# ECHODISTILL: ROBUST LARGE AUDIO LANGUAGE MODELS VIA NOISY-TO-CLEAN SELF-DISTILLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Audio Language Models (LALMs) remain vulnerable to acoustic noise, which can obscure task-relevant evidence and produce unreliable responses. We propose **`EchoDistill`**, a noisy-to-clean self-distillation framework that uses clean audio as privileged information during post-training. A noisy-input student samples candidate responses reflecting its inference-time behavior, while a frozen copy of the same backbone processes the corresponding clean audio. **`EchoDistill`** combines masked response-token distillation, task-gated consistency shaping, and teacher-referenced group-relative optimization to align noisy-input generation with clean-conditioned semantics. Only the student is retained at inference time, introducing no additional inference cost. Across three LALM backbones and three audio domains at $-10\,\mathrm{dB}$, **`EchoDistill`** improves average noisy-input accuracy by 1.63 percentage points over the strongest baseline. On Qwen2.5-Omni, it raises noisy-input accuracy from 59.33% to 62.94%, while clean-audio accuracy increases from 76.56% to 77.56%. Replacing matched audio with random, shuffled, or silent inputs reduces accuracy by 3.08–6.42 points, confirming that matched acoustic evidence contributes to its predictions. Additional evaluations show improvements on held-out additive noises and external benchmarks, while revealing that these gains do not reliably extend to non-additive distortions. These results demonstrate robust post-training improvements under severe additive noise without sacrificing clean-audio capability across diverse tasks. Our codes are available at https://anonymous.4open.science/r/EchoDistill-45DA.

## 1 INTRODUCTION

Large Audio Language Models (LALMs) Chu et al. (2024); Tang et al. (2024); Ding et al. (2025); Yang et al. (2026a) have emerged as a vital research direction, leveraging LLMs' advanced representation learning for audio processing and a wide range of applications, including automatic speech recognition An et al. (2025); Shi et al. (2026), translation Huang et al. (2023); Du et al. (2024) and text-to-speech Hu et al. (2026); Gong et al. (2026). However, audio captured in practical environments is frequently contaminated by additive environmental noise, which distorts the observed waveform and masks task-relevant acoustic cues Song et al. (2026b); Zhang et al. (2026a). Consequently, the accuracy and reliability of LALM outputs can degrade substantially Yin et al. (2024); Kumar & Mishra (2025); Luo et al. (2026b); Lin et al. (2026b); Luo et al. (2026a). Under severe noise, models may misinterpret acoustic evidence and produce unstable responses, leading to semantic drift and hallucinations Chen et al. (2025); Cheng et al. (2025); Yin et al. (2026).

Existing research on LALM denoising has predominantly focused on inference-time interventions or feature-level modifications Chen et al. (2025); Yin et al. (2026); Chen et al. (2026); Zhang et al. (2026b). These approaches suppress estimated noise or attenuate noise-related representations, but do not teach the model to recover clean-conditioned semantics from corrupted inputs, leaving its predictions vulnerable to semantic drift under severe noise. Moreover, nuisance components and task-relevant acoustic cues are entangled in the learned representation space; aggressive suppression may discard useful evidence and shift acoustic representations away from the distribution expected by the language backbone Song et al. (2026b); Zhang et al. (2026b), thereby degrading downstream accuracy or general capabilities. Although methods such as DFL Purwins et al. (2019) can mitigate



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

certain noise corruptions, a recent study SEE Zhang et al. (2026b) shows that such interventions may incur a representation mismatch and downstream degradation.

In this work, we introduce `EchoDistill`, a noisy-to-clean self-distillation framework for improving the robustness of LALMs Xiaomi (2025); Yu et al. (2026); Lin et al. (2026a) to severe additive acoustic noise. Our key observation is that paired clean and noisy recordings preserve the same task-relevant semantics, while differing in the acoustic corruption applied to the observation. Consequently, under the same response context, a change in the model's next-token preference between the two recordings directly reveals the semantic decision altered by noise López et al. (2025); Ma et al. (2025a). Based on this relationship, `EchoDistill` uses the clean-conditioned prediction as a reference for what the model would have predicted from the same content without corruption, and corrects the corresponding noisy-conditioned token preference before such local errors accumulate into sequence-level semantic drift. Unlike existing methods that estimate and suppress noise in the waveform or representation space, this approach directly corrects the semantic consequence of noise without assuming that interference can be cleanly separated from task-relevant acoustic evidence.

To implement this principle, `EchoDistill` exploits an information asymmetry available during post-training. The student receives noisy audio and generates candidate responses, whereas a frozen copy of the same backbone processes the corresponding clean audio. For each student-generated prefix, we align the paired next-token distributions through masked response-token distillation. Noisy-input rollouts expose the decisions encountered at inference time, while a teacher-referenced group-relative clipped objective reinforces trajectories corrected toward clean-conditioned semantics. A task-positive consistency term further prevents agreement with an incorrect teacher prediction from being rewarded. Clean audio and the teacher branch are used only during post-training; at inference time, `EchoDistill` retains the original single-model architecture without additional cost.

As shown in Figure 1, we evaluate `EchoDistill` under severe additive acoustic noise at $-10\,\mathrm{dB}$ across three representative LALMs: Qwen2.5-Omni Fan et al. (2025), MiniCPM-o Yao et al. (2024), and Step-Audio-2-mini Wu et al. (2025) and three audio domains: Music, Sound, and Speech Sakshi et al. (2025). We use noisy-input accuracy as the primary metric of task correctness and report Generation Success Rate (GSR) as an auxiliary measure of behavioral consistency. Averaged over the nine backbone–domain settings, `EchoDistill` improves noisy-input accuracy by 1.63 percentage points and GSR by 4.18 percentage points over the strongest corresponding baselines. These improvements show that `EchoDistill` not only stabilizes clean–noisy behavior but also converts this stability into more accurate predictions. Our contributions are as follows:

- ❶We formulate LALM robustness post-training as *on-trajectory semantic correction with privileged clean audio*. For each noisy-student response prefix, paired clean and noisy next-token distributions reveal how acoustic corruption shifts semantic decisions, rather than relying on feature suppression or offline teacher responses.
- ❷We propose `EchoDistill`, a closed-loop noisy-to-clean self-distillation framework. Noisy rollouts expose inference-time errors, masked token-level distillation provides clean-conditioned corrections along the same trajectories, and task-positive consistency shaping prevents rewarding incorrect teacher agreement. A teacher-referenced group-relative objective reinforces the corrected trajectories without additional inference cost.
- ❸We conduct a correctness-centered evaluation across three LALM backbones and three audio domains. Clean-audio retention, controlled baselines, SNR sweeps, held-out noises, cross-benchmark transfer, and audio interventions demonstrate improvements under severe additive noise while identifying the generalization boundary under non-additive distortions.

## 2 Related Work and Background

### 2.1 Related Work

**Noise robustness for speech and LALMs.** Traditional robust speech systems use data augmentation, speech enhancement, and source separation to reduce acoustic mismatch (Chen et al., 2025; Yin et al., 2026; Song et al., 2026b; Zhang et al., 2026b). Recent LALM methods further introduce task-conditioned enhancement or representation-level noise suppression (Yin et al., 2024; Chen et al., 2025; Yin et al., 2026; Zhang et al., 2026b). These approaches modify waveforms or acoustic repre-

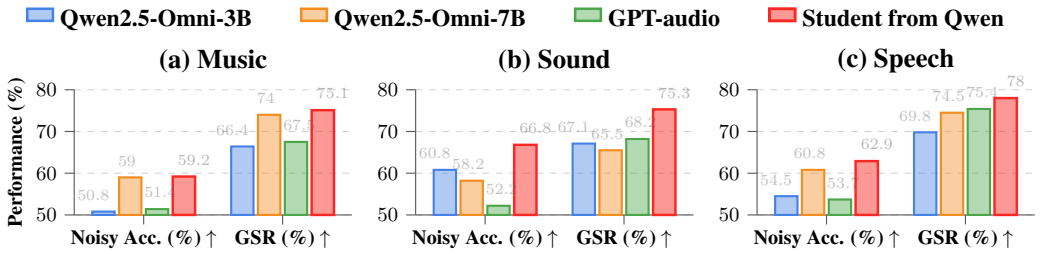


Figure 1: Task-wise noisy-input accuracy and Generation Success Rate (GSR) under severe additive noise at $-10\,\mathrm{dB}$. Accuracy measures task correctness, whereas GSR measures clean–noisy prediction consistency. EchoDistill improves both metrics across the Music, Sound, and Speech domains.

sentations and may remove task-relevant cues when signal and interference overlap (Purwins et al., 2019; Zhang et al., 2026b). Noise-augmented SFT instead provides answer-level supervision but cannot localize corruption-induced errors along noisy-input trajectories. `EchoDistill` addresses this limitation through on-trajectory semantic correction without adding an inference-time enhancement module.

**Knowledge distillation and on-policy self-distillation.** Knowledge distillation transfers predictive distributions from a teacher to a student (Hinton et al., 2015). Offline distillation may suffer from a mismatch between fixed training sequences and student-generated trajectories (Yang et al., 2026b), whereas on-policy distillation queries the teacher on the student's own outputs (Agarwal et al., 2024; Zhao et al., 2026; Ko et al., 2026). Unlike conventional teacher imitation, `EchoDistill` evaluates each student-generated prefix under paired clean and noisy audio. Their next-token discrepancy estimates the semantic shift caused by acoustic corruption and provides a trajectory-specific correction signal. Extended discussion is provided in Appendix A.1.

### 2.2 Background

**Self-distillation.** Let $p_\phi$ and $p_\theta$ denote the teacher and student distributions. Output-level self-distillation is commonly formulated as

$$\mathcal{L}_{\mathrm{SD}} = \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\left(p_\phi(\cdot \mid x, y_{<t}) \,\|\, p_\theta(\cdot \mid x, y_{<t})\right). \tag{1}$$

In `EchoDistill`, the two distributions share the same student-generated prefix but are conditioned on clean and noisy audio, respectively.

**Large Audio Language Models.** LALMs couple an audio encoder with a language backbone (Huang et al., 2024; Xu et al., 2025; Luo et al., 2026b). Given audio $a$ and prompt $x$, their generation process can be written as

$$\pi(y \mid x, a) = \prod_{t=1}^{T} \pi(y_t \mid x, g(E(a)), y_{<t}), \tag{2}$$

where $E(\cdot)$ is the audio encoder and $g(\cdot)$ maps acoustic features into the language model. Noise can corrupt this conditioning representation and consequently shift the generated response.

**Group Relative Policy Optimization.** Group Relative Policy Optimization (GRPO) (Shao et al., 2024) samples a group of responses $\{y^{(k)}\}_{k=1}^{K}$ for each input and estimates their relative advantages by normalizing rewards within the group:

$$A^{(k)} = \frac{r^{(k)} - \mu}{\sigma + \varepsilon}, \tag{3}$$

where $\mu$ and $\sigma$ denote the group reward mean and standard deviation. GRPO then optimizes a clipped importance-ratio objective between the current and old policies. However, because the same sequence-level advantage is assigned to all response tokens, standard GRPO cannot identify which token-level decisions are altered by acoustic corruption.

