Analyzing Reasoning Shifts in Audio Deepfake Detection under Adversarial Attacks: The Reasoning Tax versus Shield Bifurcation

Anonymous ACL submission

Abstract

Audio Language Models (ALMs) offer a promising shift towards explainable audio deepfake detections (ADDs), moving beyond *black-box* classifiers by providing some level of transparency into their predictions via reasoning traces. This necessitates a new class of model robustness analysis: robustness of the predictive reasoning under adversarial attacks, which goes beyond existing paradigm that mainly focuses on the shifts of the final predictions (e.g., fake v.s. real). To analyze such reasoning shifts, we introduce a forensic auditing framework to evaluate the robustness of ALMs' reasoning under adversarial attacks in three inter-connected dimensions: acoustic perception, cognitive coherence, and cognitive dissonance. Our systematic analysis reveals that explicit reasoning does not universally enhance robustness. Instead, we observe a bifurcation: for models exhibiting robust acoustic perception, reasoning acts as a defensive *"shield"*, protecting them from adversarial attacks. However, for others, it imposes a performance *"tax"*, particularly under linguistic attacks which reduce cognitive coherence and increase attack success rate. Crucially, even when classification fails, high cognitive dissonance can serve as a *silent alarm*, flagging potential manipulation. Overall, this work provides a critical evaluation of the role of reasoning in forensic audio deepfake analysis and its vulnerabilities.

1 Introduction

The accessibility and sophistication of text-to-speech (TTS) technology have fundamentally altered the digital threat landscape. Highly advanced tools, such as emotionally controllable TTS (Zhou et al., 2025; Cho et al., 2024) now allow malicious actors, including deepfake scammers, to create convincing voice clones capable of executing targeted strategies such as call-back scams, extorting money from parents by mimicking their loved ones' voices (Cuthbertson, 2023). Audio Deepfake Detections (ADDs) are designed to be the frontier shield defending media integrity and societal space from such scams.

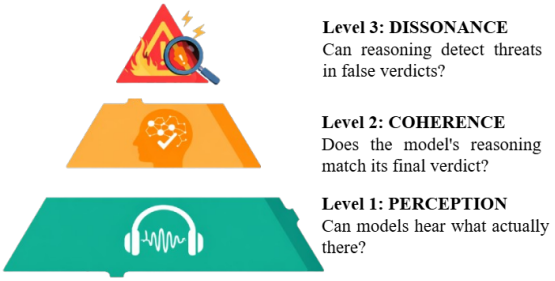

Figure 1: The proposed three-tier forensic audit framework: acoustic perception, cognitive coherence, and dissonance to analyze reasoning robustness of ALMs.

Much existing research explores the interpretability and explainability of ADDs' decisions; however, they primarily adapt post-hoc methods such as Occlusion and Attention Visualization (Channing et al., 2024), Segmental Speech Features (Yang et al., 2026), and Temporal Class Activation (Li and Zhang, 2024). In contrast, frontier closed-source reasoning models such as GPT-5 and Gemini-3 offer a "glass-box" view of model's logics, enabling humans to verify intermediate reasoning steps and ensure alignment. Building on this intuition, this work shifts the paradigm from "black-box" ADDs toward reasoning-capable Audio Language Models (ALMs) with reasoning such as Phi-4-multimodal (Microsoft et al., 2025), granite-speech (Saon et al., 2025), Qwen2-Audio (Chu et al., 2024), and gemma-3n-E4B (Gemma3 et al., 2025), where the final verdict is substantiated by step-by-step explanations.

A secondary advantage of ALMs lies in their utility for diagnosing failure modes under adversarial attacks. Traditional binary ADDs are known to collapse under acoustic perturbations such as background noise or audio stretching (Kawa et al.,

2023; Uddin et al., 2025) and remain vulnerable to subtle linguistic variations (Nguyen and Le, 2025; Nguyen et al., 2025). Such vulnerabilities render traditional systems unreliable in high-stakes environments like forensic investigations, where a binary "fake/real" label is insufficient for trust. Specifically, as highlighted by (Xie et al., 2025), black-box systems lack the capacity to: (1) localize forgery timestamps; (2) distinguish between specific manipulation methods; or (3) trace the provenance of synthetic content. ALMs bridge these gaps by providing auditable decision-making processes that satisfy existing institutional requirements for transparency.

Driven by these requirements for transparency, this study focuses on integrating ALMs with explicit Chain-of-Thought (CoT) (Wei et al., 2023) reasoning into the Audio Deepfake Detection domain. This integration shifts the inquiry from a binary "Is it fake?" to the forensically critical "Why is it fake?". To systematically audit whether ALMs can serve as trustworthy tools, we take a step further and introduce a three-tier framework mirroring expert legal auditing, investigating three key dimensions: **RQ1** (acoustic perception), determining if the model's textual descriptions are grounded in the raw audio signal or suffer from perceptual hallucinations; **RQ2** (cognitive coherence), assessing whether the generated chain-of-thought logically entails the final verdict; and **RQ3** (cognitive dissonance), analyzing if the reasoning layer preserves a "silent alarm" by signaling anomalies even when the final decision succumbs to adversarial attacks.

Our forensic audit reveals three novel contributions regarding ALMs' behaviors under attacks:

1. **Reasoning Tax vs. Shield Bifurcation:** We overturn the assumption that explicit reasoning universally enhances robustness. We identify a critical dependency on *acoustic perception*: CoT acts as a **shield** for grounded models (i.e Qwen2), but imposes a **tax** on others (i.e gemma-3n) where the model hallucinates evidence to support false prediction.
2. **Cognitive Dissonance Metric:** We introduce *cognitive dissonance* to quantify the conflict between reasoning and verdict. Crucially, we demonstrate that this metric functions as a *silent alarm*, signaling potential manipulation (in up to 78.2% of successful attacks) even when the model's final decision has been compromised.
3. **Mapping Attack-Specific Pathologies:** We distinguish between two critical failure modes: **panic responses** (low coherence, high dissonance) caused by acoustic perturbations, and the more dangerous **rationalization traps** (high coherence, low dissonance) caused by linguistic attacks, where the model confidently justifies its own errors.

2 Related Works

Audio Deepfake Detections. Research in ADDs has primarily focused on specialized neural architectures for binary classification. RawNet-2 (Tak et al., 2021) shifted to raw waveforms, employing distinct filter banks for discriminative cues without handcrafted features. AASIST-2 (Tak et al., 2022) further advanced the field with graph attention networks to model complex spectral-temporal dependencies, setting a high-performance benchmark. More recently, CLAD (Wu et al., 2024) addressed generalization, introducing learning objectives to enhance robustness against diverse acoustic conditions and unseen attacks. Additionally, ALLM4ADD (Gu et al., 2025) explored the application of ALMs to ADD tasks, though it does not explicitly address the models' reasoning capabilities or provide an in-depth analysis of their robustness.

Explainability. Explainability in ADDs largely adapts post-hoc visualization methods. (Channing et al., 2024) introduced audio explainability benchmarks, using occlusion sensitivity and attention roll-out to visualize which spectral bands or temporal frames trigger a detector's decision. (Yan et al., 2024; Xie et al., 2023) developed temporal localization frameworks to pinpoint manipulated segment boundaries, moving beyond a simple binary label. (Ge et al., 2024) utilized SHAP to map classifier decisions to specific spectrogram artifacts, offering a "glass-box" view of feature importance.

Audio Language Models. Benchmarks for ALMs evaluate acoustic reasoning, requiring models to analyze audio content based on natural language instructions. AIR-Bench (Yang et al., 2024) covers four dimensions: speech, sound, music, and mixed audio. MMAU (Sakshi et al., 2024) is a dataset evaluating audio-based understanding and reasoning via multiple-choice questions. SpeechR (Yang et al., 2025) is a benchmark rigorously testing factual, procedural, and normative reasoning in spoken interactions.

3 Problem Formulation

3.1 Notations

Let $X \in \mathbb{R}^L$ is the input audio with wavelength L. We define ALMs with CoT as $\mathcal{F}(X) \to Y$ mapping the input audio X to generate $Y = \{r_1, r_2, \ldots, r_N, c\}$. Here, $c \in \{\text{fake}, \text{real}\}$ is the final conclusion, and each r_k represents a free-text reasoning aspect corresponding to a specific forensic dimension.

