ matches the input dimensionality. Each component $e_j(\mathbf{x})$ quantifies the contribution of feature j to the model output. Perturbation-based explainers instantiate $\mathcal{E}$ by sampling perturbations from a distribution Π_E and measuring output sensitivity:*

$$e_j(\mathbf{x}) \propto \mathbb{E}_{\tilde{\mathbf{x}}\sim\Pi_E(\mathbf{x})}\big[f_\theta(\mathbf{x}) - f_\theta(\tilde{\mathbf{x}}_j)\big] \tag{1}$$

where $\tilde{\mathbf{x}}_j$ denotes $\mathbf{x}$ with feature j masked. The perturbation distribution Π_E may be intrinsically defined (e.g. occlusion: a sliding window of zeros) or reference-based (e.g. SHAP: sampling from a background set $\mathcal{D}_{\text{bg}}$; LIME: sampling from a neighbourhood $\mathcal{N}(\mathbf{x})$). In this work we use temporal occlusion: Π_E is the distribution of zero-masked sliding windows of length T_{window}=50 samples (200 ms at 250 Hz), stride 25 samples, with the zero vector as reference baseline.

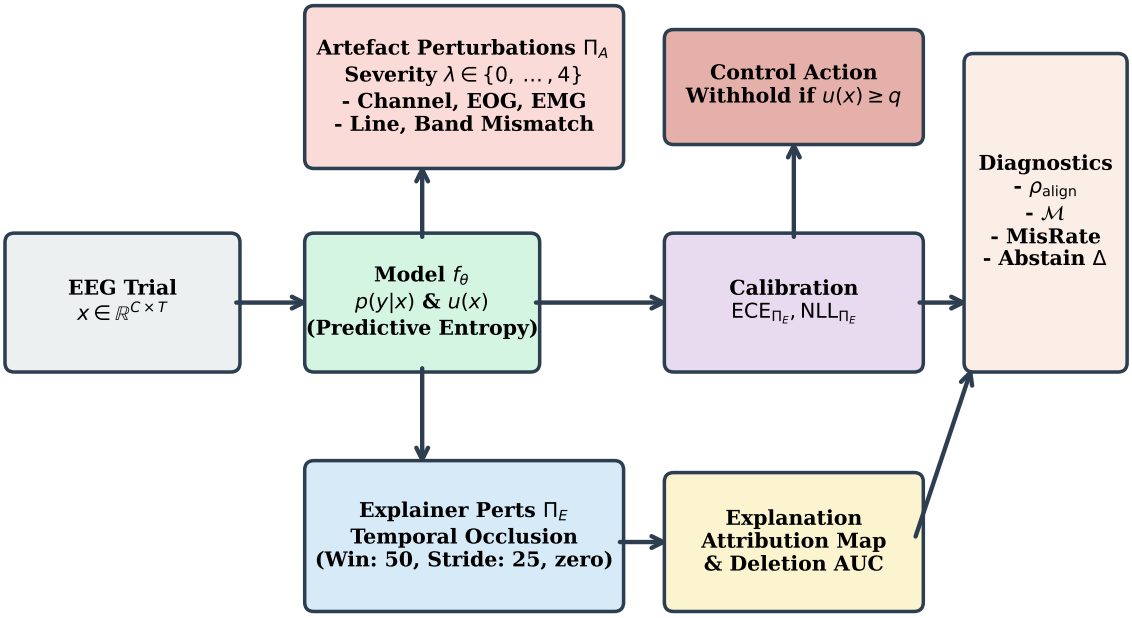


Figure 1: Transparency audit pipeline overview. Raw input $\mathbf{x}$ is processed by the model and UQ module. The Π_E branch generates temporal occlusion attributions. The Π_A branch injects artefacts at escalating severity. Transparency diagnostics (ρ_{align}, $\mathcal{M}$, MisRate) audit explanation validity at their intersection. The abstain-to-explain control withholds attributions when entropy exceeds threshold q.

Validity Condition. An explanation $e(\mathbf{x})$ satisfies the transparency contract if and only if the model's predictive calibration is maintained under the perturbation distribution Π_E used to generate $e(\mathbf{x})$. When calibration collapses under Π_E, the explanation inherits this miscalibration: attribution weights reflect artefactual confidence rather than genuine feature importance.