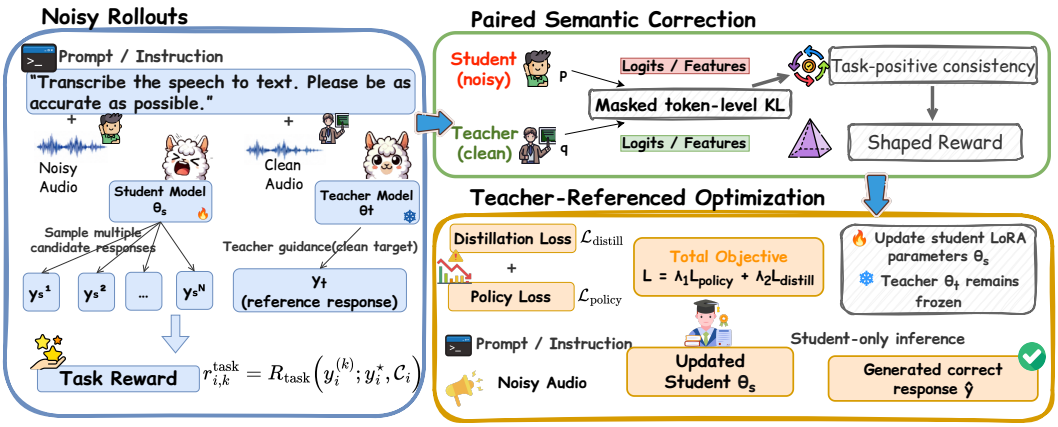


Figure 2: **Overview of `EchoDistill`.** The noisy-input student first samples candidate responses. The student and frozen clean-audio teacher produce paired token distributions, whose masked discrepancy provides token-level correction and a task-gated consistency bonus. A teacher-referenced group-relative objective updates only the student, which is used alone at inference time.

# 3 Methodology

## 3.1 Problem Formulation

We consider post-training a LALM using paired clean and noisy recordings that convey the same semantic content. Let

$$\mathcal{D} = \{(x_i, a_i^{\mathrm{n}}, a_i^{\mathrm{c}}, y_i^{\star}, \mathcal{C}_i)\}_{i=1}^{N}, \tag{4}$$

where $x_i$ is the textual instruction, $a_i^{\mathrm{n}}$ and $a_i^{\mathrm{c}}$ are the noisy and clean versions of the same audio, $y_i^{\star}$ is the ground-truth answer, and $\mathcal{C}_i$ denotes the candidate-answer set when applicable.

We initialize a trainable student policy $\pi_\theta$ and a frozen teacher $\pi_\phi$ from the same pretrained model. The student receives $(x_i, a_i^{\mathrm{n}})$, whereas the teacher receives $(x_i, a_i^{\mathrm{c}})$. The clean audio therefore serves as privileged information during post-training rather than as an additional inference-time input. Our objective is to improve $\pi_\theta(y \mid x_i, a_i^{\mathrm{n}})$ while preserving its performance on uncorrupted audio. At inference time, only the trained student and the observed audio are required.

## 3.2 `EchoDistill` Framework

`EchoDistill` combines three complementary components: noisy-input rollouts that expose the student's inference-time behavior, token-level noisy-to-clean distillation that supplies dense clean-conditioned supervision, and task-gated consistency shaping that differentiates sampled responses using their candidate-specific clean–noisy discrepancy.

**Noisy-input rollouts.** At the beginning of each update, we retain a fixed rollout policy $\pi_{\theta_{\mathrm{old}}}$ and sample $K$ candidate responses:

$$y_i^{(k)} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x_i, a_i^{n}), \qquad r_{i,k}^{\mathrm{task}} = R_{\mathrm{task}}\left(y_i^{(k)}; y_i^{\star}, \mathcal{C}_i\right), \quad k = 1, \ldots, K. \tag{5}$$

**Candidate-specific noisy-to-clean distillation.** For each sampled response, we evaluate the frozen teacher and trainable student along the same textual prefix but under different acoustic conditions. Here, $m_{i,k,t}$ selects assistant-response tokens and excludes prompt tokens, audio placeholders, and dialogue-template tokens. The stop-gradient operator $\mathrm{sg}[\cdot]$ prevents gradients from flowing

through the frozen teacher. The instance-level distillation objective is

$$D_{i,k}(\theta) = \frac{\sum_t m_{i,k,t} D_{\mathrm{KL}}\left(\mathrm{sg}[q_{i,k,t}] \| p^{\theta}_{i,k,t}\right)}{\sum_t m_{i,k,t}}, \qquad \mathcal{L}^{(i)}_{\mathrm{distill}} = \frac{1}{K}\sum_{k=1}^{K} D_{i,k}(\theta). \tag{6}$$

Unlike answer-level imitation, this objective transfers the teacher's complete next-token distribution along student-generated trajectories.

**Task-gated clean–noisy consistency shaping.** We use the candidate-specific discrepancy to augment the task reward. To keep the reward fixed within an optimization update, the score is evaluated using the rollout policy and detached from gradient computation:

$$s_{i,k} = \mathrm{sg}\left[\exp\left(-D_{i,k}(\theta_{\mathrm{old}})\right)\right]. \tag{7}$$

The shaped reward is

$$\bar{r}_{i,k} = \mathrm{clip}\left(r^{\mathrm{task}}_{i,k} + \beta b_{i,k} s_{i,k}, -1, 2\right), \tag{8}$$

where $\beta$ controls the strength of the consistency bonus. The correctness gate $b_{i,k}$ prevents an incorrect response from receiving a bonus merely because its noisy- and clean-conditioned distributions agree. Among verified correct responses, candidate-specific scores allow different trajectories to receive different bonuses.

We refer to $s_{i,k}$ as a clean–noisy consistency score rather than a direct measure of acoustic grounding. Distributional agreement alone does not prove that a prediction is causally grounded in the audio; this distinction is examined separately through controlled audio interventions.

### 3.3 Teacher-Referenced Group-Relative Optimization

We compute the mean response-token log-likelihoods of each sampled candidate under the noisy student and clean-conditioned teacher:

$$\ell^{z}_{i,k}(\psi) = \frac{\sum_{t=1}^{T_{i,k}} m_{i,k,t} \log \pi_{\psi}\left(y^{(k)}_{i,t} \mid x_i, a^{z}_i, y^{(k)}_{i,<t}\right)}{\sum_{t=1}^{T_{i,k}} m_{i,k,t}}, \quad (z,\psi) \in \{(\mathrm{n},\theta),(\mathrm{c},\phi)\}. \tag{9}$$

Their difference defines a teacher-referenced likelihood weight:

$$w^{\mathrm{TCR}}_{i,k}(\theta) = \exp\left(\ell^{\mathrm{n}}_{i,k}(\theta) - \mathrm{sg}\left[\ell^{\mathrm{c}}_{i,k}(\phi)\right]\right). \tag{10}$$

We optimize the following teacher-referenced group-relative clipped surrogate:

$$\mathcal{L}^{(i)}_{\mathrm{policy}} = -\frac{1}{K}\sum_{k=1}^{K} \min\left(w^{\mathrm{TCR}}_{i,k} A_{i,k}, \mathrm{clip}\left(w^{\mathrm{TCR}}_{i,k}, 1-\epsilon, 1+\epsilon\right) A_{i,k}\right). \tag{11}$$

Because the numerator and denominator of $w^{\mathrm{TCR}}_{i,k}$ are conditioned on different acoustic inputs, this quantity is not an importance-sampling ratio between a current and an old policy. Instead, it is a response-dependent clean-reference weight: the clean-conditioned likelihood provides a semantic reference scale, while clipping limits the influence of extreme likelihood differences. Accordingly, we describe this component as a teacher-referenced clipped surrogate rather than standard GRPO.

For a minibatch of size $B$, the complete training objective is

$$\mathcal{L}_{\mathrm{total}} = \frac{1}{B}\sum_{i=1}^{B}\left[\lambda_{\mathrm{policy}}\mathcal{L}^{(i)}_{\mathrm{policy}} + \lambda_{\mathrm{distill}}\mathcal{L}^{(i)}_{\mathrm{distill}}\right], \tag{12}$$

where $\lambda_{\mathrm{policy}}$ and $\lambda_{\mathrm{distill}}$ balance group-relative optimization and token-level distillation. Only $\theta$ is updated; $\phi$ remains frozen throughout post-training.

**Inference.** After post-training, the teacher branch, clean audio, rollout sampling, and reward computation are discarded. Given an observed audio input $a$, the model generates

$$y \sim \pi_{\theta}(\cdot \mid x, a). \tag{13}$$

Table 1: Main results under severe additive noise (SNR $= -10$ dB). `Acc` denotes noisy-input task accuracy and is our primary metric; `Noisy` and `GSR` are auxiliary clean–noisy agreement diagnostics. Red and blue values denote percentage-point increases (↑) and decreases (↓) relative to the corresponding Initial model. The best and second-best results are shown in **bold** and underlined.