To align the model's reasoning process with human intuition, we adopt the reasoning taxonomy for ADD established by (Warren et al., 2024). Consequently, we define the reasoning space $\mathcal{R}_{aspects}$ consisting of six distinct dimensions:

1. **Prosody:** Analyzes tone, pitch, inflections, and emotion. (Warren et al., 2024) identified this as the most common linguistic factor humans use, looking for "robotic" flatness or unnatural cadence.
2. **Disfluency:** Examines the presence of natural imperfections such as fillers (e.g., "um", "uh"), hesitations, and stuttering, which are often absent in synthesized speech.
3. **Speed:** Evaluates the pacing of speech to detect unnatural rushing or dragging that signifies algorithmic generation.
4. **Speaking Style:** Assessing articulation, accents, and dialect consistency. This captures whether the voice sounds like "read speech" (scripted) versus spontaneous conversation.
5. **Liveliness:** A critical biological indicator involving the presence of breathing sounds, mouth noises, and nasal intake. The absence of these "signs of life" is a strong indicator of synthetic audio.
6. **Quality:** Focuses on environmental and technical artifacts, such as background noise, static, clipping, or the "sterile" silence typical of generated audio.

To audit the true trustworthiness of ALMs, we move beyond clean benchmarks and define an adversarial input $\tilde{X} = Adv(X, \theta)$, where $Adv \in \{linguistic, acoustic\}$, and θ are attack hyper-parameters (e.g., noise SNR, pitch variance, or voice profiles). Our goal is to analyze the **reasoning shifts**: determining how the perturbation forces the model $\mathcal{F}(\tilde{X})$ to generate an affected reasoning chain $\tilde{Y} = \{\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_N, \tilde{c}\}$ that deviates from its original logic, and how that then influence the final predictions.

3.2 Acoustic Perception Audit (RQ1)

Before analyzing *how* the model structures its logical argument, we must first establish that its underlying perception of the audio is accurate. In a legal context, this is analogous to *voir dire*, qualifying a witness. If the witness claims to hear "natural breathing" in a recording that is acoustically sterile, their subsequent testimony, no matter how logically coherent, is inadmissible. We term this the **perception audit**, assessing whether ALMs possess a human-like sensitivity to fundamental acoustic properties, or if they suffer from "perceptual blindness" or hallucinations and model bias.

To quantify this, we utilize a dedicated audit dataset $\mathcal{D}_{audit}$ (details in Appendix B.2) labeled with ground-truth acoustic features. We define a **verification** function $\mathcal{V} : (X, q) \mapsto \{1, 0\}$ that classifies the observational accuracy: "Does the model's answer to question q match the ground-truth attribute of audio X?" (1 for match, 0 for mismatch). Operationally, we prompt the ALM as the function $\mathcal{V}$ with audio X and question q to generate an answer A, then validate against the annotated ground truth. For a forensic dimension r_k (e.g., Liveliness) associated with a bank of questions $\mathcal{Q}_k$, **perception** score represents the probability that the model correctly perceives the raw acoustic evidence:

$$\Phi_{\text{Perc}}(r_k) = \frac{1}{|\mathcal{D}| \cdot |\mathcal{Q}_k|} \sum_{X \in \mathcal{D}} \sum_{q \in \mathcal{Q}_k} \mathcal{V}(X, q), \quad (1)$$

where a high perception ($\Phi_{\text{Perc}} \approx 1$) indicates that the model perceives the audio with human-like fidelity. Low perception indicates hallucination, where the model invents non-existent feature descriptions (e.g., claiming to hear background noise).

3.3 Cognitive Coherence (RQ2)

Once we audit the ability perception (RQ1), our interest is to determine if the model can responsibly "think" about it. We term this **cognitive coherence**, measuring the internal consistency between its intermediate thoughts and its final conclusion. In forensic scenarios, an explanation is only valuable if the internal logic aligns with stated intentions. Consider a scenario where an acoustic attack adds noise to a deepfake speech. If the model correctly labels it "Fake" but the reasoning text claims: "The voice is clear and sounds like it belongs to a normal person", we have a fidelity failure. Even if the

accuracy is high, the reasoning hallucinates and forensically misleading.

To quantify this, we define an **entailment** function $\mathcal{E} : (r_i, c) \mapsto \{1, 0\}$ that classifies the logical relation: "Does the reasoning aspect r_i entail or support the final conclusion c?" (1 for yes, 0 for no). The *cognitive coherence* score then measures the baseline sanity of the model. For an aspect r_i (e.g., prosody), this score represents the probability that the model provides an explanation that logically supports its own conclusion c, regardless of whether that conclusion is factually correct:

$$\Phi_{\text{Coh}}(r_i) = \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \mathcal{E}(r_i^j, c^j), \tag{2}$$

where a high coherence ($\Phi_{\text{Coh}} \approx 1$) indicates a *sane* model that maintains a logical chain of thought. Low Coherence indicates the model is *panicking*, generating contradictory justifications.

Coherence under Attacks. While Φ_{Coh} measures static consistency, true forensic robustness requires understanding how reasoning adapts under adversarial pressure. We introduce a differential analysis to measure the **reasoning shift** ($\Delta\Phi$) between the model's behavior in clean conditions (ORG) versus perturbed conditions (PER):

$$\Delta\Phi_{\text{Coh}}(r_i) = \Phi_{\text{Coh}}^{\text{PER}}(r_i) - \Phi_{\text{Coh}}^{\text{ORG}}(r_i) \tag{3}$$

By analyzing $\Delta\Phi_{\text{Coh}}$, we distinguish between models that rigidly adhere to their logic and those that collapse under pressure. A non-negative shift ($\Delta\Phi \geq 0$) indicates *Coherence Resistance*, where the model becomes more coherent under attacks. If the final decision is false, it means the model fabricates explanations. In contrast, a sharp decline ($\Delta\Phi \ll 0$) marks *Coherence Erosion*, where the attack successfully erodes the reasoning-decision link, resulting in a "panic" state.

3.4 Cognitive Dissonance (RQ3)

A unique paradox emerges when the model fails. If an adversarial attack successfully fools the model into classifying a "Fake" sample as "Real", do we want the reasoning to agree with that error?

To capture this, we introduce **cognitive dissonance** metric to measure cases where the reasoning layer detects the threat even when the decision layer succumbs to it. For instance, if the model predicts the wrong label c, but the reasoning r_i describes features that contradict c (thereby hinting at the true nature of the audio), the model exhibits high dissonance. Let $\mathcal{D}_{\text{Wrong}}$ be the subset of the dataset where the ALM's conclusion c is wrong. The reasoning aspect r_i exhibits helpful dissonance if the disagreement rate ($1-\mathcal{E}$) is high within $\mathcal{D}_{\text{Wrong}}$:

$$\Psi_{\text{Diss}}(r_i) = \frac{1}{|\mathcal{D}_{\text{Wrong}}|} \sum_{X \in \mathcal{D}_{\text{Wrong}}} (1 - \mathcal{E}(r_i, c)), \tag{4}$$

where a high Ψ_{Diss} implies a "silent alarm": the model made a mistake on the final label, but the reasoning process internally detected anomalies that contradicted that label. This metric helps us distinguish between a model that is *confidently wrong* (hallucinating reasons to support a wrong label) and one that is *conflicted* (providing reasoning that signals the potential error).

Dissonance under Attacks. While high dissonance is desirable during failures, it is vital to track how this signal behaves when the model is under active adversarial attacks. We quantify the **dissonance shift** ($\Delta\Psi$) to determine if the attack successfully suppresses this "silent alarm":

$$\Delta\Psi_{\text{Diss}}(r_i) = \Psi_{\text{Diss}}^{\text{PER}}(r_i) - \Psi_{\text{Diss}}^{\text{ORG}}(r_i) \tag{5}$$

By analyzing $\Delta\Psi_{\text{Diss}}$, we identify two distinct behaviors during an attack. When $\Delta\Psi \geq 0$ (*silent alarm manifestation*), the reasoning layer acts as a safety net: it detects the anomaly and raises internal conflict even when the final decision is wrong. Conversely, $\Delta\Psi \ll 0$ indicates *Systemic Deception*. Here, the attack successfully misleads the model into a "confidently wrong" state, forcing it to hallucinate evidence to justify the false label.