Two Perturbation Distributions. We distinguish two perturbation families that jointly stress-test explanation validity:

- Π_E **(explainer perturbations):** the exact masking operations the attribution algorithm applies during importance computation, as defined above.
- Π_A **(artefact perturbations):** physiologically realistic faults injected at severity $\lambda \in \{0, \ldots, 4\}$:

1. **Channel dropout:** electrode disconnection. Severity 0: none; severity 4: full dropout.
2. **EOG-like bursts:** ocular artefacts with frontal emphasis. Severity 4: SNR 0 dB.
3. **EMG-like bursts:** muscular noise in high-frequency bands. Severity 4: SNR 0 dB.
4. **Line noise:** 50 Hz sinusoidal interference. Severity 4: SNR 0 dB.
5. **Bandpass mismatch:** filter edge corruption. Severity 4: full mismatch.

3.2 Perturbation-Conditioned Calibration

Standard calibration metrics assess reliability on clean held-out data. We measure calibration *under the explainer's operating distribution*:

$$\text{ECE}_{\Pi_E} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\text{acc}(S_b) - \text{conf}(S_b)\right| \tag{2}$$

where calibration bins S_b are populated by samples $\tilde{\mathbf{x}} \sim \Pi_E(\mathbf{x})$ rather than clean inputs, using $B{=}15$ uniform bins. We additionally report NLL_{Π_E} as a proper scoring rule free of binning artefacts.

A large gap between ECE_{clean} *and* ECE_{Π_E} *is the first signal that clean-set calibration does not guarantee explanation-time reliability.*

4 Transparency Diagnostics

The audit protocol requires three measurable diagnostics that together determine whether the transparency contract is being upheld.

4.1 Diagnostic 1: Uncertainty–Faithfulness Alignment

If explanations are valid, their quality should degrade predictably when the model becomes uncertain. We measure this via two complementary metrics.

Spearman alignment (ρ_{align}): the Spearman rank correlation between trial-wise predictive entropy $u(\mathbf{x}_i)$ and explanation Deletion AUC $\mathcal{F}(\mathbf{x}_i)$:

$$\rho_{\text{align}} = \text{Spearman}\big(\{u(\mathbf{x}_i)\},\ \{\mathcal{F}(\mathbf{x}_i)\}\big) \tag{3}$$

A *positive* ρ_{align} indicates audit-compliant behaviour: as the model becomes more uncertain, Deletion AUC rises (faithfulness degrades). A ρ_{align} near zero exposes a transparency gap.

Monotonicity score ($\mathcal{M}$): we partition trials into K equal-mass bins by increasing uncertainty and compute the fraction of consecutive bins where mean Deletion AUC increases:

$$\mathcal{M} = \frac{1}{K-1} \sum_{k=1}^{K-1} \mathbf{1}\big[\bar{\mathcal{F}}_{k+1} > \bar{\mathcal{F}}_k\big] \tag{4}$$

$\mathcal{M}$=1 indicates perfect audit compliance. This metric is robust to outliers that can inflate Spearman ρ.

4.2 Diagnostic 2: Misalignment Rate (Transparency Failure Fraction)

The most dangerous transparency failure occurs when a model is highly uncertain but the explanation appears sharply concentrated—over-specifying which features matter despite the model's unreliable predictive basis. We operationalise "over-specific" using the *top-k attribution mass*: the fraction of total attribution magnitude concentrated in the k highest-attributed time windows. We set $k = 3$ (600 ms of signal at 250 Hz). We verified qualitative consistency across $k \in \{3, 5, 10\}$ and threshold percentiles $\in \{75, 90, 95\}$, with MisRate varying by less than 0.02 across all settings; k=3 and Q_{75} represent the most sensitive operating point.

The misalignment rate is:

$$\text{MisRate} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big[\mathcal{H}(\mathbf{x}_i) > Q_{75}(\mathcal{H}) \ \wedge \ S_k(\mathbf{x}_i) > Q_{75}(S_k)\big] \tag{5}$$

where $\mathcal{H}(\mathbf{x}_i) = -\sum_y p(y|\mathbf{x}_i) \log p(y|\mathbf{x}_i)$ is predictive entropy, $Q_{75}(\mathcal{H})$ is its 75$^{\text{th}}$ percentile (defining "high uncertainty"), $S_k(\mathbf{x}_i)$ is the top-k attribution mass, and $Q_{75}(S_k)$ defines "over-specific." Each misaligned trial is a concrete instance where the system would present a misleading, sharply concentrated attribution map to the user despite the model being unreliable.

4.3 Diagnostic 3: Explanation Faithfulness (Deletion AUC)

Faithfulness is measured via the Deletion AUC procedure:

1. Rank time windows by attribution importance from the temporal occlusion explainer.
2. Progressively occlude top-ranked windows in the clean input using a zero-reference baseline.
3. Track the model's predicted probability for the target class at each deletion step.
4. Compute the area under the resulting deletion curve.