| Domain | Method | *Step-Audio-2-mini* | | | *Qwen2.5-Omni* | | | *MiniCPM-o-2.6* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Noisy | Acc | GSR | Noisy | Acc | GSR | Noisy | Acc | GSR |
| Music | **Initial** | — | 59.17 | 78.67 | — | 59.00 | 74.00 | — | 63.63 | 85.15 |
| | **WT** | 75.08 | 59.08 ↓0.09 | 77.33 ↓1.34 | 71.33 | **59.33** ↑0.33 | 74.16 ↑0.16 | 85.10 | 64.06 ↑0.43 | 86.95 ↑1.80 |
| | **STFT** | **78.16** | **60.58** ↑1.41 | 79.00 ↑0.33 | 73.91 | 58.75 ↓0.25 | **75.41** ↑1.41 | **86.49** | 64.55 ↑0.92 | 87.15 ↑2.00 |
| | **DFL** | 74.91 | 59.08 ↓0.09 | 75.66 ↓3.01 | 72.75 | 56.58 ↓2.42 | 74.50 ↑0.50 | 85.40 | **64.82** ↑1.19 | 86.57 ↑1.42 |
| | **SEEN** | 70.75 | 57.00 ↓2.17 | 77.08 ↓1.59 | 74.08 | 58.91 ↓0.09 | 73.91 ↓0.09 | 85.47 | 63.63 ↓0.00 | 85.05 ↓0.10 |
| | **EchoDistill** | **78.16** | 59.75 ↑0.58 | **86.00** ↑7.33 | **74.58** | 59.17 ↑0.17 | 75.08 ↑1.08 | 86.32 | 64.69 ↑1.06 | **87.24** ↑2.09 |
| Sound | **Initial** | — | 63.50 | 70.42 | — | 58.17 | 65.50 | — | 61.92 | 77.76 |
| | **WT** | 69.00 | 63.33 ↓0.17 | 70.50 ↑0.08 | 61.58 | 56.66 ↓1.51 | 63.41 ↓2.09 | 75.27 | 60.86 ↓1.06 | 77.66 ↓0.10 |
| | **DFL** | 71.50 | 63.50 ↓0.00 | 74.58 ↑4.16 | 64.66 | 59.75 ↑1.58 | 67.41 ↑1.91 | **78.72** | 61.55 ↓0.37 | 80.31 ↑2.55 |
| | **STFT** | 70.33 | **63.75** ↑0.25 | 69.91 ↓0.51 | 67.75 | 60.75 ↑2.58 | 68.25 ↑2.75 | 77.11 | 60.93 ↓0.99 | 77.78 ↑0.02 |
| | **SEEN** | 67.25 | 59.83 ↓3.67 | 77.16 ↑6.74 | 66.25 | 59.33 ↑1.16 | 66.42 ↑0.92 | 77.20 | 61.56 ↓0.36 | 78.04 ↑0.28 |
| | **EchoDistill** | **71.75** | 63.67 ↑0.17 | **80.08** ↑9.66 | **72.83** | **66.83** ↑8.66 | **75.33** ↑9.83 | 77.14 | **62.18** ↑0.26 | **80.42** ↑2.66 |
| Speech | **Initial** | — | 55.42 | 69.15 | — | 60.80 | 74.56 | — | 57.27 | 74.64 |
| | **WT** | 58.07 | 46.62 ↓8.80 | 59.49 ↓9.66 | 66.52 | 53.21 ↓7.59 | 66.52 ↓8.04 | 68.16 | 50.82 ↓6.45 | 69.43 ↓5.21 |
| | **DFL** | 65.55 | 51.83 ↓3.59 | 64.88 ↓4.27 | 72.65 | 58.46 ↓2.34 | 72.15 ↓2.41 | 72.66 | 55.17 ↓2.10 | 75.31 ↑0.67 |
| | **STFT** | 68.22 | 54.25 ↓1.17 | 68.47 ↓0.68 | 74.30 | 59.09 ↓1.71 | 73.97 ↓0.59 | 71.80 | 55.62 ↓1.65 | 75.11 ↑0.47 |
| | **SEEN** | 66.63 | 51.34 ↓4.08 | 70.48 ↑1.33 | 75.14 | 59.79 ↓1.01 | 74.47 ↓0.09 | 73.07 | 57.91 ↑0.64 | 73.27 ↓1.37 |
| | **EchoDistill** | **71.04** | **55.29** ↓0.13 | **74.89** ↑5.74 | **78.67** | **62.89** ↑2.09 | **78.01** ↑3.45 | **73.75** | **58.49** ↑1.22 | **76.46** ↑1.82 |
| Average | **Initial** | — | 59.36 | 72.75 | — | 59.32 | 71.35 | — | 60.94 | 79.18 |
| | **WT** | 67.38 | 56.34 ↓3.02 | 69.11 ↓3.64 | 66.48 | 56.40 ↓2.92 | 68.03 ↓3.32 | 76.18 | 58.58 ↓2.36 | 78.01 ↓1.17 |
| | **STFT** | 72.24 | 59.53 ↑0.17 | 72.46 ↓0.29 | 71.99 | 59.53 ↑0.21 | 72.54 ↑1.19 | 78.47 | 60.37 ↓0.57 | 80.01 ↑0.83 |
| | **DFL** | 70.65 | 58.14 ↓1.22 | 71.71 ↓1.04 | 70.02 | 58.26 ↓1.06 | 71.35 ↓0.00 | 78.93 | 60.51 ↓0.43 | 80.73 ↑1.55 |
| | **SEEN** | 68.21 | 56.06 ↓3.30 | 74.91 ↑2.16 | 71.82 | 59.34 ↑0.02 | 71.60 ↑0.25 | 78.58 | 61.03 ↑0.09 | 78.79 ↓0.39 |
| | **EchoDistill** | **73.65** | **59.57** ↑0.21 | **80.32** ↑7.57 | **75.36** | **62.96** ↑3.64 | **76.14** ↑4.79 | **79.07** | **61.79** ↑0.85 | **81.37** ↑2.19 |

# 4 EXPERIMENTS

We evaluate **EchoDistill** through three research questions:

- **RQ1:** Does **EchoDistill** outperform controlled baselines without degrading clean-audio performance, and what do its ablations reveal?
- **RQ2:** How well does **EchoDistill** generalize across noise conditions and benchmarks? Under which corruption conditions does the method cease to provide gains?
- **RQ3:** To what extent do the predictions of **EchoDistill** depend on acoustic evidence matched to the question, rather than clean–noisy output agreement or input-independent priors?

## 4.1 EXPERIMENTAL SETUP

**Models and data.** We use a clean-teacher/noisy-student setting. The teacher and student share the same backbone; the teacher receives clean audio and is frozen, while the student receives noisy audio and is updated by EchoDistill. Main experiments cover *Qwen2.5-Omni* Fan et al. (2025), *MiniCPM-o-2.6* Yao et al. (2024), and *Step-Audio-2-mini* Wu et al. (2025); ablations are conducted on *Qwen2.5-Omni*. Training and validation use paired clean/noisy audio from MMAR Ma et al. (2025b), with 14,397 training samples and 1,704 validation samples. The data cover 10 noise types Hu & Wang (2010) and 7 SNR levels $\{-10, -5, 0, 5, 10, 20, 30\}$Zhang et al. (2026b). Main evaluation is conducted on noisy MMAU (Sakshi et al., 2025) at SNR $= -10$.

**Baseline Methods and Evaluation Metrics.** We compare EchoDistill with four representative noise-robustness methods, including the frequency-domain processing method `STFT`, the wavelet-transform method `WT`, the deep-learning-based denoising method `DFL` Purwins et al. (2019), and the representation-level noise suppression method `SEE` Zhang et al. (2026b). Specifically, `STFT` and `WT` represent traditional signal-processing-based front-end denoising methods, `DFL` represents a learning-based speech enhancement method, and `SEE` directly suppresses noise-related components in the internal representation space of LALMs. For fair comparison, all methods follow the same

Table 2: Controlled noisy-data baselines and clean-capability retention on 3,600 paired *Qwen2.5-Omni* examples at SNR $= -10$ dB. All values are percentages except $\Delta_{\text{clean}}$, which is measured in percentage points. Invalid or unparsable outputs are counted as incorrect. Gap closed denotes the fraction of the Initial model's clean-to-noisy accuracy gap recovered by each method.

| **Method** | **Noisy-Acc.** | **Clean-Acc.** | $\boldsymbol{\Delta}_{\text{clean}}$ | **Gap-closed** |
|---|---|---|---|---|
| Initial | 59.33 | 76.56 | – | 0.00 |
| Noisy-SFT | 60.31 | 76.56 | +0.00 | 5.65 |
| NoiseAug-SFT | 60.00 | 76.56 | +0.00 | 3.87 |
| `EchoDistill` | **62.94** | **77.56** | **+1.00** | **20.97** |

Table 3: Single-run component analysis on *Qwen2.5-Omni* at SNR $= -10$ dB. Current/old denotes the conventional GRPO policy ratio, while noisy/clean denotes EchoDistill's student–teacher ratio. `Acc` is the primary metric; `Noisy` and `GSR` are auxiliary agreement diagnostics.

| Variant | Policy ratio | Teacher objective | Reward bonus | Gate | Acc | Noisy | GSR |
|---|---|---|---|---|---|---|---|
| GRPO-only | Current/old | – | – | – | 59.94 | 71.47 | 71.61 |
| Distill-only | – | Token | – | – | 61.15 | 73.27 | 74.19 |
| Standard GRPO+KL | Current/old | KL | – | – | 61.50 | 70.44 | 74.31 |
| EchoDistill-StdRatio | Current/old | Token | ✓ | ✓ | 60.97 | 73.58 | 74.22 |
| EchoDistill-Ungated | Noisy/clean | Token | ✓ | – | 61.03 | 73.36 | 74.17 |
| **EchoDistill** | Noisy/clean | Token | ✓ | ✓ | **62.94** | **75.36** | **76.14** |

answer extraction protocol. We report three metrics: `Acc` Hendrycks et al. (2020), `Noisy`Zhang et al. (2026b), and `GSR` Benesty et al. (2009). Further implementation and evaluation details are provided in Appendix A.2.

### 4.2 Effectiveness and Attribution (RQ1)

**Obs.❶: `EchoDistill` improves aggregate robustness across backbones and domains, although the gains are heterogeneous.** As shown in Table 1, `EchoDistill` achieves an average noisy-input accuracy of **61.44%** across the nine backbone–domain settings, outperforming the strongest signal- or representation-level baseline by **1.63** percentage points (pp). Compared with the strongest corresponding baseline on each backbone, the average gains are **0.04** pp on *Step-Audio-2-mini*, **3.43** pp on *Qwen2.5-Omni*, and **0.76** pp on *MiniCPM-o-2.6*. The improvement is most pronounced for *Qwen2.5-Omni* in the Sound domain, whereas signal-processing baselines remain competitive in Music. Moreover, `EchoDistill` improves the auxiliary `Noisy` and `GSR` metrics by **1.80** and **4.18** pp. These results demonstrate an aggregate improvement under severe additive noise, rather than a uniform advantage in every backbone–domain setting.

**Obs.❷: Single noisy-data SFT does not account for the improvement of `EchoDistill`.** As shown in Table 2, `EchoDistill` achieves a noisy-input accuracy of **62.94%**, compared with **60.31%** for Noisy-SFT and **60.00%** for NoiseAug-SFT. The corresponding paired advantages are **2.64** pp and **2.94** pp, respectively. Because the controlled baselines are also trained with noisy audio, these results indicate that exposure to corrupted inputs alone is insufficient to reproduce the improvement. The additional gain is consistent with the benefit of aligning noisy-input trajectories with clean-conditioned token distributions.

**Obs.❸: The robustness gain does not incur a clean-audio alignment tax on the evaluated backbone.** While noisy-input accuracy increases from **59.33%** to **62.94%**, clean-audio accuracy also increases from **76.56%** to **77.56%**, corresponding to a **1.00** pp gain. `EchoDistill` closes **20.97%** of the Initial model's clean-to-noisy accuracy gap without reducing its clean-audio performance.

Taken together, these controlled comparisons show that the improvement of `EchoDistill` is neither explained by noisy-data exposure alone nor obtained at the expense of clean-audio capability. We next examine which optimization and distillation components contribute to this advantage.

**Obs.❹: The full formulation achieves the strongest point estimates among the controlled optimization variants.** As shown in Table 3, `EchoDistill` achieves an average accuracy of

Table 5: Accuracy (%) under audio interventions on the same 1,200 *Qwen2.5-Omni* questions at −10 dB. The final three columns report matched-minus-control differences in percentage points; bracketed values on the second line are paired 95% bootstrap confidence intervals.

| Model | Matched | Random | Shuffled | Silence | Matched −Random | Matched −Shuffled | Matched −Silence |
|---|---|---|---|---|---|---|---|
| Initial | 57.33 | 52.00 | 54.00 | 55.00 | +5.33<br>[3.17, 7.58] | +3.33<br>[1.58, 5.08] | +2.33<br>[−0.42, 5.08] |
| **EchoDistill** | **63.33** | 56.92 | 60.25 | 59.33 | **+6.42**<br>[4.25, 8.58] | **+3.08**<br>[1.42, 4.75] | **+4.00**<br>[1.50, 6.50] |

**62.94%**, compared with **59.94%** for GRPO-only, **61.15%** for Distill-only, and **61.50%** for Standard GRPO+KL. It therefore exceeds the standard GRPO formulation with clean-teacher KL regularization by **1.46** pp. Replacing the noisy/clean student–teacher policy ratio with the conventional current/old ratio reduces accuracy to **60.97%**, while removing the task-positive gate yields **61.03%**.

### 4.3 Generalization and Limits (RQ2)

**Obs.❺: Gains concentrate under severe noise and transfer weakly to unseen additive noise.** In Figure 3, EchoDistill improves over Initial by 6.00 pp at −10 dB, but the gain falls to **2.08, 1.17, and 0.08** pp at −5, 0, and 5 dB, respectively, and becomes **-1.25** pp at 10 dB. None of its paired differences from Standard GRPO+KL has a **95%** confidence interval excluding zero. Thus, the benefit over Initial is concentrated at low SNR rather than uniform across noise levels.