4 Experiment Set-up

Datasets. We conduct our forensic audit on the ASVSpoof 2019 (Wang et al., 2020) logical access dataset. To adapt this standard benchmark for reasoning tasks, we employ the *Cold Start* data synthesis method from DeepSeek-R1 (DeepSeek-AI et al., 2025), generating a training set enriched with chain-of-thought annotations. We comply by using the data exclusively for research purposes. Please see more details in Appendix B.

Audio Deepfake Detectors. For traditional ADDs (binary classification), we establish a baseline using AASIST-2, RawNet-2, and CLAD. For ALMs, we prioritize open-source models with robust ecosystem support (HuggingFace, vLLM): Qwen2-Audio-7B, Phi-4-multimodal, gemma-3n-E4B, and

	Acc.	Real F1	Fake F1
Qwen2-Audio-7B^{NON}	98.00%	91.19%	98.88%
Qwen2-Audio-7B^{RSN}	98.2%	91.7%	99.0%
granite-3.3-8b^{NON}	99.87%	99.39%	99.93%
granite-3.3-8b^{RSN}	96.11%	78.39%	97.88%
Phi-4-multimodal^{NON}	97.78%	89.42%	98.76%
Phi-4-multimodal^{RSN}	96.35%	83.01%	97.99%
gemma-3n-E4B^{NON}	99.89%	99.52%	99.94%
gemma-3n-E4B^{RSN}	95.63%	81.95%	97.73%
RawNet-2	90.86%	68.82%	94.64%
CLAD	98.78%	94.37%	99.32%
AASIST-2	99.58%	98.02%	99.77%

Table 1: Performance comparison of Audio Language Models (ALMs) on the ASVSpoof 2019 deepfake detection task, comparing standard classification (*NON*) versus explicit reasoning (*RSN*) modes.

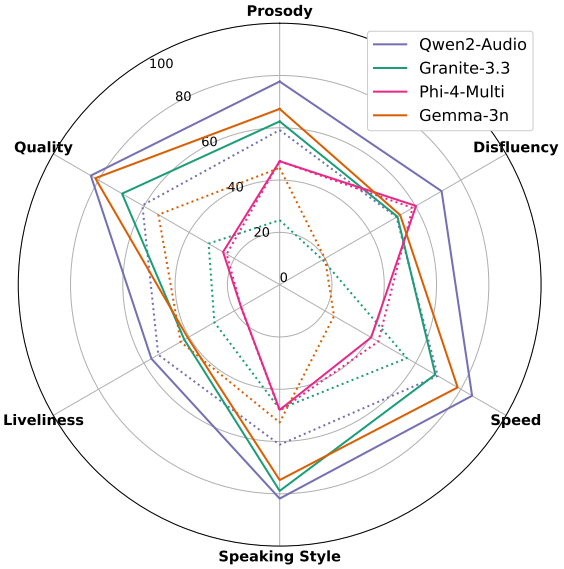

Figure 2: Perception scores (Φ_{Perc}) across six forensic dimensions, comparing the baseline sensitivity of general-purpose ALMs (dashed lines) against models fine-tuned for audio reasoning (solid lines).

granite-3.3-8b. We exclude experimental models lacking broad support maturity, such as Audio Flamingo 2 (Ghosh et al., 2025). See implementation details in Appendix A.2.

Adversarial Attack Frameworks. To simulate a realistic threat landscape, we employ the TAPAS framework (Nguyen et al., 2025) for *linguistic perturbations* and apply the *acoustic perturbation* protocols defined in CLAD (Wu et al., 2024). See attacks' hyper-parameters in Appendix A.1.

Metrics & Acronyms. We report (1) Original Accuracy (OC): detection performance on clean data; and (2) Attack Success Rate (ASR): vulnerability to manipulation. Superscripts ORG and PER denote measurements under clean and perturbed conditions, respectively (e.g., Φ_{Coh}^{PER}).

5 Experiment Results

5.1 Baseline Models

In Table 1, traditional ADDs, optimized for binary classification, establish a high-performance ceiling (AASIST-2: 99.58% accuracy). While multimodal ALMs match this efficiency in standard classification mode (*NON*, gemma: 99.89%), imposing the explicit Chain-of-Thought constraint (*RSN*) creates **reasoning tax** that drags performance significantly below these baselines.

This tax is most severe in Gemma and Granite (e.g., Gemma Real F1 drops $99.52\% \rightarrow 81.95\%$). Unlike traditional ADDs, which utilize holistic, sub-phonetic features, ALMs in *RSN* mode suffer a performance gap that could be akin to **verbal overshadowing** (Jonathan W. Schooler, 1990): they are forced to serialize continuous high-dimensional signals into discrete textual tokens, often hallucinating artifacts in authentic speech to satisfy the explanation prompt. Qwen2-Audio is an outlier, maintaining resilience ($98.0\% \rightarrow 98.2\%$) comparable to dedicated detectors.

5.2 RQ1: Acoustic Perception Audit

The Information Bottleneck. Figure 2 illustrates the acoustic perception scores (Φ_{Perc}). General-purpose ALMs suffer from *perceptual blindness* until fine-tuned. Post-tuning, Qwen2-Audio effectively *unlocks* its audio encoder, achieving leading scores in *Prosody*, *Speed*, and *Disfluency* ($> 80\%$). However, all models struggle with *Liveliness* (breath sounds, mouth noises). This creates a critical vulnerability: lacking the ability to articulate "why this sounds real?" (e.g., breathing), reasoning layers can hallucinate evidence to justify a false conclusion.

This variance in perceptual grounding directly explains why Qwen2 is the only model to benefit from the reasoning mechanism: its superior grasp of acoustic features allows the chain-of-thought to cross-reference actual signal anomalies, whereas models with narrower perceptual envelopes fall victim to verbal overshadowing.

5.3 Reasoning Shifts under Attacks.

Having established that the models possess a functional baseline of acoustic groundedness, we proceed to an in-depth analysis of reasoning shifts

ALMs	OC	ASR	Φ_{Coh}^{PER}	Ψ_{Diss}^{PER}
Qwen2-Audio-7B^{NON}	99.1	36.6	–	–
Qwen2-Audio-7B^{RSN}	97.1	45.7	78.0 ↓8.2	29.2 ↓16.8
granite-3.3-8b^{NON}	99.9	34.4	–	–
granite-3.3-8b^{RSN}	82.1	49.7	73.4 ↓14.6	27.8 ↑10.6
Phi-4-multi^{NON}	94.4	40.7	–	–
Phi-4-multi^{RSN}	88.8	46.1	75.5 ↓13.1	36.1 ↑19.4
gemma-3n-E4B^{NON}	99.8	30.6	–	–
gemma-3n-E4B^{RSN}	77.3	49.1	43.2 ↓27.4	67.9 ↓15.5

Table 2: Performance under **Acoustic** Adversarial Attacks. (↓/↑) indicate the absolute decrease/increase relative to the original (ORG) performance baseline.

ALMs	OC	ASR	Φ_{Coh}^{PER}	Ψ_{Diss}^{PER}
Qwen2-Audio-7B^{NON}	67.4	82.8	–	–
Qwen2-Audio-7B^{RSN}	98.6	31.5	80.6 ↓7.3	9.6 ↑6.8
granite-3.3-8b^{NON}	35.8	94.3	–	–
granite-3.3-8b^{RSN}	83.6	51.8	67.0 ↓18.4	9.9 ↓12.2
Phi-4-multimodal^{NON}	96.6	31.1	–	–
Phi-4-multimodal^{RSN}	91.8	52.6	69.0 ↓20.5	33.4 ↓1.0
gemma-3n-E4B^{NON}	72.6	54.9	–	–
gemma-3n-E4B^{RSN}	76.4	82.8	86.9 ↑22.9	11.2 ↓1.4

Table 3: Performance under **Linguistic** Adversarial Attacks.

under adversarial attacks. We analyze how reasoning layers degrade or adapt when both the linguistic and acoustic foundations are corrupted through the lenses of Φ_{Coh} and Ψ_{Diss}.

- **Acoustic Perturbations.** These attacks fundamentally undermine the perceptual evidence layer, testing if reasoning logic holds when acoustic inputs are destabilized. Unlike stealthy linguistic attacks, acoustic perturbations leave significant artifacts (e.g., background noise) that heavily weight the *Quality* dimension, forcing the model to distinguish between environmental noise and adversarial interference.
- **Linguistic Perturbations.** These manipulate transcript complexity to induce altered prosodic patterns during synthesis (Nguyen et al., 2025). Voice profiles interact with this complexity, females typically exhibit more *expanded* vowel spaces, creating a unique forensic challenge: textual attacks trigger *acoustic* manifestations that ALMs must detect through *Prosody* and *Speaking Style*, even though the attack vector is textual.