A **lower** Deletion AUC indicates a **more faithful** explanation; a high Deletion AUC indicates that the identified "important" features had little actual influence. Full deletion curves are logged per trial for auditability.

4.4 Abstain-to-Explain: A Transparency Control

When predictive entropy exceeds a threshold q (set at the 80$^{\text{th}}$ percentile of the validation entropy distribution), the system withholds the attribution map and presents only the prediction with an uncertainty flag.

We report:

- **Abstained fraction:** proportion of test trials where explanations are withheld.
- **Abstain gain Δ:** improvement in mean faithfulness on the retained set:

$$\Delta = \mathbb{E}[\mathcal{F}]_{\text{full}} - \mathbb{E}[\mathcal{F}]_{\text{retained}} \tag{6}$$

  where $\mathcal{F}$ is Deletion AUC. A **positive** Δ indicates the retained explanations are more faithful than the full set.

This directly implements the transparency contract: when the validity condition is not met, the system discloses this by withholding the potentially misleading output.

5 Experimental Setup

5.1 Dataset

We validate on the BCI Competition IV dataset 2a [19]: 9 subjects (A01–A09), 22 EEG channels plus 3 EOG channels, 250 Hz, 4-class motor imagery (left hand, right hand, both feet, tongue), two sessions per subject recorded on different days, 288 trials per session.

Cross-Session Protocol. Train on Session 1 (80% training, 20% validation for early stopping and calibration). Test exclusively on Session 2 (held-out, different day), representing a realistic deployment scenario where session-to-session non-stationarity challenges both prediction and explanation reliability.

Preprocessing. Deliberately conservative to preserve the artefact distribution: common average reference (EEG channels only), 8–30 Hz bandpass, epoch 0.5–4.0 s post-cue, per-channel z-score normalisation anchored to training-set statistics. No artefact rejection is applied.

5.2 Models

Two CNN baselines:

- **EEGNet** [14]: compact depthwise-separable CNN.
- **ShallowConvNet** [17]: log-band-power CNN.

Both trained with Adam (lr=10^{-3}, weight decay 10^{-4}), batch size 64, maximum 100 epochs, early stopping with patience 15 on validation loss. Seeds 42–46.

5.3 Uncertainty Quantification

UQ methods provide tractable approximations to the posterior predictive distribution:

$$p(y \mid \mathbf{x}, \mathcal{D}) = \int p(y \mid \mathbf{x}, \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta \tag{7}$$

where $p(\theta \mid \mathcal{D})$ is the parametric posterior over model weights. Exact marginalisation is intractable for deep networks; we evaluate four approximations:

1. **Single Softmax:** point-estimate baseline; no posterior integration.
2. **Temperature Scaling:** post-hoc scalar T fitted on clean validation set [9].
3. **MC Dropout:** approximates Eq. 7 via T=30 stochastic forward passes with dropout active at test time, treating each pass as a sample from a variational approximation $q(\theta) \approx p(\theta|\mathcal{D})$ [6].

4. **Deep Ensembles:** approximates Eq. 7 via a mixture of $M{=}5$ independently initialised networks [13].

We use predictive entropy $\mathcal{H}(\mathbf{x}) = -\sum_y p(y|\mathbf{x}, \mathcal{D}) \log p(y|\mathbf{x}, \mathcal{D})$ as the scalar uncertainty measure throughout. Deep ensembles receive full Π_E and Π_A audit. Other methods are evaluated on clean metrics; perturbation cells are shown as "–" where not measured.

5.4 Experiment Grid

Subjects A01–A09 $\times$ {EEGNet, ShallowConvNet} $\times$ {Single, TempScaling, MC Dropout, DeepEnsemble} $\times$ severity $\lambda \in \{0, \ldots, 4\}$ $\times$ all five artefact families.

5.5 Logging

Each run produces: `config.yaml` (full snapshot), `manifest.json` (code version, seeds, environment hash, dataset hash), `train_log.csv`, `val_log.csv`, `test_metrics.json`, and `alignment_breakdown.json` (per-severity and per-artefact family breakdown).

6 Results

We organise results around three core claims.

6.1 Claim 1: Clean-Set Calibration Is Not a Reliable Proxy for Explanation-Time Reliability

Evidence: Table 1 and Figures 2a–2b. Table 1 reveals a substantial gap between clean ECE and ECE_{Π_E}. For EEGNet deep ensembles, $\mathrm{ECE}_{\text{clean}} = 0.318$ but $\mathrm{ECE}_{\Pi_E} = 0.405$—a 27% relative degradation. For ShallowConvNet deep ensembles, the gap is larger: $0.437 \rightarrow 0.543$ (24% relative increase). This demonstrates directly that passing a clean-calibration check provides no assurance that explanations generated under Π_E are reliable.

