AUDITEAR: EXPERT AUDIO QUALITY ASSESSMENT WITH LARGE AUDIO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing speech-quality assessment systems typically reduce a recording to a scalar score, offering little insight into what is wrong, how severe the problem is, or how it should be fixed. We introduce PROQA, a dataset of 634 speech recordings assessed by professional audio engineers across 18 production-quality axes. Its 11,412 axis-level annotations pair ordinal severity ratings with issue tags, evidence-grounded assessments, and recommended fixes. To expand this costly expert supervision, we apply controlled augmentations across ten axes and compound their severity increments onto the original expert ratings, producing 15,524 audio variants with known quality changes. We use this corpus to train AUDITEAR, an audio-language model that diagnoses quality issues and recommends corrective actions. Our three-stage training paradigm combines Supervised Fine-Tuning, On-Policy Self-Distillation from privileged expert context, and GRPO with Verifiable Rewards. On a held-out test set, AUDITEAR achieves relative improvements of 41.2% in per-axis severity accuracy, 64.1% in pairwise accuracy, and 86.4% in holistic fix-F1 over its zero-shot backbone. It also outperforms Audio-Flamingo-Next, Kimi-Audio-7B-Instruct, and MiDashengLM-7B across all evaluated tasks, with gains of up to 2× over the strongest of these baselines. On external benchmarks, AUDITEAR achieves 3×–45× the correlation of published zero-shot audio-language baselines on QualiSpeech and reaches 74.3% preference accuracy on SpeechEval, compared with 70.4% for audio-language models evaluated zero-shot. These results show that expert, per-axis supervision can turn audio-quality assessment into a grounded and actionable diagnosis.

1 INTRODUCTION

Audio quality assessment and diagnosis are critical to professional quality control, content ingestion, evaluation, and audio and video production. Reliable assessment, however, still depends heavily on human listening tests and expert review, which are costly, time-consuming, and difficult to scale to modern audio data. Most automated approaches provide only a partial substitute by framing the problem as score prediction. Classical objective metrics such as PESQ (Rix et al., 2001) and STOI (Taal et al., 2011), and learned predictors of subjective ratings such as DNSMOS (Reddy et al., 2022) and NISQA (Mittag et al., 2021), reduce a recording to a single quality score or a small, fixed set of scores. These systems distinguish broad quality differences, but the same score can conceal fundamentally different acoustic conditions. They do not identify specific issues and their severity, explain the audible evidence underlying a judgement, recommend corrective actions, or adapt to a particular context. A recording with excessive reverberation, for example, may receive a score similar to one containing synthesis artifacts, clipping, or aggressive noise suppression. Yet downstream applications may weigh these failures differently, and each demands a different remedy. Individual differences in listener preference introduce further unexplained variance, limiting the practical use of predicted scores. In contrast, audio professionals conduct diagnostic assessments: an engineer identifies what stands out through listening and inspection and translates those observations into actionable recommendations.

Rapid progress in audio-language models provides a promising foundation for combining these capabilities. Such models inherit broad knowledge and flexible reasoning abilities from language models and can express their judgements in natural language. However, they are not generally trained for the specialized listening required in professional audio-quality assessment and may generate

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.

plausible diagnoses without grounding them in the input signal. Existing data also focus largely on average listeners rather than the nuanced distinctions made by professionals. We address these gaps with PROQA, an expert corpus of professional diagnostic judgements, and AUDITEAR, a model trained on it. Our contributions are:

1. **Expert annotations for audio-quality assessment.** We introduce PROQA, a dataset of 634 speech recordings annotated by professional audio engineers across 18 production-level audio quality axes. Its $11{,}412$ axis-level annotations combine issue tags, ordinal severity, written diagnoses, and corrective recommendations (Sec. 3.1).
2. **Controlled augmentation for scalable supervision.** We generate $15{,}524$ variants through controlled augmentations across ten axes, adding severity increments to existing expert ratings. This produces structured supervision and same-source comparison pairs with known changes in the assigned severity labels (Sec. 3.2).
3. **Expert audio assessment through three-stage training.** We train AUDITEAR through supervised fine-tuning, RGSD+OPSD self-distillation, and GRPO with verifiable rewards. Compared with its zero-shot backbone, AUDITEAR achieves relative improvements of 41.2% in per-axis severity accuracy, 64.1% in pairwise accuracy, and 86.4% in holistic fix-F1. It also outperforms Audio-Flamingo-Next, Kimi-Audio-7B-Instruct, and MiDashengLM-7B across all evaluated tasks on our held-out benchmark (Sec. 5).
4. **Generalization to external speech-quality benchmarks.** Without benchmark-specific training, AUDITEAR achieves up to $45\times$ the correlation with human ratings of published zero-shot audio-language baselines on QualiSpeech. On SpeechEval, it achieves 74.3% preference accuracy, compared with 70.4% for other zero-shot audio-language models (Sec. 5.4).

2 RELATED WORK

Speech quality assessment. Speech quality assessment systems typically predict a mean opinion score (MOS) (Saeki et al., 2022; Baba et al., 2024) or scores along a few perceptual axes (Reddy et al., 2022; Mittag et al., 2021), using large crowd-rated corpora from listening-test challenges (Cooper et al., 2022; Huang et al., 2022). Absolute ratings require listeners with different internal standards to use a shared numerical scale. Alternative approaches learn from comparative judgments through just-noticeable-difference labels (Manocha et al., 2020), contrastive content and acoustic embeddings (Manocha et al., 2021), or synthetic preference pairs used to train a no-reference score predictor (Manocha et al., 2022). Despite differing supervision, these approaches summarize perceptual evidence as scalar scores, typically relying on non-expert listeners. Even expert-listener datasets such as ODAQ (Torcoli et al., 2024), with 26 MUSHRA listeners, provide scalar ratings. Such scores support benchmarking and ranking but offer limited guidance for repair: a MOS of 2.4 does not distinguish microphone coloration or bandwidth limitations from noise, reverberation, or uneven dynamics. We instead capture professional judgment through ordinal severity ratings across 18 production-quality axes, paired with diagnoses and corrective recommendations.