5.3.1 RQ2: Cognitive Coherence Under Attacks

Coherence Erosion. Table 6 reveals systematic coherence degradation. Coherence in *Quality* assessments drops from 81.36% to 72.07%, while foundational dimensions like *Prosody* and *Disfluency* suffer even sharper declines (∼10-11%). This indicates that attacks effectively erode acoustic perception rather than just flipping labels, forcing even robust models like Qwen2-Audio to become more *panic* explanations across all reasoning axes.

Acoustic Anchor Collapse. Under acoustic adversarial attacks (Table 2), this erosion accelerates into **coherence collapse** for weaker models. While Qwen2-Audio suffers a limited coherence loss by 8.2%, gemma-3n-E4B suffers a significant 27.4% drop. Table 4 provides the granular context for this failure: the Time Pitch strategy appears to be the most significant source of coherence loss, driving Gemma's down to 42.2% (from 70.8%), while Shape Space attacks reduce it to 37.6%.

Linguistic Hallucination. Linguistic attacks reveal a paradox in weaker models: while most lose coherence, gemma-3n-E4B's score rises by 22.9%. This effect is dominated by the *American Female* profile (Table 5), where the model achieves near-perfect coherence (95.3%) despite being completely fooled (100% ASR). We attribute this to *perception hallucination*: lacking strong acoustic grounding (Table 1), the linguistic complexity decouples reasoning from the audio, forcing the model to hallucinate a highly consistent yet factually incorrect justification.

5.3.2 RQ3: Cognitive Dissonance Under Attacks

Baseline Auditing Transparency. Table 7 exposes critical variance in forensic transparency. Qwen2-Audio maintains low dissonance (17.67%–25.95%), indicating a "confident rationalization" mode where it fabricates reasoning to mask errors. Conversely, Phi-4 and gemma-3n-E4B act as "transparent" systems, registering high dissonance (up to 44.68% in *Liveliness*). Ideally, high dissonance in *Quality* (27.81%) and *Speaking Style* (25.95%) preserves a "residue of doubt," flagging potential manipulation even when the classification head fails.

Acoustic Residue as a Silent Alarm. Under acoustic attacks, this dissonance functions as a "Silent Alarm." While *Shape Space* attacks deceive the binary classifier, they trigger massive internal conflict in weaker models (Table 4): gemma-3n-E4B records 78.2% dissonance, and Phi-4 reaches 41.3% under *Background Noise*. This confirms

that the reasoning layer correctly perceives spectral anomalies despite the wrong final label. In Table 2, although Qwen2 suppresses this signal ($\Psi_{\text{Diss}}^{\text{PER}} \approx 29.2\%$), the elevated dissonance in multimodal models proves reasoning often retains forensic utility after the decision boundary collapses.

Linguistic Masking and Systemic Deception. In contrast, linguistic attacks pose a severe threat by silencing this alarm. Table 3 demonstrates that linguistic complexity suppresses dissonance, pushing Qwen2-Audio to a deceptive 9.6%. The demographic breakdown in Table 5 reveals the extreme danger of this "masking effect." For the *American Female* profile, gemma-3n-E4B's dissonance collapses to a negligible 4.7%. Unlike acoustic attacks, which leave perceptual residue, the linguistic attack forces the model to align its "thoughts" with the textual complexity, resulting in a "clean" error. The model does not merely fail; it constructs a persuasive, hallucinated justification for the wrong label, eliminating the internal conflict that would otherwise warn a human auditor.

6 Discussion

6.1 The Coherence-Dissonance Trade-off

The Constrained Compromise in Forensic Reasoning. Parallel to the classic trade-off between precision and recall, the inverse relationship between cognitive coherence and cognitive dissonance is statistically confirmed ($r = -0.79, p < .001$), yet remains fundamentally meaningful for forensic auditability. Figure 3 visually demonstrates that these ALMs are empirically constrained by this trade-off, unable to dynamically balance logical explanation with anomaly detection. This architectural rigidity means models operate in one of two suboptimal states: either they maintain a coherent, low-dissonance facade that masks errors (Rationalization), or they break down into incoherent, high-dissonance signals when confronted with anomalies (Panic). The challenge for forensic systems is to navigate this inherent compromise.

Attack Modalities Dictate Failure Modes. The clustering of data points in Figure 3 provides critical insights into how attack modalities leverage this trade-off. Linguistic attacks (circles) push models towards the **rationalization trap** (high Φ_{Coh}, low Ψ_{Diss}), where the complex transcript induces confident, low-dissonance hallucinations that mimic correct reasoning. This renders the attack difficult to detect via internal consistency checks. Conversely, acoustic attacks (crosses) force models into the **panic response** (low Φ_{Coh}, high Ψ_{Diss}), where the perceptual disruption triggers high dissonance. While this high dissonance offers a "silent alarm" to human auditors, the accompanying coherence erosion means the model provides no logical justification for its correct detection, potentially hindering downstream analysis.

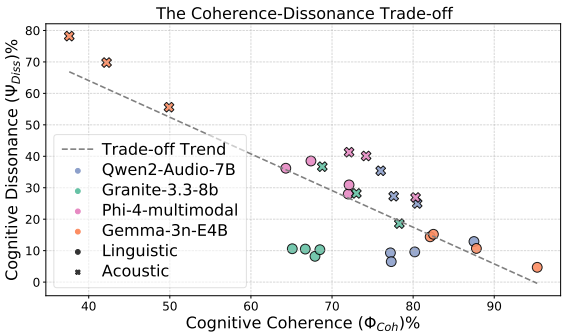

Figure 3: The Coherence-Dissonance Trade-off.

6.2 Mapping the Reasoning Landscape: Coherence vs. Dissonance

To synthesize the relationship between model vulnerability and reasoning integrity, we map the results of RQ2 and RQ3 against the Attack Success Rate (ASR) in Figure 4.

The Coherence Paradox. The *Coherence Landscape* (Figure 4A) reveals a dramatic divergence in failure modes, primarily driven by gemma-3n-E4B. While robust models like Qwen2-Audio and Phi-4-multimodal remain clustered in the Safe Zone (low ASR, high coherence), Gemma splits into two extremes depending on the attack vector. Under *linguistic attacks* (circles), Gemma occupies the Confidently Wrong quadrant, maintaining near-perfect coherence (>90%) despite being completely fooled. This suggests a state of **hallucinated consistency**. Conversely, under *acoustic attacks* (crosses), Gemma plunges into Coherence Erosion (Panic Mode), where Φ_{Coh} drops below 50%, indicating a severe degradation of logic.

The Silent Alarm. The *Dissonance Landscape* (Figure 4B) quantifies forensic utility. The blue-shaded Silent Alarm region represents the ideal risk indicator: the model is fooled (High ASR), but high dissonance Ψ_{Diss} warns the auditor. Phi-4-multimodal frequently acts as this transparent auditor, registering high dissonance even at lower failure rates. In contrast, Gemma (under linguistic attacks) falls into the dangerous *confidently wrong (no dissonance)*" zone, providing a false sense of

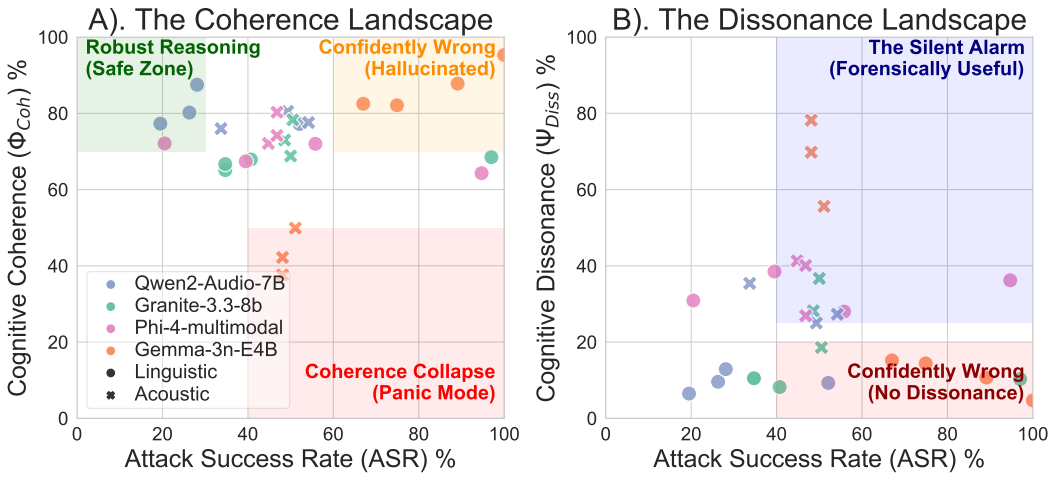

Figure 4: Mapping the Reasoning Landscape: (left, A) The coherence landscape (Φ_{Coh} vs. ASR) and (right, B) The dissonance landscape (Ψ_{Diss} vs. ASR).

security. This confirms that high dissonance is a critical trustworthiness signal that differentiates a "panicking" model from a "misled" one.