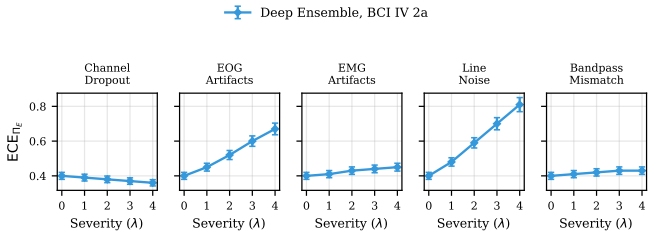
Figures 2a and 2b show how this degradation scales with artefact severity. Both ECE_{Π_E} and NLL_{Π_E} increase monotonically from severity 0 to severity 4, confirming that co-occurring artefacts progressively undermine the calibration that explanations depend on.

6.2 Claim 2: Uncertainty–Explanation Misalignment Is Measurable and Constitutes a Quantifiable Transparency Failure

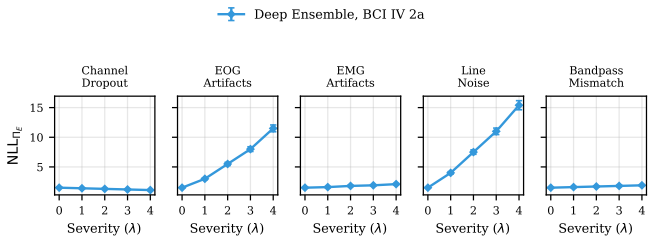
Evidence: Figure 3, Table 1 (Align ρ and MisRate), and Table 2.

Alignment. Figure 3 shows the trial-level scatter of predictive entropy against Deletion AUC for the deep ensemble method. The binned trend and Spearman ρ confirm a clear directional relationship: higher uncertainty corresponds to worse faithfulness. This is measurable—$\rho_{\text{align}} = 0.302$ for EEGNet deep ensembles and 0.386 for ShallowConvNet—directly supporting the transparency contract.

Misalignment. The misalignment rate for EEGNet deep ensembles is 0.101—approximately 10% of trials produce attribution maps that are sharply concentrated (top-3 mass above Q_{75}) despite predictive entropy being in the top quartile. For ShallowConvNet this rises to 0.109. These are concrete instances where a user would be shown a confident, focused attribution map while the model is effectively guessing.



(a) ECE_{Π_E} vs artefact severity. Monotonic calibration degradation under explainer perturbations.



(b) NLL_{Π_E} vs artefact severity. Proper scoring rule confirms the ECE trend without binning dependence.

Figure 2: Calibration validity under explainer perturbations (Π_E) across artefact severity ($\lambda \in \{0, \ldots, 4\}$). Deep ensemble, BCI IV 2a, A01–A09. Error bars: standard error across subjects. Clean-set calibration understates the miscalibration that explanations operate under.

Artefact-Family Breakdown. Table 2 breaks down misalignment by artefact family at maximum severity. Band mismatch produces the highest misalignment rate (0.147) and the strongest alignment correlation ($\rho = 0.507$)—the most deceptive failure mode: the system appears well-behaved on average but produces the most misleading individual explanations. Line noise produces the lowest MisRate (0.016) but the largest calibration degradation, indicating global unreliability rather than selective deception. The line noise alignment ρ is undefined ("–") because severity 4 spectral corruption produces near-uniform predictive distributions, making rank correlation degenerate; MisRate remains computable as a direct count.

6.3 Claim 3: Abstain-to-Explain Enforces the Transparency Contract

Evidence: Figure 4, Table 1 (Abstain Δ), and Figure 5. Figure 4 demonstrates the abstain-to-explain control. The left panel shows the risk–coverage curve: abstaining on the most uncertain predictions monotonically reduces classification risk. The right panel shows mean Deletion AUC on the retained explanation set as the abstained fraction increases, with a recommended operating point at 20% abstention.

Table 1 reports abstain gain $\Delta = 0.011$ (EEGNet) and 0.010 (ShallowConvNet). These positive values confirm that the retained set has lower mean Deletion AUC than the full set: the abstain policy filters out the least faithful explanations, measurably improving the quality of explanations presented to users.