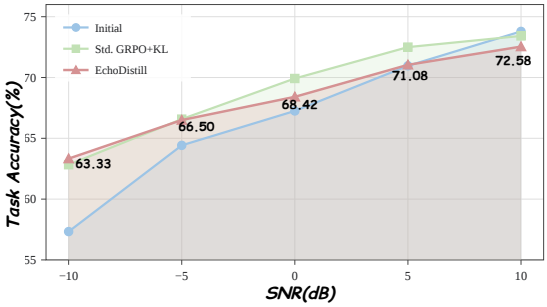


Figure 3: Task accuracy across SNR levels on *Qwen2.5-Omni*. All methods use the same noisy waveforms at each SNR ($N = 1{,}200$); annotations report `EchoDistill` experiment results.

The held-out categories show the same pattern (Appendix Figure 4): the gain is **+2.00** pp at −10 dB, but only **+0.81** pp when pooling all three SNRs. This provides suggestive evidence of transfer under severe unseen additive noise, not uniform out-of-distribution generalization.

Table 4: Cross-benchmark accuracy (%) at −10 dB. $\Delta_b$ denotes EchoDistill minus baseline $b$ in percentage points; paired 95% bootstrap CIs are reported below.

| Method | AIR-Bench | DREAM-TTS | MuChoMusic |
|---|---|---|---|
| Init. | 40.73 | 62.00 | 55.25 |
| GRPO+KL | 41.74 | 60.75 | 55.25 |
| **EchoDistill** | **42.65** | **66.50** | **57.50** |
| $\Delta_{\text{Init.}}$ | +1.92<br>[−0.25, 4.09] | +4.50<br>[0.75, 8.50] | +2.25<br>[−1.25, 5.75] |
| $\Delta_{\text{GRPO+KL}}$ | +0.92<br>[−1.09, 3.01] | +5.75<br>[2.50, 9.25] | +2.25<br>[−0.75, 5.25] |

**Obs.❻: Transfer is benchmark-dependent and does not reliably extend to the tested non-additive distortions.** Table 4 shows positive point differences over both Initial and Standard GRPO+KL on AIR-Bench Foundation (Yang et al., 2024), DREAM-TTS, and MuChoMusic (Wang et al., 2025). Only DREAM-TTS has intervals excluding zero: **+4.50** pp over Initial and **+5.75** pp over Standard GRPO+KL. Sample-weighted clean accuracy is nearly unchanged (**72.57% vs. 72.52%**), so the evidence supports benchmark-dependent rather than uniform transfer. The non-additive stress test further marks the method's boundary. Across 15 bandwidth, filtering, clipping, codec, and reverberation conditions, EchoDistill changes macro accuracy by **-0.47** pp, with lower point estimates in 10 conditions.

Table 6: Complementary acoustic-evidence checks in percentage points with 95% bootstrap CIs.

| Evaluation | $N$ | Metric / setting | Reference | Effect | 95% CI |
|---|---|---|---|---|---|
| AIR-Bench Chat | 300 | Matched accuracy | Random | +12.00 | [5.67, 18.67] |
| | | Matched accuracy | Silence | +13.67 | [7.00, 20.67] |
| ECAC | 470 | Critical accuracy | GRPO+KL | +4.26 | [1.91, 6.60] |
| | | Drop reduction | GRPO+KL | +2.98 | [0.21, 5.74] |
| | | Evidence preservation index | GRPO+KL | +3.19 | [0.85, 5.53] |
| | | Critical accuracy | Initial | −0.21 | [−2.55, 2.13] |
| Resampling | 600×2 | Realization 1 | GRPO+KL | +1.00 | [−2.00, 4.00] |
| | | Realization 2 | GRPO+KL | +1.33 | [−1.67, 4.17] |
| | | Pooled | GRPO+KL | +1.17 | [−1.33, 3.67] |

### 4.4 Dependence on Matched Acoustic Evidence (RQ3)

**Obs.❼: Predictions depend measurably on question-matched acoustic evidence.** As shown in Table 5, replacing matched audio with Random, Shuffled, or Silence reduces `EchoDistill` accuracy by **6.42, 3.08, and 4.00** points, and all paired intervals exclude zero. Because these effects use ground-truth correctness rather than GSR or clean–noisy agreement, an audio-independent explanation is insufficient. However, Initial also benefits from matched audio: its corresponding gaps are **5.33, 3.33, and 2.33** points. The increase in the matched-minus-control gap after training is therefore only **+1.09, −0.25, and +1.67** points. The controls establish that matched evidence contributes to `EchoDistill` predictions, but not that stronger acoustic dependence explains its entire post-training gain; mismatched audio may itself be harmful.

**Obs.❽: The auxiliary controls support a qualified grounding claim and expose its boundary.** Table 6 shows that, on open-ended AIR-Bench Chat, matched audio improves judged correctness over Random and Silence by **12.00 and 13.67** points, reducing the likelihood that the main effect is solely an answer-position shortcut. Under ECAC, EchoDistill exceeds Standard GRPO+KL by **4.26** points on the Critical condition, reduces the clean-to-Critical drop by **2.98** points, and improves EPI by **3.19** points. Its Critical accuracy is, however, statistically indistinguishable from Initial (−0.21 points), so ECAC supports robustness over the controlled optimizer rather than universally stronger grounding. Table 7 further shows that, when the clean teacher is wrong, `EchoDistill` recovers the correct answer more often than it imitates the teacher (**42.11% vs. 34.15%**), while the task-positive gate blocks **97.79%** of verified-wrong teacher signals. Finally, both new noise realizations retain positive point differences, but the pooled interval for the **+1.17-point** gain includes zero. Thus, matched evidence and selective teacher guidance contribute to the learned behavior. Full transition counts, ECAC results, option-rotation results, and teacher-confidence breakdowns are reported in the Appendix D.

Table 7: Teacher-error and task-positive-gate audit at −10 dB on 3,596 paired examples. WT imitation denotes agreement with an incorrect teacher; WT recovery denotes accuracy when the teacher is wrong. Blocked denotes verified-wrong teacher signals excluded from the reward.

| Method | Acc. (%) | WT imit. | WT recover | Wrong blocked |
|---|---|---|---|---|
| Initial | 59.40 | 33.87 | 39.80 | – |
| GRPO+KL | 61.54 | 32.68 | **43.30** | 98.01 |
| **EchoDistill** | **62.99** | 34.15 | 42.11 | 97.79 |

## 5 Conclusion

In this work, we propose `EchoDistill`, a noisy-to-clean self-distillation framework for improving LALM robustness under severe acoustic noise. By combining a frozen clean-audio teacher with noisy student rollouts, token-level alignment, and audio-aware reward shaping, `EchoDistill` improves noisy-input accuracy and generation stability without degrading clean-audio capability or adding inference-time costs. Experiments across three LALMs, together with controlled baselines and audio interventions, demonstrate the effectiveness of noisy-to-clean semantic alignment and confirm that matched acoustic evidence contributes to model predictions.

## AI USE STATEMENT

We used AI tools to assist with language polishing and manuscript editing. The authors remain fully responsible for the technical content, experimental results, analyses, citations, and claims presented in this paper.

## ETHICS STATEMENT

`EchoDistill` uses existing audio datasets and pretrained models to study the noise robustness of large audio language models. Model predictions and clean-conditioned teacher signals may contain errors or reflect biases in the underlying data and models. Applications involving personal audio should respect data licenses, consent, and privacy, and should include safeguards against unsupported or sensitive inferences about individuals.

## REPRODUCIBILITY STATEMENT

The model roles, datasets, baselines, and evaluation settings are described in Section 4. The training objective and inference procedure are detailed in Section 3. Additional information on data construction, metric definitions, statistical analysis, controlled audio interventions, and supplementary results is provided in Appendices A.2–D.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Keyu An, Yanni Chen, Zhigao Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Bo Gong, Xiangang Li, Yabin Li, et al. Fun-asr technical report. *arXiv preprint arXiv:2509.12508*, 2025.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.

Yanxi Chen, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Xin Li, Peijie Qiu, Hao Wang, Xuanzhao Dong, Yujian Xiong, Anderson Schneider, et al. Aha: Aligning large audio-language models for reasoning hallucinations via counterfactual hard negatives. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29294–29306, 2026.

Yuanjian Chen, Yang Xiao, Han Yin, Yadong Guan, and Xubo Liu. Noise-robust sound event detection and counting via language-queried sound separation. *IEEE Signal Processing Letters*, 2025.

Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Bai Jionghao, Zehan Wang, Shengpeng Ji, Rongjie Huang, et al. Omnichat: Enhancing spoken dialogue systems with scalable synthetic data for diverse scenarios. *arXiv preprint arXiv:2501.01384*, 2025.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Yexing Du, Ziyang Ma, Yifan Yang, Keqi Deng, Xie Chen, Bo Yang, Yang Xiang, Ming Liu, and Bing Qin. Cot-st: Enhancing llm-based speech translation with multimodal chain-of-thought. *arXiv preprint arXiv:2409.19510*, 2024.

Yang Fan, Xuemei Li, Jiannan Xiang, Xiang Li, Shuzheng Si, Xiaoyang Wang, Zhipeng Chen, Wenjie Li, Jiangwei Li, Jianghao Li, Hao Zhou, Jun Wang, Xinglin Yang, Xudong Wang, Zhicheng Zhang, Lei Li, Yanghan Wu, Wenhua Liu, Jian Yang, Jinze Bai, Xiang Li, Ruiqi Zhang, Yiheng Zhang, Changxing Li, Dongjie Ma, Yang Li, Hongzhuo Zhou, Xingwang Hu, Zonghan Yang, Jinghui Wang, Jian Zhang, Jiaming Liu, Zhichao Lu, Guangwei Song, Kun Liang, Xiaoyuan Zhang, Zhenyu Xu, Yang Song, Jianing Wang, Xudong Zhou, Xiaohe Li, Wenfeng Shang, Wei Wu, Jiaqi Wang, Jianwei Yang, Jiayi Wu, Xianbin Ye, Tong Zhang, Bin Lu, Xinjiang Wang, Lei Zhang, Zhigang Yuan, Yuanchun Shen, Ji Zhang, Xuanjing Shen, Hongyi Lv, Jingyang Lin, Han Yang, Jiarong Lu, Bowen Zhang, Xin Li, Jianhui Li, Yu Qiao, Yufeng Ding, Wenhao Zhang, Hao Lu, Jize Liu, Xingdong Ma, Kai Chen, Jiayan Sun, Tianyu Huang, Ruijie Song, Xinyu Li, Xin Hong, Zhihua Wu, Jun Wu, Zelin Wu, Wenjuan Han, Yongqiang Chen, Xiaodan Zhu, Jun Huang, and Yang Yang. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025. URL https://arxiv.org/abs/2503.20215.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, et al. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Guanyu Hou, Jiaming He, Yinhang Zhou, Ji Guo, Yitong Qiao, Rui Zhang, and Wenbo Jiang. Evaluating robustness of large audio language models to audio injection: An empirical study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25671–25687, 2025.

Guoning Hu and DeLiang Wang. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2067–2079, 2010.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23802–23804, 2024.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*, 2023.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Interspeech*, volume 2015, pp. 3586, 2015.