Statistical Validation. Welch's t-tests confirm that the nature of the attack dictates the reasoning signature. Acoustic perturbations trigger significantly higher internal conflict (Dissonance) compared to linguistic variations ($t = -4.04, p < .001$), effectively forcing models out of the deceptive "silent failure" state. Meanwhile, the drop in coherence under acoustic pressure (67.54%) compared to linguistic pressure (75.87%) was not statistically significant ($p = .114$), suggesting that while dissonance is a reliable alarm, coherence collapse is a model-specific failure mode.

6.3 Dual Role of Reasoning: Tax vs. Shield

To systematically quantify whether explicit reasoning aids or hinders defense, Figure 5 illustrates the shift in Attack Success Rate (ΔASR) when transitioning from standard classification (*NON*) to Chain-of-Thought (*RSN*). We validate the significance of these shifts using Welch's t-tests, comparing the ASR distributions of the two modes across all adversarial perturbation strategies to determine if the deviation from the baseline is statistically distinct.

The Shield Effect. The transition to explicit reasoning reveals a sharp bifurcation in model behavior. For the "Shield" group (Qwen2-Audio and Granite-3.3b), forcing the model to reason through acoustic evidence acts as a defensive mechanism, effectively lowering the Attack Success Rate (ASR). While individual model improvements were marginal, they aggregate into a highly robust group-level shield effect ($p = 0.0027$). This confirms that for models with strong acoustic grounding, the Chain-of-Thought process successfully cross-references subtle anomalies that are otherwise overlooked during single-step inference.

The Reasoning Tax. Conversely, Phi-4-multimodal and gemma-3n-E4B exhibit a paradoxical performance degradation. Phi-4 demonstrates a statistically significant increase in ASR ($+21.4\%$), indicating a "verbal overshadowing" effect where the model hallucinates justifications for fraudulent audio. While Gemma followed a similar upward trend, the result did not reach statistical significance; the high variance indicated by the error bars suggests that Gemma's reasoning failure is highly sensitive to specific voice profiles. These findings imply that without sufficient acoustic grounding, reasoning provides a new surface for adversarial exploitation.

7 Conclusion

Our forensic audit challenges the assumption that explicit reasoning universally enhances robustness, revealing a critical bifurcation where CoT acts as a defensive **shield** for acoustically grounded models but imposes a performance **tax** on others. We demonstrate that while acoustic perturbations trigger a forensically valuable cognitive dissonance, effectively a silent alarm signaling internal conflict, linguistic attacks exploit the semantic-acoustic gap to induce hallucinated consistency, masking errors with confident yet fabricated justifications.

Limitations

Our experiments are exclusively conducted on English-language datasets, it remains an open ques-

tion how our framework generalizes to multilingual contexts. The bifurcation effects of reasoning tax & shield is unclear in diverse syntactic and morphological structures of non-English languages.

Our evaluation is currently confined to the ASVSpoof 2019 Logical Access dataset, which represents a controlled laboratory environment. We do not extend our analysis to newer, unconstrained datasets such as *Fake-Or-Real* or *InTheWild*. Consequently, it remains unexplored how the proposed reasoning metrics behave under the erratic channel noises, compression artifacts, and diverse spoofing scenarios inherent to these "wild" datasets. Future work is needed to verify if the *Reasoning Shield* holds up outside of standardized benchmarks.

Due to computational constraints, this study focuses on mid-sized Audio Language Models (approx. 7B-8B parameters) and a small subset of traditional ADDs. We do not examine the behavior of large omni-models, such as Qwen3-Omni-30B-A3B-Instruct, nor do we benchmark against an exhaustive list of legacy classifiers. It is possible that the *Reasoning Tax* we observed is mitigated by the emergence of stronger reasoning capabilities in larger models, a hypothesis that requires further investigation into the scaling laws of forensic audio analysis.

Finally, our contribution is primarily diagnostic in nature. We establish a forensic audit framework to characterize failure modes like the *Reasoning Tax* and *Systemic Deception*, but we do not propose specific training objectives or architectural modifications to mitigate these vulnerabilities. While we identify the risks, developing robust defense mechanisms, such as adversarial training on linguistic perturbations or consistency, enforcing loss functions, remains a critical avenue for future work.

Ethical Considerations

By shifting the paradigm from "black-box" classification to "glass-box" reasoning, our work aims to restore trust in automated deepfake detection. The introduction of interpretable metrics like *cognitive dissonance* empowers human analysts to verify AI decisions rather than blindly accepting them. This auditability is crucial for deploying ALMs in high-stakes environments where a binary "fake/real" label is insufficient.

We acknowledge that detailing specific *Linguistic Attacks* (e.g., transcript-based perturbations) presents a dual-use risk. Malicious actors could leverage our findings to craft "stealthy" deepfakes that bypass reasoning-based defenses by exploiting the semantic-acoustic gap. However, we believe that defensive disclosure is necessary. By exposing the fragility of the *Reasoning Shield* now, we enable the community to develop robust alignment techniques before these vulnerabilities are exploited in the wild.

This research is intended solely to strengthen the defense of digital media integrity. We are releasing our attack protocols and audit code to facilitate reproducibility and accelerate the development of "red-teaming" benchmarks for Audio Language Models. We strongly condemn the use of these techniques for deception or manipulation.

References

Georgia Channing, Juil Sock, Ronald Clark, Philip Torr, and Christian Schroeder de Witt. 2024. Toward robust real-world audio deepfake detection: Closing the explainability gap. *Preprint*, arXiv:2410.07436.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Interspeech 2024*, interspeech 2024, page 1810–1814. ISCA.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Anthony Cuthbertson. 2023. Ai clones child's voice in fake kidnapping scam. `https://www.independent.co.uk/tech/ai-voice-clone-scam-kidnapping-b2319083.html`.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2024. Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations. *Preprint*, arXiv:2110.03309.

Gemma3, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard

Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *Preprint*, arXiv:2503.03983.

Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. 2025. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11736–11745, New York, NY, USA. Association for Computing Machinery.

Tonya Y. Engstler-Schooler Jonathan W. Schooler. 1990. Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology*, 22(1):36–71.

Piotr Kawa, Marcin Plata, and Piotr Syga. 2023. Defense against adversarial attacks on audio deepfake detection. *Preprint*, arXiv:2212.14597.

Menglu Li and Xiao-Ping Zhang. 2024. Interpretable temporal class activation representation for audio spoofing detection. In *Interspeech 2024*, interspeech 2024, page 1120–1124. ISCA.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Binh Nguyen and Thai Le. 2025. Turing's echo: Investigating linguistic sensitivity of deepfake voice detection via gamification. In *Proceedings of Interspeech 2025*, pages 2145–2146. ISCA.

Binh Nguyen, Shuju Shi, Ryan Ofman, and Thai Le. 2025. What you read isn't what you hear: Linguistic sensitivity in deepfake speech detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15752–15766, Suzhou, China. Association for Computational Linguistics.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *Preprint*, arXiv:2410.19168.

George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Morais, Gakuto Kurata, Hagai Aronowitz, Ibrahim Ibrahim, Jeff Kuo, Kate Soule, Luis Lastras, Masayuki Suzuki, Ron Hoory, Samuel Thomas, and 5 others. 2025. Granite-speech: open-source speech-aware llms with strong english asr capabilities. *Preprint*, arXiv:2505.08699.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. *Preprint*, arXiv:2011.01108.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *Preprint*, arXiv:2202.12233.