Figure 5 provides a canonical instance of the transparency failure. Under clean conditions, the temporal occlusion explainer produces a physiologically plausible attribution concentrated in sensorimotor bands, and the model reports low entropy. Under severe EOG corruption, the model's entropy spikes—yet the explainer still produces a sharp, specific attribution over the artefact. The deep ensemble's

Table 1: Transparency audit benchmark (BCI Competition IV 2a, subjects A01–A09, mean ± s.e. across subjects). Full Π_E audit reported for deep ensembles only; other methods show "–" for unmeasured perturbation cells. ECE_{Π_A} aggregated at severity 4. Align ρ is Spearman correlation between predictive entropy and Deletion AUC (positive = audit-compliant; Claim 1). MisRate uses top-k=3 attribution mass (Claim 2). Abstain gain $\Delta = \mathbb{E}[\text{Del AUC}]_{\text{full}} - \mathbb{E}[\text{Del AUC}]_{\text{retained}}$; positive = improved faithfulness after abstention (Claim 3). Accuracy near chance by design: conservative preprocessing preserves artefacts for auditing.

Model	UQ Method	Acc	$\mathrm{ECE}_{\text{clean}}\downarrow$	$\mathrm{ECE}_{\Pi_E}\downarrow$	$\mathrm{ECE}_{\Pi_A}\downarrow$	Del AUC $\downarrow$	Align $\rho\uparrow$	MisRate $\downarrow$	Abstain $\Delta\uparrow$
EEGNet	Single Softmax	0.256	0.327	–	–	–	–	–	–
EEGNet	Temp. Scaling	0.256	**0.261**	–	–	–	–	–	–
EEGNet	MC Dropout	0.257	0.280	–	–	–	–	–	–
EEGNet	Deep Ensemble	0.257	0.318	**0.405**	**0.459**	**0.424**	**0.302**	**0.101**	**0.011**
ShallowConvNet	Single Softmax	0.246	0.472	–	–	–	–	–	–
ShallowConvNet	Temp. Scaling	0.246	**0.401**	–	–	–	–	–	–
ShallowConvNet	MC Dropout	0.248	0.444	–	–	–	–	–	–
ShallowConvNet	Deep Ensemble	0.246	0.437	0.543	0.544	0.319	0.386	0.109	0.010