Pankaj Kumar and Subhankar Mishra. Robustness in large language models: A survey of mitigation strategies and evaluation metrics. *arXiv preprint arXiv:2505.18658*, 2025.

Bohan Li, Wenbin Huang, Yuhang Qiu, Yiwei Guo, Hankun Wang, Zhihan Li, Jing Peng, Ziyang Ma, Xie Chen, and Kai Yu. Isa-bench: Benchmarking instruction sensitivity for large audio language models. *arXiv preprint arXiv:2510.23558*, 2025.

Jinyu Li, Li Deng, Reinhold Haeb-Umbach, and Yifan Gong. Robust automatic speech recognition: a bridge to practical applications. 2015.

Bin Lin, Bo Zhao, Boyang Zhang, Boyong Wu, Chao Yan, Chen Geng, Chen Wu, Cheng Yi, Chengli Feng, Chenglin Zhu, et al. Stepaudio 3 realtime technical report. *arXiv preprint arXiv:2609.14005*, 2026a.

Liang Lin, Miao Yu, Kaiwen Luo, Yibo Zhang, Lilan Peng, Dexian Wang, Xuehai Tang, Yuanhe Zhang, Xikang Yang, Zhenhong Zhou, et al. Hidden in the noise: Unveiling backdoors in audio llms alignment through latent acoustic pattern triggers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32015–32023, 2026b.

Fernando López, Santosh Kesiraju, and Jordi Luque. Robustness assessment of large audio language models in multiple-choice evaluation. *arXiv preprint arXiv:2510.04584*, 2025.

Kaiwen Luo, Liang Lin, Yibo Zhang, Moayad Aloqaily, Jialiang Tao, Dexian Wang, Zhenhong Zhou, Junwei Zhang, Kun Wang, Li Sun, et al. Chronosaudio: A comprehensive long-audio benchmark for evaluating audio-large language models. *arXiv preprint arXiv:2601.04876*, 2026a.

Kaiwen Luo, Zhenhong Zhou, Leyan Wang, Liang Lin, Tianyu Shao, Yuanhe Zhang, Yang Xiao, Yuxuan Li, Miao Yu, Kailin Lyu, et al. A survey of large audio language models: Generalization, trustworthiness, and outlook. *arXiv preprint arXiv:2605.20266*, 2026b.

Ziyang Ma, Xiquan Li, Yakun Song, Wenxi Chen, Chenpeng Du, Jian Wu, Yuanzhe Chen, Zhuo Chen, Yuping Wang, Yuxuan Wang, et al. Towards reliable large audio language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1000–1014, 2025a.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025b.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13 (2):206–219, 2019.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Yuda Song, Lili Chen, Fahim Tajwar, Remi Munos, Deepak Pathak, J Andrew Bagnell, Aarti Singh, and Andrea Zanette. Expanding the capabilities of reinforcement learning via text feedback. *arXiv preprint arXiv:2602.02482*, 2026a.

Yuhan Song, Linhao Zhang, Chuhan Wu, Aiwei Liu, Wei Jia, Houfeng Wang, and Zhou Xiao. Stabletoken: A noise-robust semantic speech tokenizer for resilient speechllms. In *International Conference on Learning Representations*, volume 2026, pp. 49543–49569, 2026b.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025.

DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM transactions on audio, speech, and language processing*, 26(10):1702–1726, 2018.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

LCT Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR, abs/2512.23808*, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026a.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Han Yin, Yang Xiao, Jisheng Bai, and Rohan Kumar Das. Leveraging llm and text-queried separation for noise-robust sound event detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Satellite Workshop on Speech and Audio Language Models (SALMA)*, 2024.

Han Yin, Yang Xiao, Younghoo Kwon, Ting Dang, and Jung-Woo Choi. Focus then listen: Exploring plug-and-play audio enhancer for noise-robust large audio language models. *arXiv preprint arXiv:2603.04862*, 2026.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A standalone speech llm without codec injection for full-duplex conversation. *Advances in Neural Information Processing Systems*, 38: 24617–24643, 2026.

Yibo Zhang, Kaiwen Luo, Liang Lin, Shilinlu Yan, Jin Wang, Yaoqi Guo, Yitian Chen, Yalan Qin, Zhenhong Zhou, Kun Wang, et al. Rsa-bench: Benchmarking audio large models in real-world acoustic scenarios. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18350–18374, 2026a.

Yuanhe Zhang, Jiayu Tian, Yibo Zhang, Shilinlu Yan, Liang Lin, Zhenhong Zhou, Li Sun, and Sen Su. See: Signal embedding energy for quantifying noise interference in large audio language models. *arXiv preprint arXiv:2601.07331*, 2026b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

# A Appendix

## A.1 Extended Related Work

**Group Relative Policy Optimization.** Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Song et al., 2026a) samples a group of responses $\{y^{(k)}\}_{k=1}^{K}$ and normalizes their rewards within the group:

$$A^{(k)} = \frac{r^{(k)} - \mu}{\sigma + \varepsilon}, \tag{14}$$

where $\mu$ and $\sigma$ are the group reward mean and standard deviation. For

$$\rho_t^{(k)} = \frac{\pi_\theta(y_t^{(k)} \mid x, y_{<t}^{(k)})}{\pi_{\theta_{\text{old}}}(y_t^{(k)} \mid x, y_{<t}^{(k)})}, \tag{15}$$

the clipped objective is

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|y^{(k)}|}\sum_{t=1}^{|y^{(k)}|}\min\Big(\rho_t^{(k)}A^{(k)}, \text{clip}(\rho_t^{(k)}, 1-\epsilon, 1+\epsilon)A^{(k)}\Big). \tag{16}$$

GRPO avoids training a separate value model, but its sequence-level advantage is shared across tokens and therefore does not identify which token decisions are altered by acoustic corruption.

**Robust speech processing.** Robust speech systems have long addressed acoustic mismatch through multi-condition training, waveform augmentation, spectral masking, speech enhancement, and source separation (Li et al., 2015; Ko et al., 2015; Park et al., 2019; Pascual et al., 2017; Wang & Chen, 2018). These methods are effective when improved signal fidelity or recognition accuracy is closely coupled with the downstream objective. However, LALMs must map acoustic observations into open-ended semantic responses, so recovering a perceptually cleaner signal does not necessarily preserve the specific evidence required by a downstream question. Moreover, aggressive enhancement may suppress weak but discriminative acoustic cues when those cues overlap with interference (Purwins et al., 2019).

**Noise robustness for LALMs.** Acoustic corruption can change the high-level predictions of LALMs even when their generated language remains fluent (Li et al., 2025; Hou et al., 2025; Wang et al., 2025). Recent methods use text-queried separation, plug-and-play audio enhancement, and representation-level suppression to reduce this sensitivity (Yin et al., 2024; Chen et al., 2025; Yin et al., 2026; Zhang et al., 2026b). These approaches mainly intervene in the waveform, encoder features, or internal noise representations. Such intervention can be complementary to `EchoDistill`, but it does not explicitly optimize the semantic decisions made along noisy-input generation trajectories. `EchoDistill` instead retains the original input and model architecture at inference time and uses paired clean audio only to construct privileged post-training supervision.

**Knowledge distillation and on-policy learning.** Classical knowledge distillation matches softened predictions or internal representations from a teacher (Hinton et al., 2015). For autoregressive models, offline teacher forcing can expose the student to prefixes that differ from those encountered during its own generation, creating a distribution mismatch related to exposure bias (Ross et al., 2011). On-policy distillation addresses this issue by evaluating teacher guidance on student-generated trajectories (Agarwal et al., 2024). Recent studies further explore self-distillation between checkpoints or different states of the same model (Zhao et al., 2026; Hübotter et al., 2026; Shenfeld et al., 2026; Ko et al., 2026). `EchoDistill` combines on-policy guidance with asymmetric acoustic information: the clean and noisy branches evaluate the same student-generated prefix but receive different versions of the audio. This paired comparison estimates where corruption changes the model's next-token preference. Task-positive gating further prevents agreement with an incorrect clean-conditioned prediction from being rewarded indiscriminately.

## A.2 Experimental Protocol and Metrics

**Models and paired data.** We evaluate Qwen2.5-Omni, MiniCPM-o-2.6, and Step-Audio2 (Fan et al., 2025; Yao et al., 2024; Wu et al., 2025). For each backbone, the teacher and student are initialized from the same checkpoint, after which the teacher is frozen. Training and validation contain

14,397 and 1,704 paired clean/noisy examples from MMAR (Ma et al., 2025b). Training corruptions comprise ten additive-noise categories and seven SNR levels, $\{-10, -5, 0, 5, 10, 20, 30\}$ dB. The primary evaluation contains 3,600 MMAU examples (Sakshi et al., 2025), equally divided among Music, Sound, and Speech. Unless stated otherwise, results use additive noise at $-10$ dB.

**Primary and auxiliary metrics.** Noisy-input task accuracy is the primary metric. For a fixed evaluation set of size $N$, invalid or unparsable answers are counted as incorrect:

$$\mathrm{Acc}_d^{\mathrm{n}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i^{d,\mathrm{n}} = y_i^{\star}\right]. \tag{17}$$

Clean-audio accuracy $\mathrm{Acc}_d^{\mathrm{c}}$ is defined analogously. Let $d = 0$ denote the initial model. We report

$$\Delta_{\mathrm{clean}}(d) = \mathrm{Acc}_d^{\mathrm{c}} - \mathrm{Acc}_0^{\mathrm{c}}, \tag{18}$$

$$\mathrm{GapClosure}(d) = \frac{\mathrm{Acc}_d^{\mathrm{n}} - \mathrm{Acc}_0^{\mathrm{n}}}{\mathrm{Acc}_0^{\mathrm{c}} - \mathrm{Acc}_0^{\mathrm{n}}} \times 100\%. \tag{19}$$

For comparison with prior work, we retain `Noisy` and `GSR` as auxiliary agreement diagnostics. Let $\hat{y}_i^{0,\mathrm{c}}$ be the initial model's clean-audio prediction and let $\mathcal{V}_{\mathrm{Noisy}}$ and $\mathcal{V}_{\mathrm{GSR}}$ denote the corresponding valid paired-output sets. We compute

$$\mathrm{Noisy}_d = \frac{1}{|\mathcal{V}_{\mathrm{Noisy}}|} \sum_{i \in \mathcal{V}_{\mathrm{Noisy}}} \mathbb{I}\left[\hat{y}_i^{d,\mathrm{n}} = \hat{y}_i^{0,\mathrm{c}}\right], \tag{20}$$

$$\mathrm{GSR}_d = \frac{1}{|\mathcal{V}_{\mathrm{GSR}}|} \sum_{i \in \mathcal{V}_{\mathrm{GSR}}} \mathbb{I}\left[\hat{y}_i^{d,\mathrm{n}} = \hat{y}_i^{d,\mathrm{c}}\right]. \tag{21}$$

These diagnostics quantify agreement with clean-conditioned behavior. They are not correctness or grounding metrics: predictions may agree while both are wrong, and an input-independent model may also obtain high agreement.

**Statistical analysis.** Unless otherwise stated, intervals are paired 95% bootstrap confidence intervals based on 5,000 resamples with seed 42. When several corruptions are derived from the same clean source, resampling is clustered by source. Corrected/broken transitions are additionally evaluated with the exact McNemar test. The independent noise-resampling experiment uses 10,000 paired cluster-bootstrap resamples. These intervals capture evaluation-sample uncertainty conditional on one trained checkpoint and the evaluated waveforms; they do not capture post-training seed variability.