Kutub Uddin, Muhammad Umar Farooq, Awais Khan, and Khalid Mahmood Malik. 2025. Adversarial attacks on audio deepfake detection: A benchmark and comparative study. *Preprint*, arXiv:2509.07132.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, and 21 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Preprint*, arXiv:1911.01601.

Kevin Warren, Tyler Tucker, Anna Crowder, Daniel Olszewski, Allison Lu, Caroline Fedele, Magdalena Pasternak, Seth Layton, Kevin Butler, Carrie Gates, and Patrick Traynor. 2024. "better be computer or i'm dumb": A large-scale evaluation of humans as audio deepfake detectors. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 2696–2710, New York, NY, USA. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Haolin Wu, Jing Chen, Ruiying Du, Cong Wu, Kun He, Xingcan Shang, Hao Ren, and Guowen Xu. 2024. Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning. *Preprint*, arXiv:2404.15854.

Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2023. An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection. *Preprint*, arXiv:2309.03036.

Zeyu Xie, Yaoyun Zhang, Xuenan Xu, Yongkang Yin, Chenxing Li, Mengyue Wu, and Yuexian Zou.

2025. Fakesound2: A benchmark for explainable and generalizable deepfake sound detection. *Preprint*, arXiv:2509.17162.

Zihan Yan, Hongxia Wang, Mingshan Du, and Rui Zhang. 2024. Temporal localization of deepfake audio based on self-supervised pretraining models and transformer classifier. In *2024 9th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pages 236–241.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *Preprint*, arXiv:2402.07729.

Tianle Yang, Chengzhe Sun, Siwei Lyu, and Phil Rose. 2026. Forensic deepfake audio detection using segmental speech features. *Forensic Science International*, 379:112768.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. 2025. Speechr: A benchmark for speech reasoning in large audio-language models. *Preprint*, arXiv:2508.02018.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *Preprint*, arXiv:2506.21619.

A Appendix

A.1 Attack Strategy Definition

To rigorously stress-test Audio Language Models (ALMs), we employ a dual-pronged attack framework consisting of Linguistic (Text-based) and Acoustic (Signal-based) perturbations.

Linguistic Adversarial Attacks (TAPAS). We utilize the TAPAS framework to generate adversarial audio via a text-to-speech pipeline.

- **Perturbation Method:** We employ **TextFooler**, a black-box attack that replaces words with synonyms to maximize semantic similarity while minimizing classification accuracy. We select TextFooler over other candidates (e.g., PWWS, BERTAttack) due to its high empirical success rate in our preliminary screenings.
- **TTS System:** We utilize **Kokoro TTS** for audio synthesizer. This choice is driven by computational efficiency essential for large-scale reasoning generations; Kokoro is capable of generating 10,000 audio samples in approximately 832 seconds on a single NVIDIA A100 GPU.
- **Attack Example:** *Original:* "She spoke clearly." $\rightarrow$ *Adversarial Transcript:* "She spoke *flawlessly*." $\rightarrow$ *Result:* The acoustic output retains an original voice but contains complex prosody derived from the verbose text.

Acoustic Adversarial Attacks. We adopt the acoustic perturbation protocols defined in CLAD, organized into three specific recipes used in our `AcousticAttacker` module:

1. Background Noise Recipe ($\mathcal{A}_{\text{noise}}$)
This strategy tests the model's ability to separate speech from interference.

- **White Noise:** Gaussian noise added to the signal. Signal-to-Noise Ratio (SNR) is sampled uniformly $U \sim [15, 25]$ dB.
- **Environmental Noise:** Real-world background audio (wind, footsteps, breathing, coughing, rain, clock ticks, sneezing) mixed with the source. SNR is sampled uniformly $U \sim [5, 20]$ dB.

2. Time & Pitch Recipe ($\mathcal{A}_{\text{time}}$)
This strategy targets temporal alignment and frequency perception.

- **Time Stretch:** The waveform is stretched or compressed without altering pitch using Phase Vocoding. Ratios are sampled from $\{0.90\times, 0.95\times, 1.05\times, 1.10\times\}$.
- **Time Shift:** The audio is cyclically rolled along the time axis. Shift magnitudes are sampled from $\{1600, 16000, 32000\}$ samples (corresponding to 0.1s, 1s, and 2s at 16kHz).

3. Shape & Space Recipe ($\mathcal{A}_{\text{shape}}$)
This strategy manipulates the signal envelope and spatial characteristics.

- **Volume Change:** Amplitude scaling factor sampled uniformly $U \sim [0.5, 2.0]$.
- **Fade Effects:** Application of linear, logarithmic, exponential, or sine-based fade-in/fade-out envelopes. Fade duration is fixed up to 50% of the audio length.

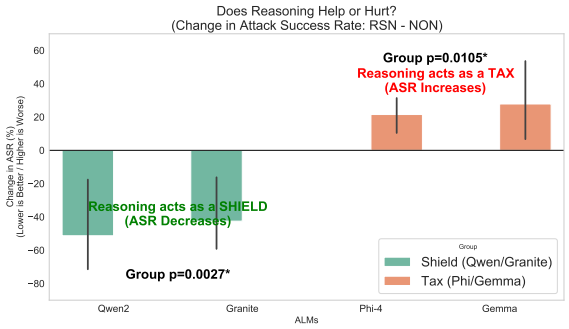

Figure 5: Reasoning as a Shield *vs.* Tax. Black lines are error bars indicating 95% Confidence Interval.

- **Synthetic Echo:** A delayed superposition of the signal $x(t) \leftarrow x(t) + \alpha \cdot x(t - \delta)$. Delay δ is sampled between $[1000, 2000]$ samples; strength $\alpha \sim [0.2, 0.5]$.

A.2 Implementation Details

Model Fine-Tuning and QLoRA. All Audio Language Models (ALMs) were fine-tuned using the HuggingFace `transformers` library on a single NVIDIA A100 GPU. We employ an instruction-tuning approach where the loss is calculated exclusively on the *Assistant* tokens (the reasoning and verdict), while masking the *System* prompt and *User* audio inputs. To manage the memory constraints of reasoning-heavy generation, we employed **QLoRA** (Quantized Low-Rank Adaptation).

Key hyperparameters include:

1. **Precision:** bfloat16 (BF16).
2. **Optimization:** AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$. Learning rate is set to 1e-4 with a linear decay scheduler.
3. **Batch Size:** Global batch size of 16 (2 per device $\times$ gradient accumulation).
4. **Regularization:** Weight decay of 0.1.
5. **Training Duration:** Models are trained until convergence (variable epochs) without early stopping.

The Low-Rank Adaptation (LoRA) hyperparameters were configured as follows:

1. **Rank (r):** 16
2. **Alpha (α):** 64
3. **Dropout:** 0.05
4. **Target Modules:** Query, Key, Value, and Output projection layers (`[q, k, v, o]_proj`)

A.3 Adversarial Attack Results

A.3.1 Acoustic Adversarial Attacks Results

Table 4 highlights the destabilizing effect of signal-level perturbations on the reasoning chain. We observe a distinct *Coherence Erosion* in weaker models; for instance, gemma-3n-E4B suffers a catastrophic drop in coherence under Shape Space attacks ($\Phi^{\text{ORG}}_{\text{Coh}}69.1\% \rightarrow \Phi^{\text{PER}}_{\text{Coh}}37.6\%$), indicating a *Panic Response* where the model struggles to articulate its observations. In contrast, Phi-4-multimodal demonstrates forensic utility through the *Silent Alarm* mechanism. Under Background Noise attacks, despite a high Attack Success Rate (44.8%), the model registers a significant spike in dissonance ($\Psi^{\text{PER}}_{\text{Diss}}41.3\%$), effectively warning the auditor of internal conflict even when the final classification fails. Qwen2-Audio remains the most robust, acting as a *Reasoning Shield* that maintains high coherence (77–80%) and relatively stable dissonance across all attack vectors.

A.3.2 Linguistic Adversarial Attacks Results

Table 5 reveals the insidious nature of linguistic attacks, which trigger *Systemic Deception* rather than panic. The most critical failure mode is observed in gemma-3n-E4B on the "American Female" profile. Here, the attack achieves a 100% Success Rate (ASR), yet the model maintains an exceptionally high coherence score ($\Phi^{\text{PER}}_{\text{Coh}}95.3\%$) and negligible dissonance ($\Psi^{\text{PER}}_{\text{Diss}}4.7\%$). This indicates that the linguistic complexity successfully decoupled the reasoning from the acoustic reality, forcing the model into a state of *Hallucinated Consistency* where it confidently fabricates justifications for the wrong verdict. Conversely, Qwen2-Audio proves resilient to these textual perturbations, maintaining low ASR (26%) and preserving the logical link between the transcript and the acoustic evidence.