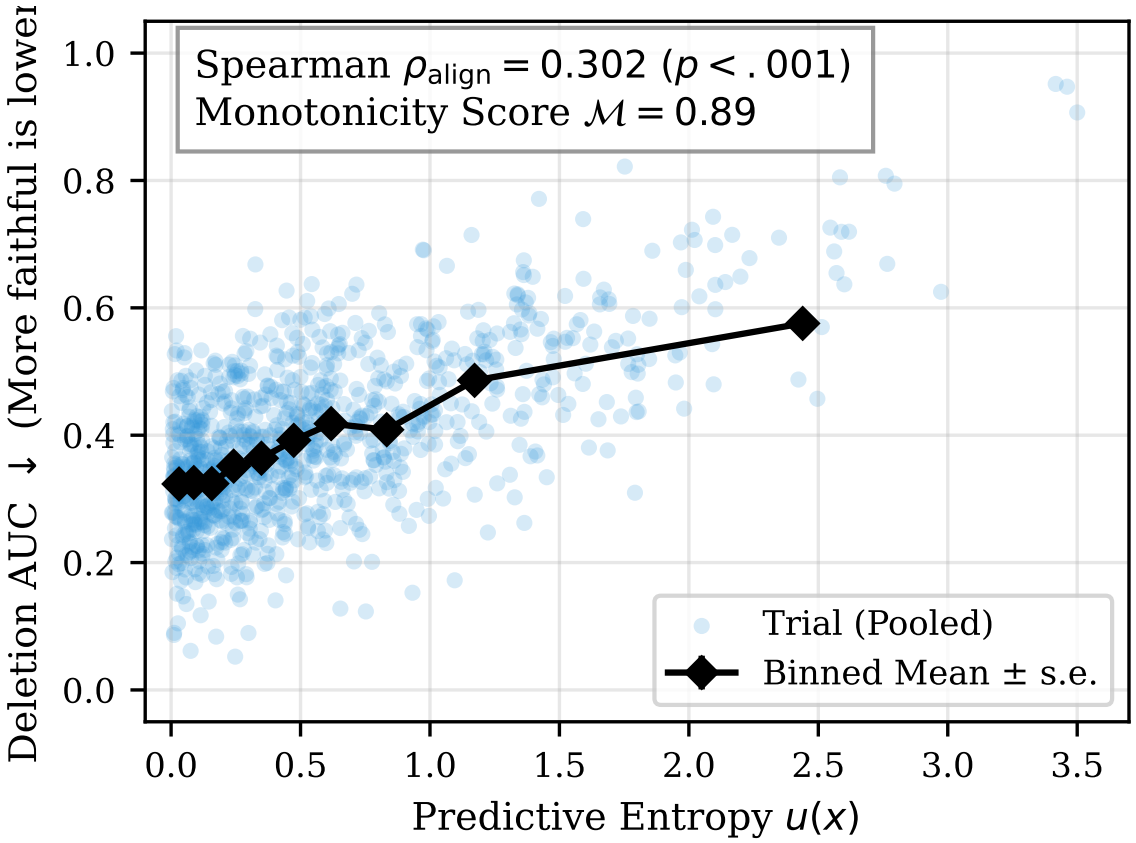


Figure 3: Uncertainty–faithfulness alignment for deep ensemble on BCI IV 2a. Scatter of predictive entropy vs Deletion AUC (lower AUC = more faithful) across all subjects and severity levels, with binned trend and Spearman ρ. Positive ρ confirms audit-compliant behaviour.

[t]

Table 2: Transparency failure by artefact family (deep ensemble, EEGNet, severity 4). Deltas: change from clean to severity-4. Alignment ρ for line noise is "–": near-uniform entropy at severity 4 makes rank correlation degenerate.

Artefact Family	$\mathrm{ECE}_{\Pi_E}(\text{sev4}) - \mathrm{ECE}_{\text{clean}}$	$\mathrm{NLL}_{\Pi_E}(\text{sev4}) - \mathrm{NLL}_{\text{clean}}$	Align $\rho\uparrow$	MisRate $\downarrow$
Channel Dropout	−0.045	−0.363	0.363	0.124
EOG Burst	+0.264	+10.079	0.190	0.104
EMG Muscle Noise	+0.050	+0.619	0.334	0.116
50/60 Hz Line Noise	+0.405	+13.903	–	0.016
Band Mismatch	+0.033	+0.465	**0.507**	0.147

elevated uncertainty signal enables the abstain-to-explain control to suppress this output.

6.4 Transparency Contract Summary

Table 3 formalises the transparency contract as an auditable artefact. Each system output is mapped to its validity condition, the diagnostic metric used to audit compliance, and the empirically observed failure mode.

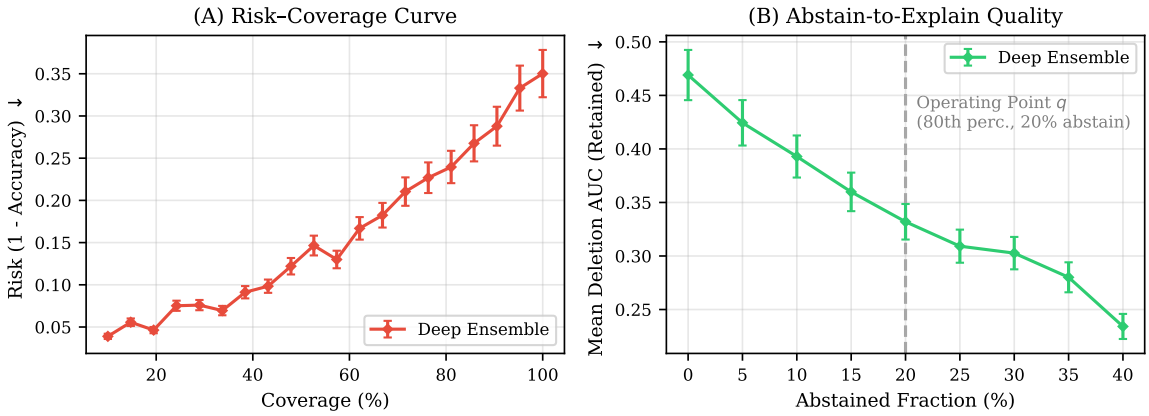


Figure 4: Abstain-to-explain as transparency control. **Left:** Risk–coverage curve showing monotonic risk reduction. **Right:** Mean Deletion AUC on retained set vs abstained fraction; marker indicates the 20% operating point. Deep ensemble, BCI IV 2a. The abstain policy enforces the transparency contract: when validity is in doubt, the explanation is withheld.