**Training-instance format.** Each record contains a paired waveform, the task prompt, candidate choices, the target, and corruption metadata. Student candidates, teacher scores, and rewards are produced online rather than stored as fixed targets.

**Cross-table denominators.** Table 1 reports 62.96% because its Average row is the macro-average of the three domain-level accuracies. Tables 2, 3, and 8 instead report pooled accuracy on the same fixed set of 3,600 examples, yielding 62.94%, with invalid or unparsable outputs counted as incorrect. Table 13 reports 62.99% because the teacher-reliability audit is restricted to 3,596 eligible examples. These small differences arise from aggregation and evaluation eligibility rather than different model checkpoints.

### A.3 Limitations

`EchoDistill` relies on paired clean–noisy audio and an imperfect clean-conditioned teacher, while noisy rollouts and token-level scoring increase post-training cost. Its gains are strongest under severe additive noise and do not reliably extend to the tested non-additive distortions. Moreover, several analyses use a single trained checkpoint, so future work should examine seed variability, broader acoustic conditions, and more natural open-ended audio interactions.

### A.4 ILLUSTRATIVE TRAINING RECORD

To qualitatively examine whether the model grounds its predictions in the acoustic input rather than relying primarily on linguistic priors, we present three representative case studies. Case Study 1 illustrates the structure of a noisy-audio training record. Case Studies 2 and 3 compare the responses of the initial model, standard GRPO+KL, and `EchoDistill` under the same noisy input. We additionally replace the relevant waveform with random audio or remove it entirely. These controlled variants reveal whether a correct response is supported by the acoustic evidence: `EchoDistill` recovers the reference answer from the noisy waveform, whereas its predictions change when the audio is irrelevant or unavailable.

The examples were manually selected for interpretability and should not be viewed as an unbiased sample of model successes or failures.

**Case Study 1**

**Example ID: 19452**

***Prompt:*** What is producing the sound in the audio?

***Choices:*** Airplane, Motorcycle, Train, Sports car

***Acoustic condition:*** Water noise at an SNR of 30 dB.

***Target:*** **Airplane**

**Training Record (Raw Input Format):**


```
{
  "id": 19452,
  "prompt": "What is producing the sound in the audio?",
  "noisy_audio_path": ".../noise/water/snr_30/...wav",
  "clean_audio_path": ".../audio/...wav",
  "choices": ["Airplane", "Motorcycle", "Train", "Sports car"],
  "target": "Airplane",
  "noise_type": "water",
  "snr": 30
}
```


*Paths are abbreviated for readability. The target field provides the supervision signal used during training.*

**Case Study 2 (1/2)**

**Example ID: 1907**

***Question:*** In which show did the male speaker mention a guest appearance was made?

***Reference answer:*** Little Britain.

**Answer❶: Initial (Noisy Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (noisy audio tokens) ...AUDIOAUDIOaudio_eos
In which show did the male speaker mention a guest appearance was
made?  im_end
im_startassistant
", "predict":  "The male speaker mentioned that a guest appeared on
the show 'The Tonight Show Starring Jimmy Fallon'." }
```

**Answer❷: Standard GRPO+KL (Noisy Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (noisy audio tokens) ...AUDIOAUDIOaudio_eos
In which show did the male speaker mention a guest appearance was
made?  im_end
im_startassistant
", "predict":  "The male speaker mentioned a guest appearance was
made in the show 'The X-Files'." }
```

**Answer❸: `EchoDistill` (Noisy Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (noisy audio tokens) ...AUDIOAUDIOaudio_eos
In which show did the male speaker mention a guest appearance was
made?  im_end
im_startassistant
", "predict":  "The male speaker mentioned a guest appearance was
made in Little Britain." }
```

**Case Study 2 (2/2)**

**Example ID: 1907**

***Question:*** In which show did the male speaker mention a guest appearance was made?

***Reference answer:*** Little Britain.

**Answer❹: `EchoDistill` (Random Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (random audio tokens) ...AUDIOAUDIOaudio_eos
In which show did the male speaker mention a guest appearance was
made?  im_end
im_startassistant
", "predict":  "The male speaker mentioned a guest appearance was
made in the show 'The Best of the Best'." }
```

**Answer❺: `EchoDistill` (No Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
In which show did the male speaker mention a guest appearance was
made?  im_end
im_startassistant
", "predict":  "I'm sorry, but I cannot answer this question
without more context.  Can you please provide me with more
information about the male speaker and the show in question?" }
```

**Takeaway.** Under the matched noisy waveform, `EchoDistill` recovers the reference *Little Britain*, whereas replacing or removing the waveform changes the answer. This example illustrates sensitivity to question-matched acoustic evidence, but does not by itself establish a population-level effect.

**Case Study 3 (1/2)**

**Example ID: 97**

***Question:*** What is the speaker's final statement about the future of peace in the area?

***Reference answer:*** The speaker's final statement is that there's never going to be peace over there.

**Answer❶: Initial (Noisy Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (noisy audio tokens) ...AUDIOAUDIOaudio_eos
What is the speaker's final statement about the future of peace in
the area?  im_end
im_startassistant
", "predict":  "The speaker believes that peace will always be a
possibility in the area." }
```

**Answer❷: Standard GRPO+KL (Noisy Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (noisy audio tokens) ...AUDIOAUDIOaudio_eos
What is the speaker's final statement about the future of peace in
the area?  im_end
im_startassistant
", "predict":  "The speaker believes that peace will always be a
possibility in the area." }
```

**Answer❸: `EchoDistill` (Noisy Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (noisy audio tokens) ...AUDIOAUDIOaudio_eos
What is the speaker's final statement about the future of peace in
the area?  im_end
im_startassistant
", "predict":  "The speaker believes that peace will never be
achieved in the area." }
```

**Case Study 3 (2/2)**

**Example ID: 97**

***Question:*** What is the speaker's final statement about the future of peace in the area?

***Reference answer:*** The speaker's final statement is that there's never going to be peace over there.

**Answer❹: `EchoDistill` (Random Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
audio_bosAUDIOAUDIO...  (random audio tokens) ...AUDIOAUDIOaudio_eos
What is the speaker's final statement about the future of peace in
the area?  im_end
im_startassistant
", "predict":  "The speaker believes that peace will be achieved in
the future." }
```

**Answer❺: `EchoDistill` (No Audio)**