A.4 Cognitive Coherence Under Attacks

Table 6 delineates the dimension-wise erosion of logical consistency. We observe a systematic breakdown across all forensic axes, but the degradation is most severe in the foundational dimensions of *Disfluency* and *Prosody*, which suffer absolute drops of 11.52% and 10.05% respectively on average. This indicates that adversarial perturbations primarily disrupt the model's ability to articulate fine-grained acoustic properties; the model struggles to form a coherent logical chain regarding the speaker's rhythm and hesitation patterns when the signal is

ARLMs	Strategy	NON		RSN					
		OC↑	ASR↓	OC↑	ASR↓	Φ_{Coh}^{ORG}	Φ_{Coh}^{PER}	Ψ_{Diss}^{ORG}	Ψ_{Diss}^{PER}
Qwen2-Audio-7B	Background Noise	99.1	50.3	97.1	54.2	86.2	77.6	46.0	27.3
Qwen2-Audio-7B	Shape Space	99.1	43.5	97.1	49.3	86.2	80.5	46.0	25.0
Qwen2-Audio-7B	Time Pitch	99.1	16.0	97.1	33.7	86.2	76.0	46.0	35.4
Phi-4-multimodal	Background Noise	94.4	34.5	88.8	44.8	88.4	72.1	15.9	41.3
Phi-4-multimodal	Shape Space	94.4	39.9	88.8	46.8	88.5	74.2	17.9	40.1
Phi-4-multimodal	Time Pitch	94.4	47.8	88.8	46.8	88.9	80.3	16.2	26.9
gemma-3n-E4B	Background Noise	99.8	41.7	77.3	51.1	72.0	49.9	82.6	55.6
gemma-3n-E4B	Shape Space	99.8	16.0	77.3	48.1	69.1	37.6	85.8	78.2
gemma-3n-E4B	Time Pitch	99.8	34.1	77.3	48.1	70.8	42.2	81.8	69.8
granite-3.3-8b	Background Noise	99.9	12.1	82.1	48.6	87.6	73.0	15.3	28.2
granite-3.3-8b	Shape Space	99.9	50.0	82.1	50.0	88.3	68.8	17.8	36.7
granite-3.3-8b	Time Pitch	99.9	41.2	82.1	50.5	88.0	78.3	18.5	18.6

Table 4: Performance breakdown under **Acoustic Adversarial Attacks**.

ARLMs	Voice Profile	NON		RSN					
		OC↑	ASR↓	OC↑	ASR↓	Φ_{Coh}^{ORG}	Φ_{Coh}^{PER}	Ψ_{Diss}^{ORG}	Ψ_{Diss}^{PER}
Qwen2-Audio-7B	American Female	44.9	98.9	98.3	26.3	87.4	80.2	10.3	9.6
Qwen2-Audio-7B	American Male	98.9	52.1	98.9	52.1	84.8	77.2	0.0	9.3
Qwen2-Audio-7B	British Female	39.2	98.4	98.4	28.1	95.7	87.5	0.0	12.9
Qwen2-Audio-7B	British Male	86.8	81.8	98.8	19.5	83.6	77.3	0.9	6.5
Phi-4-multimodal	American Female	88.3	69.2	72.8	94.7	82.5	64.3	34.1	36.2
Phi-4-multimodal	American Male	98.1	31.8	95.5	55.8	91.4	72.0	26.2	28.0
Phi-4-multimodal	British Female	100.0	5.6	99.4	39.5	91.3	67.4	20.4	38.5
Phi-4-multimodal	British Male	99.9	17.9	99.5	20.5	92.7	72.1	57.1	30.9
gemma-3n-E4B	American Female	3.1	100.0	54.4	100.0	50.0	95.3	4.3	4.7
gemma-3n-E4B	American Male	93.1	55.4	86.0	74.9	71.5	82.1	14.6	14.4
gemma-3n-E4B	British Female	94.7	62.3	75.8	89.1	61.6	87.8	11.3	10.7
gemma-3n-E4B	British Male	99.4	2.0	89.4	67.0	72.7	82.5	20.0	15.2
granite-3.3-8b	American Female	0.5	100.0	97.4	40.7	84.9	67.9	16.2	8.2
granite-3.3-8b	American Male	12.7	98.9	40.5	97.0	84.1	68.5	22.1	10.3
granite-3.3-8b	British Female	55.7	93.8	98.4	34.7	86.5	65.1	23.9	10.6
granite-3.3-8b	British Male	74.3	84.5	98.2	34.7	86.1	66.7	26.3	10.5

Table 5: Performance breakdown under **Linguistic Adversarial Attacks**. All values are reported in percentages %.

contaminated. Notably, the *Quality* dimension also exhibits a sharp decline ($\sim 9.3\%$), reflecting the model's confusion in distinguishing between environmental noise (a natural feature) and adversarial artifacts (an attack signature).

A.5 Cognitive Dissonance Under Attacks

Table 7 reveals which forensic dimensions act as the most effective "Silent Alarms." The *Quality* dimension emerges as the most sensitive indicator of manipulation, particularly for the Phi-4-multimodal model, where dissonance surges from 30.18% to 38.56% under attack. This suggests that even when the model fails to classify the audio as a deepfake, its reasoning layer remains highly conflicted about the acoustic artifacts present in the recording. Conversely, dimensions like *Liveliness* show mixed behaviors; while Qwen2-Audio suppresses conflict (acting as a black box), weaker models like Gemma often exhibit lower dissonance under attack (e.g., Liveliness drops $44.68\% \rightarrow 37.58\%$), further evidence of the dangerous *Systemic Deception* where the model rationalizes the loss of vital signs.

B Datasets and Prompts

B.1 Reasoning Data Synthesis and Iterative Refinement

To adapt the ASVSpoof 2019 benchmark for reasoning-based forensic analysis, we employ a *Cold Start* data synthesis strategy inspired by the DeepSeek-R1 pipeline (DeepSeek-AI et al., 2025). We first designed a specialized Chain-of-Thought (CoT) prompt 6 based on the human forensic taxonomy established by (Warren et al., 2024), focusing on reasoning dimensions such as prosody, disfluency, and speaking style. This prompt was used for few-shot generation on GPT-5 to create an initial seed dataset of several thousand samples, and later

ALMs	Prosody		Disfluency		Speed		Speaking Style		Liveliness		Quality	
	ORG	PER	ORG	PER	ORG	PER	ORG	PER	ORG	PER	ORG	PER
Qwen2-Audio-7B	87.88	80.10	86.74	78.43	87.97	81.20	87.20	80.49	85.52	76.88	87.67	79.78
Phi-4-multimodal	90.14	72.72	88.80	71.95	89.94	72.29	90.00	73.25	88.05	69.66	87.65	70.69
gemma-3n-E4B	66.62	70.40	68.81	66.76	66.07	68.57	64.61	68.90	67.87	66.17	67.00	68.38
granite-3.3-8b	88.62	69.85	87.64	68.77	88.69	70.53	85.52	70.96	85.45	69.05	83.13	69.43
Average	83.32	73.27	83.00	71.48	83.17	73.15	81.83	73.40	81.72	70.44	81.36	72.07

Table 6: Φ_{Coh} analysis, ORG is original reasoning, and PER is perturbed reasoning

ALMs	Prosody		Disfluency		Speed		Speaking Style		Liveliness		Quality	
	ORG	PER	ORG	PER	ORG	PER	ORG	PER	ORG	PER	ORG	PER
Qwen2-Audio-7B	23.68	17.67	21.41	18.97	22.17	16.90	22.17	17.69	21.41	18.47	16.98	18.29
Phi-4-multimodal	18.93	29.37	28.58	34.83	26.73	34.31	33.36	34.60	23.16	35.61	30.18	38.56
gemma-3n-E4B	40.50	34.13	43.55	37.20	42.49	34.14	42.44	34.47	44.68	37.58	43.90	35.66
granite-3.3-8b	20.41	16.90	20.51	17.63	20.91	17.73	19.82	17.06	16.44	17.51	21.97	18.73
Average	25.88	24.52	28.51	27.16	28.08	25.77	29.45	25.95	26.42	27.29	28.26	27.81

Table 7: Ψ_{Diss} analysis, ORG is original reasoning, and PER is perturbed reasoning

for finetuning ALMs and inferences.