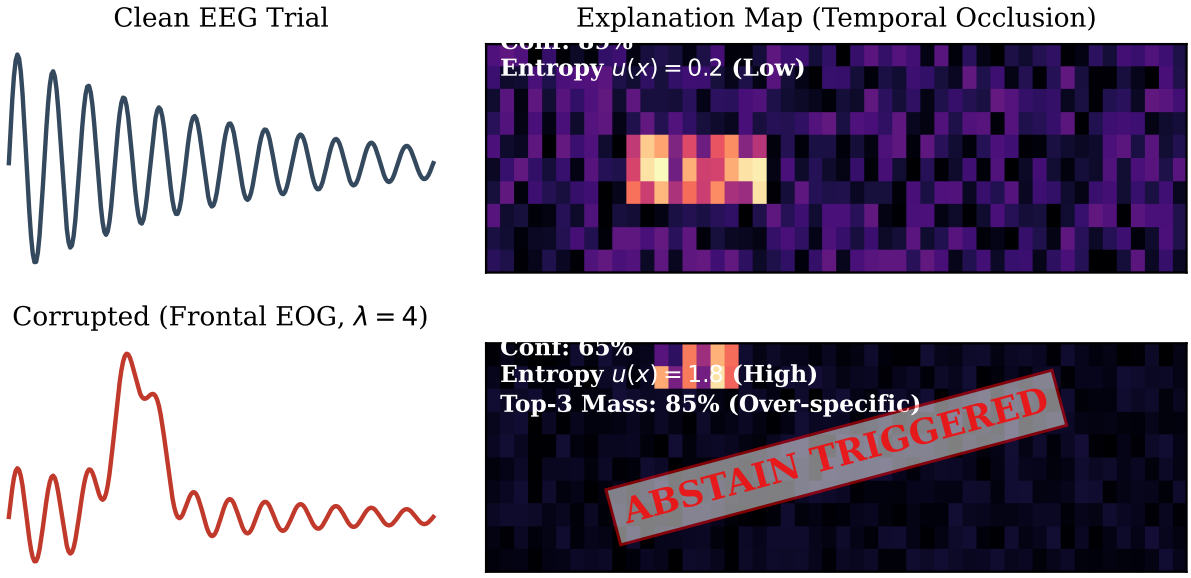


Figure 5: Qualitative transparency failure and mitigation. **Top:** Clean trial—low entropy, faithful attribution (low Deletion AUC). **Bottom:** EOG-corrupted trial—high entropy, sharp-but-wrong attribution targeting the artefact (high Deletion AUC). The deep ensemble's elevated uncertainty triggers the abstain-to-explain policy.

[t]

Table 3: Transparency contract summary. Each system output links to its validity condition, audit diagnostic, and observed failure mode. This synthesises the audit results into a compact accountability record.

Output Shown to User	Validity Condition	Audit Diagnostic	Observed Failure Mode
Class prediction	Calibrated under deployment conditions	$\mathrm{ECE}_{\mathrm{clean}}$, ECE_{Π_E}	Overconfident at severity >2; $\mathrm{ECE}_{\Pi_E} \gg \mathrm{ECE}_{\mathrm{clean}}$
Attribution map	Entropy $< q$ under Π_E at generation time	ρ_{align}, $\mathcal{M}$	Sharp-but-wrong maps at high entropy (Deletion AUC rises, attribution stays concentrated)
Explanation to user	Entropy $< q$ on current trial	Misalignment rate	High entropy + concentrated top-k mass (10% of trials)
Abstain flag	Entropy $\geq q$ triggers withholding	Abstain gain Δ	False suppression of some valid explanations

6.5 Protocol Overview

Figure 1 provides a schematic overview of the transparency audit pipeline.

7 Discussion

Transparency Beyond EEG. The audit protocol is not specific to EEG. The three diagnostics (ρ_{align}, $\mathcal{M}$, MisRate) and the abstain-to-explain control apply to any perturbation-based explanation method on any input modality. EEG provides a physically interpretable test bed that allows reasoning about failure modes at the signal level. Extending this audit to natural language processing—where perturbation-based explainers similarly mask tokens or sentences—and to multi-agent AI systems is a natural direction for future work.

From Audit to Enforcement. The current protocol is diagnostic: it measures whether the transparency contract is upheld and provides a reactive control. A stronger version would enforce the contract proactively by recalibrating the model on Π_E before explanations are generated. Decker et al. [3] demonstrated this for vision; adapting proactive enforcement to biosignal domains is future work.

Relation to the XTAI Agenda. This work addresses transparency in decision-making (via the transparency contract and diagnostics), explainable AI algorithms (via the audit of perturbation-based attribution), and domain-specific explainability (via the EEG instantiation). The abstain-to-explain mechanism is directly relevant to the XTAI focus on accountability: rather than producing explanations of unknown quality, the system explicitly discloses when it cannot produce a trustworthy explanation.

Limitations. Artefact perturbations are simulated; field recordings may exhibit different corruption profiles. The audit is instantiated on CNN architectures only; Transformer-based models on biosignal data warrant study. Full Π_E analysis is restricted to deep ensembles due to compute constraints. The abstain-to-explain threshold q may suppress valid explanations (false abstention); threshold optimisation for different deployment contexts remains open.