```
{ "prompt":  "im_startsystem
You are a helpful assistant.im_end
im_startuser
What is the speaker's final statement about the future of peace in
the area?  im_end
im_startassistant
", "predict":  "The speaker's final statement about the future of
peace in the area is that it is uncertain and depends on the
actions of the people living there." }
```

**Takeaway.** The matched noisy waveform allows `EchoDistill` to preserve the negative meaning of the speaker's statement, while the Random-Audio and No-Audio conditions reverse or weaken that conclusion. The example is illustrative; the aggregate grounding claim is based on the paired controls in Table 11.

Table 8: Fixed-denominator Qwen2.5-Omni audit on 3,600 paired examples. Accuracy and gap closure are percentages; gains are percentage points. Corrected/Broken counts paired noisy-input transitions.

| Input | Initial Acc. | Echo Acc. | Gain | Gap Closure | 95% CI | Corrected/Broken |
|---|---|---|---|---|---|---|
| Noisy (−10 dB) | 59.33 | **62.94** | +3.61 | 20.97 | [2.44, 4.81] | 297/167 |
| Clean | 76.56 | **77.56** | +1.00 | – | [0.33, 1.67] | – |

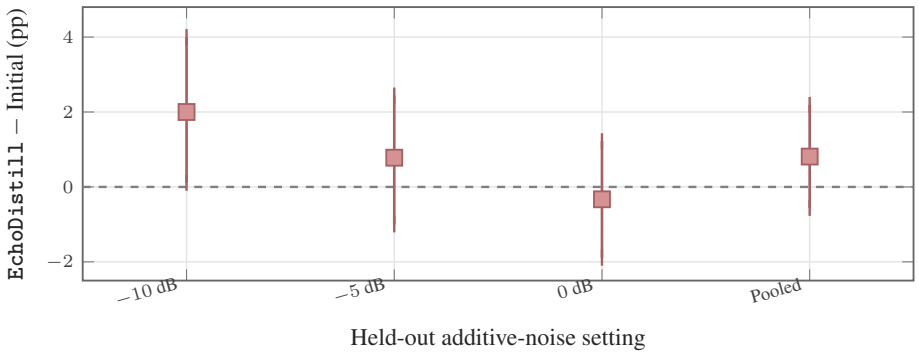


Figure 4: Transfer to six additive-noise categories excluded from training on Qwen2.5-Omni. Error bars indicate paired 95% cluster-bootstrap confidence intervals.

## B Additional Results for RQ1

**Fixed-denominator transition audit.** Table 8 examines paired correctness transitions on the same 3,600 Qwen2.5-Omni examples used for the clean-capability comparison. Invalid or unparsable outputs count as incorrect for both models. Under −10 dB additive noise, `EchoDistill` corrects **297** errors made by Initial while breaking **167** initially correct predictions. The resulting net improvement is **+3.61 percentage points** (95% CI: $[2.44, 4.81]$), recovering **20.97%** of Initial's clean-to-noisy accuracy gap. On the paired clean inputs, accuracy increases from **76.56%** to **77.56%** (95% CI for the difference: $[0.33, 1.67]$). Thus, the noisy-input improvement is not accompanied by a clean-audio accuracy loss on this evaluation. This clean-retention result applies to Qwen2.5-Omni; equivalent fixed-denominator audits are not available for the other two backbones.

## C Additional Results for RQ2

### C.1 Held-Out Additive-Noise Categories

Figure 4 and Table 9 test six additive-noise categories excluded from training. The two models receive identical, reproducibly mixed waveforms, so each comparison is paired at the source-task level. At −10 dB, `EchoDistill` improves accuracy from **64.78%** to **66.78%**: a **+2.00-point** difference with a paired 95% cluster-bootstrap CI of $[0.11, 4.00]$. The exact McNemar test yields $p = 0.0535$, however, so the two uncertainty analyses should not be summarized as an unqualified significant improvement.

The advantage becomes smaller as the corruption weakens: **+0.78 points** at −5 dB and −0.33 **points** at 0 dB. Across all three SNRs, the pooled difference is **+0.81 points** (95% CI: $[-0.56, 2.19]$). Although the six category-level point estimates at −10 dB are nonnegative, their individual intervals include zero. These results suggest transfer to unseen additive noise under the most severe tested corruption, but do not establish a uniform benefit across noise categories or intensities. None of the six category-specific confidence intervals excludes zero; the category-level breakdown is therefore descriptive and should not be used to rank the relative transferability of individual noise types.

Table 9: Held-out additive-noise evaluation on 900 source tasks. The pooled row contains three SNR observations per task and clusters its confidence interval by source. C/B denotes corrected/broken predictions.

| Setting | $N$ | Initial | `EchoDistill` | Gain | 95% CI | C/B | McNemar $p$ |
|---|---|---|---|---|---|---|---|
| *Aggregate by SNR* | | | | | | | |
| $-10$ dB | 900 | 64.78 | **66.78** | +2.00 | [0.11, 4.00] | 48/30 | 0.0535 |
| $-5$ dB | 900 | 67.89 | 68.67 | +0.78 | [$-1.00$, 2.44] | 36/29 | 0.4570 |
| 0 dB | 900 | 71.00 | 70.67 | $-0.33$ | [$-1.89$, 1.22] | 24/27 | 0.7798 |
| Pooled | 2,700 | 67.89 | 68.70 | +0.81 | [$-0.56$, 2.19] | 108/86 | 0.1314 |
| *Breakdown at* $-10$ dB | | | | | | | |
| Clock tick | 150 | 62.67 | 66.67 | +4.00 | [$-1.33$, 9.33] | 11/5 | 0.2101 |
| Crackling fire | 150 | 68.67 | 70.00 | +1.33 | [$-3.33$, 6.00] | 7/5 | 0.7744 |
| Keyboard typing | 150 | 66.67 | 70.00 | +3.33 | [$-0.67$, 8.00] | 8/3 | 0.2266 |
| Rain | 150 | 56.67 | 59.33 | +2.67 | [$-2.67$, 8.00] | 10/6 | 0.4545 |
| Sea waves | 150 | 62.67 | 62.67 | 0.00 | [$-4.67$, 4.67] | 7/7 | 1.0000 |
| Thunderstorm | 150 | 71.33 | 72.00 | +0.67 | [$-3.33$, 4.67] | 5/4 | 1.0000 |

Table 10: Non-additive stress test on 300 clean source tasks per condition. Values are accuracy percentages or `EchoDistill`-minus-Initial percentage points. Positive drop reduction favors `EchoDistill`.

| | Mean accuracy | | Δ by severity | | | Mean robustness change | |
|---|---|---|---|---|---|---|---|
| Distortion | Initial | Echo | Mild | Medium | Severe | ΔAcc | Drop reduction |
| Clean reference | 76.33 | 76.00 | – | – | – | $-0.33$ | – |
| Bandwidth | 75.44 | 74.44 | $-0.67$ | $-0.67$ | $-1.67$ | $-1.00$ | $-0.67$ |
| Channel filter | 75.78 | 75.33 | $-0.33$ | $-0.67$ | $-0.33$ | $-0.44$ | $-0.11$ |
| Clipping | 76.44 | 76.33 | $-0.33$ | $-0.67$ | +0.67 | $-0.11$ | +0.22 |
| Codec | 73.33 | 72.89 | $-1.67$ | 0.00 | +0.33 | $-0.44$ | $-0.11$ |
| Reverberation | 75.78 | 75.44 | 0.00 | 0.00 | $-1.00$ | $-0.33$ | 0.00 |
| **Macro over 15 conditions** | **75.36** | **74.89** | 2 improved / 3 tied / 10 lower | | | **$-0.47$** | **$-0.13$** |

Macro ΔAcc 95% cluster-bootstrap CI: [$-2.24$, 1.31].

### C.2 Non-Additive Robustness Boundary

Table 10 evaluates 15 non-additive conditions formed from five distortion families and three severity levels. Unlike the held-out additive-noise experiment, these conditions change the signal through bandwidth restriction, filtering, clipping, codec processing, or reverberation. `EchoDistill` has lower accuracy point estimates in **10** conditions, ties in **3**, and improves in **2**. Its macro accuracy difference relative to Initial is $-0.47$ **points** (95% cluster-bootstrap CI: [$-2.24$, 1.31]). The interval does not establish systematic harm, but neither does this stress test support reliable transfer to non-additive distortions. We therefore treat robustness to these transformations as a boundary of the present training setup rather than extending the additive-noise claim to all acoustic perturbations.

## D Additional Results for RQ3

Clean–noisy agreement alone cannot establish reliance on acoustic evidence: a model may answer consistently from the question or answer choices. We therefore hold the prompt, choices, and decoding procedure fixed while changing only the audio. The main intervention study uses 1,200 Qwen2.5-Omni examples at $-10$ dB and evaluates ground-truth task accuracy. Random supplies an unrelated waveform, Shuffled permutes the audio–question pairing, and Silence removes the waveform. These conditions assess whether the *matched* acoustic input contributes to a correct prediction, rather than treating agreement itself as a grounding measure.

Table 11: Paired audio-control transitions on the 1,200-example diagnostic set at $-10\,\mathrm{dB}$. Fixed/Broken compares Matched with Silence; Matched/Random win compares correctness under the two audio assignments. Gaps are percentage points with paired 95% bootstrap intervals.

| | Silence control | | | Random-audio control | | | Matched–Random gap | |
|---|---|---|---|---|---|---|---|---|
| Model | Fixed/Broken | Net | $p$ | Matched/Random win | Net | $p$ | Model-specific | Shared failed ($N = 468$) |
| Initial | 157/129 | +28 | 0.110 | 123/59 | +64 | $2.40\times10^{-6}$ | 10.00 [6.48, 13.33] | 10.68 [7.05, 14.32] |
| `EchoDistill` | 138/90 | +48 | 0.0018 | 129/52 | +77 | $9.75\times10^{-9}$ | 11.27 [7.79, 14.96] | 12.39 [8.76, 16.03] |
| `EchoDistill` − Initial | −19/ − 39 | +20 | – | +6/−7 | +13 | – | +1.27 | +1.71 |

Table 12: ECAC results on 470 preregistered eligible pairs. The upper table reports accuracy and changes from Initial. The lower table reports paired contrasts. Critical drop reduction is the comparator's clean-to-Critical drop minus `EchoDistill`'s; EPI is the Critical gain minus the Irrelevant gain.

**Panel A: Accuracy under different acoustic conditions**

| | Accuracy (change vs. Initial) | | | | |
|---|---|---|---|---|---|
| Method | Clean | Critical | Irrelevant | Global | Mean |
| Initial | 72.13 | 68.30 | 68.72 | 47.87 | 64.26 |
| Noisy-SFT | 68.09 (−4.04) | 64.89 (−3.40) | 64.89 (−3.83) | 44.89 (−2.98) | 60.69 (−3.56) |
| NoiseAug-SFT | 67.87 (−4.26) | 64.47 (−3.83) | 65.53 (−3.19) | 44.68 (−3.19) | 60.64 (−3.62) |
| Standard GRPO+KL | 68.94 (−3.19) | 63.83 (−4.47) | 65.74 (−2.98) | 47.23 (−0.64) | 61.44 (−2.82) |
| `EchoDistill`-StdRatio | 69.79 (−2.34) | 65.32 (−2.98) | 66.38 (−2.34) | 47.02 (−0.85) | 62.13 (−2.13) |
| `EchoDistill`-Ungated | 69.79 (−2.34) | 65.96 (−2.34) | 65.74 (−2.98) | 47.02 (−0.85) | 62.13 (−2.13) |
| `EchoDistill` | **70.21** (−1.91) | **68.09** (−0.21) | **66.81** (−1.91) | **49.36** (+1.49) | **63.62** (−0.64) |

**Panel B: Paired statistical contrasts**

| Contrast | Critical gain [95% CI] | Global gain | Drop reduction [95% CI] | EPI [95% CI] |
|---|---|---|---|---|
| `EchoDistill` − Initial | −0.21 [−2.55, 2.13] | +1.49 | +1.70 [−0.85, 4.47] | +1.70 [−0.64, 4.04] |
| `EchoDistill` − Standard GRPO+KL | **+4.26 [1.91, 6.60]** | **+2.13** | **+2.98 [0.21, 5.74]** | **+3.19 [0.85, 5.53]** |

### D.1 Paired Audio-Control Transitions

The main paper reports absolute accuracy under the four audio conditions. Table 11 additionally tracks paired changes in correctness. For `EchoDistill`, matched audio corrects **138** cases that are wrong under Silence and breaks **90**, a net difference of **48** cases ($p = 0.0018$). Relative to Random audio, matched audio wins **129** cases and loses **52** ($p = 9.75 \times 10^{-9}$). Initial also benefits from matched audio, so these tests demonstrate that `EchoDistill`'s predictions use relevant acoustic evidence; they do not establish that increased audio dependence, by itself, explains the full post-training gain. The model-specific gap is computed on each checkpoint's own subset of questions answered incorrectly under Silence. The shared-failed gap restricts both checkpoints to the same 468 questions for which their recorded Silence predictions are incorrect. Because these subsets are selected using observed Silence outcomes, both estimates are descriptive conditional quantities rather than unbiased population effects.

### D.2 Evidence-Critical Acoustic Corruption

The evidence-critical acoustic corruption (ECAC) test probes a more localized intervention. It compares matched-energy corruption of a model-localized critical window, an irrelevant window, or the full waveform. Table 12 reports results on 470 preregistered eligible pairs. In Panel A, `EchoDistill` attains **68.09%** accuracy under critical-window corruption, compared with **63.83%** for Standard GRPO+KL. Panel B quantifies this as a **+4.26-point** paired difference (95% CI: $[1.91, 6.60]$). `EchoDistill` also reduces the clean-to-Critical accuracy drop by **2.98 points** (95% CI: $[0.21, 5.74]$) relative to that optimizer and improves the evidence preservation index by **3.19 points** (95% CI: $[0.85, 5.53]$).