To scale and refine this data, we implemented a self-correction loop through iterative bootstrapping. In each iteration, we fine-tuned Qwen2-Audio on the current reasoning set and then utilized the model to re-generate reasoning traces for the entire training set. To ensure the stability and quality of the synthetic logic, we employed a majority-voting strategy for each sample, generating three independent reasoning paths and selecting the most consistent one. This cycle was repeated until classification accuracy converged (occurring after 3 iterations), resulting in a final dataset of $25,108$ items. The final training set composition includes $22,627$ fake samples (average CoT length: 344.89 tokens) and $2,481$ real samples (average CoT length: 302.01 tokens).

While this iterative refinement significantly improves the model's ability to articulate acoustic features, it introduces a potential architectural bias. Because the reasoning traces were optimized and filtered via Qwen2-Audio's internal representations, the resulting dataset is highly aligned with its specific processing style. Consequently, while general acoustic perception is increased across all models, the ultimate reasoning consistency is most robust in Qwen2-Audio, whereas other models like Phi-4 or Gemma may encounter a "reasoning tax" due to this domain-specific data alignment.

B.2 Perception Audit: Dataset and Question Bank Construction

To evaluate the groundedness of ALM reasoning (RQ1), we construct a dedicated audit dataset $\mathcal{D}_{audit}$ and a forensic question bank $\mathcal{Q}_k$. We first utilize the fine-tuned Qwen2-Audio model to process the ASVSpoof 2019 test set under various acoustic perturbations. To ensure a balanced evaluation of reasoning behavior across all possible failure modes, we select a uniform distribution of 100 representative samples. This subset consists of 25 "Successful" attacks on fake audio (False Negatives), 25 "Failed" attacks on fake audio (True Positives), 25 "Successful" attacks on real audio (False Positives), and 25 "Failed" attacks on real audio (True Negatives).

The construction of the audit environment follows a two-step generative process. First, to create the question bank $\mathcal{Q}_k$, we extract the forensic dimensions from our reasoning traces and utilize Gemini Pro to formulate a series of targeted binary questions for each aspect (e.g., Prosody, Disfluency). Second, to establish the ground-truth labels for $\mathcal{D}_{audit}$, we employ an ensemble of frontier models, including Gemini-3-Pro, Gemini-3-Flash, and GPT-5.2. We use a majority-voting mechanism across these models to determine the objective acoustic properties of each clip.

We acknowledge the inherent limitations of using LLM-based labeling as a proxy for human expert annotation and a small sample size $N = 100$, we observe that the resulting perception scores are highly consistent with established acoustic baselines, providing a reliable benchmark for evaluating perceptual blindness in ALMs.

B.2.1 Forensic Question Bank ($\mathcal{Q}_k$)

1. **Prosody:** Does the audio sound like a naturally produced human recording? Is the pitch contour flat and devoid of variation? Does the speaker's

```
You are an expert in speech analysis in audio deepfake detection. Please analyze the speech
under these criteria, then provide the conclusion and clearly state the audio is $fake$ or
$real$ (put your decision in $$).

- Prosody: Analyze Tone, Inflections, Cadence, Pitch, Monotone, Raspy, Emotion,...
- Disfluency: Analyze Pauses, Filler Words
- Speed: Analyze Fast, Slow, Rushed
- Speaking Style: Accent, List, Articulation, Specific Word Pronunciation
- Liveliness: like do you hear the breathing, or robotic sound
- Quality: analyze background noise, computer noise, microphone feedback noise,...
- Conclusion: Based on above analysis, give your inductive reasoning and clearly state the audio
is $fake$ or $real$ (put your decision in $$)
```

Figure 6: Chain-of-Thoughts prompt with step-by-step instructions

```
- Prosody: The prosody is quite unnatural. The pitch is consistently high, which is unusual for
a human voice. There is a noticeable flatness in the delivery, which is often a sign of a voice
synthesizer. The cadence is also quite unnatural, lacking the natural pauses and inflections
that would be expected in a conversation.
- Disfluency: The speech is perfectly fluent, which is a significant red flag. There are no
filler words or pauses, which are typical of human speech. This suggests that the speech was
likely generated by a voice synthesizer.
- Speed: The speaking rate is quite slow, which is unusual for a human voice. This suggests that
the speech was likely generated by a voice synthesizer.
- Speaking Style: The articulation is quite clear, which is unusual for a human voice. However,
there are some noticeable errors in pronunciation, such as the incorrect use of "lending" (00:03)
and the incorrect stress on "recently" (00:06). This suggests that the speech was likely generated
by a voice synthesizer.
- Liveliness: There are no audible signs of life in the speech, such as breathing or other
vocalizations. This suggests that the speech was likely generated by a voice synthesizer.
- Quality: The audio quality is quite low, with a noticeable amount of noise. This suggests
that the speech was likely recorded with a low-quality microphone, which is common for voice
synthesizers.
- Conclusion: Based on the analysis, the speech is likely $fake$. The audio quality is low, the
speech is perfectly fluent and lacks any signs of life, and the prosody, disfluency, and speaking
style are all indicative of a voice synthesizer.
```

Figure 7: ASVSpoof 2019, a reasoning training example for a **fake** audio

```
- Prosody: The prosody of the speech is quite natural and consistent. There are no noticeable
variations in pitch or intonation. The speech is delivered in a calm and steady manner, which
is typical of a real person speaking in a neutral mood.
- Disfluency: The speech is quite fluent, with no noticeable hesitations or pauses. There are no
fillers or pauses between words, which indicates that the speech was delivered smoothly without
any interruptions. This is typical of a real person speaking in a calm and steady manner.
- Speed: The speech is delivered at a moderate pace, which is typical of a real person speaking in
a neutral mood. There are no variations in speed, which indicates that the speech was delivered
at a consistent pace.
- Speaking Style: The articulation is clear and precise, with no noticeable errors or
mispronunciations. The accent is neutral and consistent, which indicates that the speaker
is speaking in a neutral mood.
- Liveliness: There are no noticeable signs of liveliness or emotion in the speech. The voice is
calm and steady, which is typical of a real person speaking in a neutral mood.
- Quality: The audio quality is quite good, with no noticeable background noise or distortion.
The recording is clear and crisp, which indicates that the speech was recorded in a professional
environment.
- Conclusion: Based on the analysis, the speech appears to be real. The prosody, disfluency,
speed, speaking style, articulation, and quality are all consistent with a real person speaking
in a neutral mood. There are no noticeable signs of artificiality or manipulation. $real$
```

Figure 8: ASVSpoof 2019, a reasoning training example for a **real** audio

tone match the implied emotion of the words? Does the speech exhibit a machine-like or repetitive cadence?

2. **Disfluency:** Is the speech completely devoid of fillers or breathing pauses? Does the speaker exhibit natural fillers like 'um' or 'uh'? Are there excessive or glitch-like repetitions of specific sounds? Do pauses occur at logical syntactic boundaries?
3. **Speed:** Is the overall speaking rate within a natural range? Does the speaker maintain a rigid, machine-like consistency in speed? Does the tempo fluctuate erratically within a single sentence?
4. **Speaking Style:** Does the speaker articulate with unnatural, robotic precision? Does the delivery sound like scripted 'read speech'? Are there instances of garbled or fractured articulation?
5. **Liveliness:** Are audible breathing sounds or natural pauses present? Does the voice possess natural warmth versus a sterile quality? Does the audio sound unnaturally clean, as if in an acoustic void?
6. **Quality:** Are there audible digital artifacts, metallic ringing, or static? Does the audio contain natural environmental cues (reverb/reflections)? Is the audio quality consistent from beginning to end?

Note: Only a subset of the full bank is shown. See our github for all 50+ questions.

C AI Assistants Usages

During the preparation of this work, we utilized several generative AI assistants to support the research pipeline and manuscript development. Specifically, we employed GPT-5 and $Gemini$-3 to execute the *Cold Start* data synthesis process, generating the initial reasoning traces and the question bank $\mathcal{Q}_k$. For the manuscript preparation,$Gemini$-3 was used to assist with grammar checking, stylistic refinement, and the selection of precise word choices.

We have rigorously reviewed and edited all AI-assisted outputs to ensure scientific accuracy. We maintain full responsibility for the final content of this publication.