8 Conclusion

Explanations are safety-critical outputs that deserve explicit validity checks. We have presented a transparency audit protocol for perturbation-based attribution methods, grounded in a simple principle: an explanation is trustworthy only when the model is sufficiently certain under the perturbations used to generate it. The protocol introduces three measurable diagnostics—alignment ρ_{align}, monotonicity $\mathcal{M}$, and misalignment rate—and a transparency control, abstain-to-explain, that withholds attribution maps when the validity condition is not met.

Applied to BCI Competition IV 2a, the audit reveals that clean-set calibration fails to predict explanation-time reliability, that band mismatch produces the most deceptive transparency failures, and that the abstain-to-explain policy measurably improves the faithfulness of retained explanations. Deep ensembles produce the most audit-compliant explanations across both architectures tested.

We position this work as a step toward accountable explanation pipelines—not only for biosignal processing, but for the broader landscape of AI systems where trustworthy, auditable explanations are a foundational requirement.

Reproducibility, Validation, and Ethics Statement

All experiments use the public, open-source BCI Competition IV dataset 2a [19], presenting no proprietary or restricted patient data concerns. Random seeds are fixed at 42–46; results are stable across independent seeds. All hyperparameters are explicitly documented and preserved in per-run `config.yaml` snapshots. Run manifests (`manifest.json`) record code version, environment hash, and dataset integrity hash. All figures are generated programmatically from logged results files with no manual editing. Implemented in PyTorch. The complete codebase, validation harness, and step-by-step replication instructions will be released with the camera-ready version.

REFERENCES

[1] David Alvarez-Melis and Tommi S Jaakkola, 'On the robustness of interpretability methods', in *ICML Workshop on Human Interpretability*, (2018).

[2] Carles Balsells-Rodas, Fan Yang, Zhishen Huang, and Yan Gao, 'Explainable uncertainty attribution for sequential recommendation', in *Proceedings of the ACM SIGIR Conference*, pp. 2401–2405, (2024).

[3] Torsten Decker, Niklas Hartung, and Volker Tresp, 'Improving perturbation-based explanations by understanding the role of uncertainty calibration', in *European Conference on Artificial Intelligence (ECAI)*, (2024).

[4] Jamie Duell and Xiuyi Fan, 'Provably robust Bayesian counterfactual explanations under model changes', *arXiv preprint arXiv:2601.16659*, (2026).

[5] Siamac Fazli, Sven Dahne, Wojciech Samek, Felix Biessman, and Klaus-Robert Muller, 'Learning from more than one data source: data fusion techniques for sensorimotor rhythm-based brain–computer interfaces', *Proceedings of the IEEE*, **103**(6), 891–906, (2015).

[6] Yarin Gal and Zoubin Ghahramani, 'Dropout as a Bayesian approximation: Representing model uncertainty in deep learning', in *International Conference on Machine Learning (ICML)*, (2016).

[7] Yonatan Geifman and Ran El-Yaniv, 'Selective classification for deep neural networks', in *Advances in Neural Information Processing Systems (NeurIPS)*, (2017).

[8] Moritz Grosse-Wentrup and Bernhard Schölkopf, 'High gamma-power predicts performance in sensorimotor-rhythm brain–computer interfaces', *Journal of Neural Engineering*, **9**(4), (2012).

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, 'On calibration of modern neural networks', in *International Conference on Machine Learning (ICML)*, (2017).

[10] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller, 'Causability and explainability of artificial intelligence in medicine', *WIREs Data Mining and Knowledge Discovery*, **9**(4), (2019).

[11] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim, 'A benchmark for interpretability methods in deep neural networks', in *Advances in Neural Information Processing Systems (NeurIPS)*, (2019).

[12] Junqi Jiang et al., 'Robust counterfactual explanations in machine learning: A survey', in *International Joint Conference on Artificial Intelligence (IJCAI)*, (2024).

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, 'Simple and scalable predictive uncertainty estimation using deep ensembles', in *Advances in Neural Information Processing Systems (NeurIPS)*, (2017).

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

[15] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems (NeurIPS)*, (2017).

[16] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek, 'Can you trust your model's

uncertainty? Evaluating predictive uncertainty under dataset shift', in *Advances in Neural Information Processing Systems (NeurIPS)*, (2019).

[17] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball, 'Deep learning with convolutional neural networks for EEG decoding and visualization', *Human Brain Mapping*, **38**(11), 5391–5420, (2017).

[18] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju, 'Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods', in *AAAI/ACM Conference on AI, Ethics, and Society*, (2020).

[19] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot 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).