The comparison with Initial is more limited: `EchoDistill`'s Critical accuracy is $-0.21$ **points** different (95% CI: $[-2.55, 2.13]$), and its Clean accuracy in this subset is also below Initial's. Thus,

Table 13: Teacher-reliability audit on 3,596 paired examples at $-10\,\mathrm{dB}$. Correct/Wrong denotes teacher correctness, high/low denotes the audit's confidence partition, WT imitation is agreement with an incorrect teacher, and WT recovery is accuracy when that teacher is wrong. All values are percentages.

| Method | Overall | Correct-high | Correct-low | Wrong-high | Wrong-low | WT imitation | WT recovery |
|---|---|---|---|---|---|---|---|
| Initial | 59.40 | 79.50 | 67.17 | 42.09 | 39.16 | 33.87 | 39.80 |
| Noisy-SFT | 60.37 | 79.50 | 65.81 | 39.24 | 44.62 | 31.08 | 43.44 |
| NoiseAug-SFT | 60.07 | 79.50 | 65.50 | 38.61 | 44.18 | 31.15 | 42.95 |
| Standard GRPO+KL | 61.54 | 81.80 | 67.65 | 41.46 | 43.82 | 32.68 | **43.30** |
| `EchoDistill`-StdRatio | 61.04 | 82.46 | 69.01 | 40.82 | 40.32 | 35.06 | 40.43 |
| `EchoDistill`-Ungated | 61.10 | 82.57 | 68.93 | 40.51 | 40.59 | 34.85 | 40.57 |
| `EchoDistill` | **62.99** | **84.76** | **71.01** | **43.04** | 41.85 | 34.15 | 42.11 |

Table 14: Task-positive gate audit. All values are percentages. Verified-wrong is a conservative lower bound because unresolved free-form responses are not labeled incorrect.

| Method | Positive | Verified-wrong/positive | Wrong-teacher blocked |
|---|---|---|---|
| Standard GRPO+KL | 52.84 | 1.62 | 98.01 |
| `EchoDistill`-StdRatio | 54.16 | 2.50 | 97.28 |
| `EchoDistill` | **56.33** | 2.09 | 97.79 |

Table 12 supports improved resistance to evidence-critical corruption *relative to the controlled optimizer*; it does not show uniformly better acoustic grounding than the initial checkpoint.

For a comparator $b$, we define

$$\Delta_{\mathrm{crit}}(b) = \mathrm{Acc}_{\mathrm{crit}}(\texttt{EchoDistill}) - \mathrm{Acc}_{\mathrm{crit}}(b), \tag{22}$$

$$\Delta_{\mathrm{irr}}(b) = \mathrm{Acc}_{\mathrm{irr}}(\texttt{EchoDistill}) - \mathrm{Acc}_{\mathrm{irr}}(b), \tag{23}$$

$$\mathrm{EPI}(b) = \Delta_{\mathrm{crit}}(b) - \Delta_{\mathrm{irr}}(b), \tag{24}$$

$$\mathrm{DropRed.}(b) = [\mathrm{Acc}_{\mathrm{clean}}(b) - \mathrm{Acc}_{\mathrm{crit}}(b)] - [\mathrm{Acc}_{\mathrm{clean}}(\texttt{EchoDistill}) - \mathrm{Acc}_{\mathrm{crit}}(\texttt{EchoDistill})]\,. \tag{25}$$

Positive EPI and drop reduction indicate better preservation of evidence-critical information by `EchoDistill` relative to comparator $b$.

### D.3 Teacher Reliability and Reward Gating

The clean-conditioned teacher provides useful guidance but can also make incorrect predictions. Table 13 stratifies 3,596 paired examples by teacher correctness and the audit's confidence partition. `EchoDistill` reaches **62.99%** overall accuracy, compared with **61.54%** for Standard GRPO+KL. Its accuracy is higher in both correct-teacher partitions (**84.76%** versus **81.80%** for Correct-high, and **71.01%** versus **67.65%** for Correct-low). The pattern is less uniform when the teacher is wrong: `EchoDistill`'s wrong-teacher recovery is **42.11%**, below Standard GRPO+KL's **43.30%**. These descriptive strata therefore do not justify a claim that `EchoDistill` reliably corrects teacher errors.

Importantly, `EchoDistill`'s wrong-teacher imitation rate is **34.15%**, while its task accuracy on wrong-teacher examples is **42.11%**; it does not deterministically reproduce the teacher's answer. Table 14 further shows that the task-positive gate excludes **97.79%** of *verified-wrong* teacher signals from the reward. Because unresolved free-form answers are not certified as wrong, this is an audit of identifiable errors, not a guarantee that all erroneous teacher guidance is removed. The confidence partitions are likewise descriptive and should not be interpreted as calibrated selective-risk estimates.

### D.4 Open-Ended and Option-Position Controls

Table 15 extends the audio intervention to 300 AIR-Bench Chat examples evaluated by an audio-conditioned automatic judge. For `EchoDistill`, matched audio increases judged correctness over Random by **12.00 points** (95% CI: $[5.67, 18.67]$) and over Silence by **13.67 points** (95% CI:

Table 15: Auxiliary open-ended evaluation on 300 AIR-Bench Chat examples at $-10\,\mathrm{dB}$. Rates come from an audio-conditioned automatic judge. The last two columns are `EchoDistill`-minus-comparator paired differences with 95% bootstrap confidence intervals.

| Model | Valid | Noisy correct | Noisy relevant | Noisy UAC ↓ | Δ correct | Δ relevant |
|---|---|---|---|---|---|---|
| Initial | 100.00 | 36.00 | 48.00 | 17.00 | +2.00 [−4.33, 8.33] | +7.67 [1.33, 14.33] |
| Noisy-SFT | 100.00 | 35.00 | 48.33 | 16.33 | +3.00 [−3.00, 9.00] | +7.33 [1.00, 13.67] |
| NoiseAug-SFT | 100.00 | 35.00 | 48.33 | 17.67 | +3.00 [−3.33, 9.00] | +7.33 [0.67, 14.00] |
| Standard GRPO+KL | 100.00 | 34.33 | 48.00 | **11.33** | +3.67 [−2.33, 10.00] | +7.67 [0.67, 14.67] |
| `EchoDistill`-StdRatio | 100.00 | 34.33 | 47.33 | 12.67 | +3.67 [−2.33, 10.00] | +8.33 [1.33, 15.33] |
| `EchoDistill`-Ungated | 100.00 | 36.67 | 49.00 | 16.00 | +1.33 [−4.67, 7.33] | +6.67 [0.00, 13.33] |
| `EchoDistill` | 100.00 | **38.00** | **55.67** | **11.33** | – | – |

`EchoDistill` Matched−Random correctness: +12.00 [5.67, 18.67]; Matched−Silence: +13.67 [7.00, 20.67].

Table 16: Answer-option position audit at $-10\,\mathrm{dB}$ on 120 source questions under the original order and three cyclic rotations. Mean accuracy averages all orders; All-4 requires correctness in every order; Consistency requires the same answer content in every order.

| Model | Original | Rot. 1 | Rot. 2 | Rot. 3 | Mean Acc. | All-4 | Consistency | Valid |
|---|---|---|---|---|---|---|---|---|
| Initial | 69.17 | 65.00 | 70.83 | 65.83 | 67.71 | 53.33 | 62.50 | 100.00 |
| Noisy-SFT | 69.17 | 61.67 | 61.67 | 64.17 | 64.17 | 49.17 | 54.17 | 100.00 |
| NoiseAug-SFT | 70.00 | 60.83 | 62.50 | 64.17 | 64.38 | 49.17 | 53.33 | 100.00 |
| Standard GRPO+KL | 71.67 | 63.33 | 65.83 | 64.17 | 66.25 | 50.83 | 55.83 | 100.00 |
| `EchoDistill`-StdRatio | 69.17 | 64.17 | 70.00 | 66.67 | 67.50 | 52.50 | 60.83 | 100.00 |
| `EchoDistill`-Ungated | 69.17 | 63.33 | 70.00 | 66.67 | 67.29 | 52.50 | 60.00 | 100.00 |
| `EchoDistill` | **72.50** | **66.67** | **70.83** | **68.33** | **69.58** | **54.17** | 61.67 | 100.00 |

| | | | |
|---|---|---|---|
| `EchoDistill` − Initial | Mean: +1.88 [−1.67, 5.42] | All-4: +0.83 [−5.83, 7.50] | Consistency: −0.83 [−8.33, 6.67] |
| `EchoDistill` − Standard GRPO+KL | Mean: +3.33 [0.00, 6.67] | All-4: +3.33 [−3.33, 10.00] | Consistency: +5.83 [−2.50, 14.17] |

$[7.00, 20.67]$). However, its matched-audio correctness exceeds Initial by only **2.00 points**, with a confidence interval crossing zero ($[-4.33, 8.33]$). This auxiliary, automatically judged analysis supports sensitivity to matched audio; it is not an official AIR-Bench score or evidence of a reliable cross-model accuracy gain. `EchoDistill` also reaches **55.67%** judged relevance, a **+7.67**-point difference from Initial (95% CI **[1.33, 14.33]**), while its UAC rate of **11.33%** ties Standard GRPO+KL. Thus, the auxiliary advantage is clearest for judged relevance rather than uniform across all judge dimensions.

Table 16 examines a different shortcut: dependence on answer-option position. Across the original order and three cyclic rotations of 120 source questions, `EchoDistill`'s mean accuracy exceeds Initial's by **1.88 points** (95% CI: $[-1.67, 5.42]$), while answer-content consistency differs by $-0.83$ **points** (95% CI: $[-8.33, 6.67]$). Neither interval establishes a reliable change. The option-order test therefore does not show that this shortcut explains the matched-audio effect, but it also does not establish improved option-order robustness. Relative to Standard GRPO+KL, the mean-accuracy difference is **+3.33** points, but its interval touches zero (**[0.00, 6.67]**), while the All-4 and consistency intervals both cross zero. We therefore treat all option-position comparisons as inconclusive.

### D.5 Independent Noise Resampling

Table 17 tests whether the comparison with Standard GRPO+KL depends on a particular noise realization. We hold 600 clean source tasks fixed and generate two new noise waveforms per source; both methods receive the same waveform in each comparison. `EchoDistill`'s accuracy differences are **+1.00 point** (95% CI: $[-2.00, 4.00]$) and **+1.33 points** (95% CI: $[-1.67, 4.17]$) on the two realizations. The prespecified pooled analysis covers 1,200 waveforms clustered by the 600 clean sources and yields **+1.17 points** (95% CI: $[-1.33, 3.67]$), corresponding to **89** corrected and **75** broken predictions relative to Standard GRPO+KL. The point-estimate direction repeats across these two draws, but all reported intervals include zero; Table 17 does not establish a statistically reliable resampled advantage.

Table 17: Independent noise-resampling retest at $-10\,\mathrm{dB}$. Both methods receive the same 1,200 waveforms. Gain is `EchoDistill` minus Standard GRPO+KL; intervals use 10,000 paired bootstrap resamples clustered by clean source. C/B denotes corrected/broken cases.

| Setting | $N$ | Sources | GRPO+KL | Echo | Gain | 95% CI | C/B |
|---|---|---|---|---|---|---|---|
| *Independent realizations* | | | | | | | |
| Realization 1 | 600 | 600 | 60.67 | 61.67 | +1.00 | [−2.00, 4.00] | 45/39 |
| Realization 2 | 600 | 600 | 60.17 | 61.50 | +1.33 | [−1.67, 4.17] | 44/36 |
| *Prespecified pooled analysis* | | | | | | | |
| **Overall** | 1,200 | 600 | 60.42 | **61.58** | +1.17 | [−1.33, 3.67] | 89/75 |
| Music | 400 | 200 | 60.25 | 63.50 | +3.25 | [−1.25, 8.00] | 34/21 |
| Sound | 400 | 200 | 63.75 | 63.50 | −0.25 | [−4.50, 4.00] | 24/25 |
| Speech | 400 | 200 | 57.25 | 57.75 | +0.50 | [−3.75, 4.75] | 31/29 |

**Summary of the RQ3 evidence.** The matched-audio controls, including the auxiliary open-ended test, show that question-relevant audio contributes to `EchoDistill`'s predictions. ECAC indicates an advantage over the controlled optimizer under localized evidence corruption, while the teacher audit shows both the value and the limits of clean-conditioned guidance. At the same time, performance under mismatched audio remains nontrivial, ECAC does not establish an advantage over Initial, and the independent noise-resampling interval includes zero. These analyses support a bounded claim of acoustic-evidence dependence under the tested interventions, not a claim that every gain is caused by stronger grounding.

### D.6 Optimization Dynamics

Figure 5 reports clean–noisy prediction consistency during post-training. The curves show an overall upward trend in clean–noisy consistency during post-training, although the trajectories are not monotonic. Because consistency does not imply correctness, this figure is used only as an optimization diagnostic; task accuracy is evaluated separately.

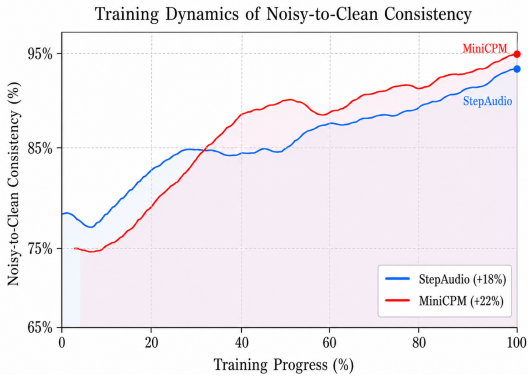


Figure 5: Clean–noisy consistency over post-training progress. MiniCPM-o improves rapidly during the early stage, whereas Step-Audio2 exhibits a steadier increase.