Audio LLMs for quality assessment. Recent work extends speech-quality assessment to natural-language explanations. Wang et al. (2025b) combine per-axis scores, while ALLD (Chen et al., 2025) produces multi-axis judgments. Audiobox Aesthetics (Tjandra et al., 2025) predicts four no-reference aesthetic scores, while SpeechLLM-as-Judges (Wang et al., 2025a) and UniSRM (Wang et al., 2026) use judge-based consistency rewards. Preference-based reward models also expose weaknesses on synthetic speech and small-margin comparisons (Cao et al., 2026). Broader advances span general audio LLMs (Ghosh et al., 2025b; Goel et al., 2025; Ghosh et al., 2026; 2025a; Xu et al., 2025), speech representations (Seth et al., 2024), prompting and retrieval (Anand et al., 2025; 2026a), and paralinguistic understanding (Anand et al., 2026b), with perception, reasoning, and reliability evaluated through audio and multimodal studies (Sakshi et al., 2024; Goel et al., 2026; Kumar et al., 2026; Seth et al., 2025; 2026; Selvakumar et al., 2025). However, these advances leave a gap in professional assessment that connects each issue's severity to an actionable remedy. PROQA addresses this gap with audio-engineer annotations linking per-axis severity, diagnosis, and corrective recommendations. AUDITEAR learns from these annotations and controlled

Table 1: Annotation coverage of audio-quality datasets. Ticks indicate explicitly provided descriptions or improvement advice.

Dataset	Rated axes	Quality descriptions	Improvement advice	Annotation source
NISQA (Mittag et al., 2021)	5	✗	✗	Crowd listeners
ODAQ (Torcoli et al., 2024)	1	✗	✗	Expert listeners
AES-Natural (Tjandra et al., 2025)	4	✗	✗	Screened listeners
ALLD dataset (Chen et al., 2025)	5	✓	✗	Human ratings; LLM text
QualiSpeech (Wang et al., 2025b)	7	✓	✗	Human + LLM-assisted
SpeechEval (Wang et al., 2025a)	8	✓	✓	Human + LLM-assisted
UniSRM-Data/Bench (Wang et al., 2026)	7	✓	✗	Task-dependent human/LLM
PROQA (ours)	**18**	✓	✓	**Professional audio engineers**

Rated axes include overall quality where applicable; PROQA provides issue-severity ratings. Advice denotes an explicit annotation field or task. UniSRM's utterance-level quality task uses QualiSpeech's seven axes; its other tasks use different criteria.

augmentations through gold-conditioned self-distillation and verifiable rewards, with explicit controls to assess audio grounding.

Model-based evaluation and self-distillation. LLM judges evaluate open-ended responses (Zheng et al., 2023), while on-policy distillation trains on student-generated outputs to reduce the training–inference mismatch (Agarwal et al., 2024). Recent approaches use the same model as a teacher with privileged context (Zhao et al., 2026) or rubric guidance (Rezaei et al., 2026). In audio, UniSRM uses external rewards for per-axis reasoning (Wang et al., 2026). These approaches leave open how to transfer professional perceptual judgments into detailed audio-quality critiques. We address this gap by conditioning the teacher on expert assessments and augmentation-defined severity rubrics, providing axis-specific supervision for a student that must assess quality from the audio without access to those annotations. Table 1 compares the supervision and diagnostic capabilities provided by existing resources and PROQA.

3 METHODOLOGY

We first introduce PROQA, a speech-quality dataset annotated by professional audio engineers. We expand the corpus through controlled multi-axis augmentations that build on the expert severity ratings, formulate nine assessment tasks, and train AUDITEAR through supervised fine-tuning, on-policy self-distillation, and reinforcement learning with verifiable rewards.

3.1 THE PROQA DATASET

PROQA is, to our knowledge, the first corpus to capture professional audio-quality judgement at the granularity an audio engineer acts on. It contains 634 English speech recordings, each evaluated by one of 22 professional audio engineers across 18 production-quality axes spanning microphone capture, signal fidelity, room and space, noise, tonal balance, and level and dynamics (Appendix A.6). For each axis, the annotator provides applicable issue tags, an ordinal severity rating (*none*, *mild*, *moderate*, *significant*, or *severe*), a written audio-quality assessment grounded in audible evidence, and recommended fixes. This yields 11,412 axis-level ratings, 10,141 written assessments, and 8,823 written fixes (Table 11). The recordings cover read, conversational, expressive, audiobook, and interview speech, including acoustically simulated material, and range from studio-quality audio to severely compromised recordings. Approximately 80% come from public corpora: Common Voice (Ardila et al., 2020) (28%), Emilia-YODAS (He et al., 2025) (19%), EARS (Richter et al., 2024) (8%), Expresso (Nguyen et al., 2023) (8%), LibriVox (LibriVox) (8%), VoxCeleb2 (Chung et al., 2018) (4.5%), and VoxCeleb1 (Nagrani et al., 2017) (4.5%). The remaining ∼20% are high-quality recordings from a closed-source collection. Together, these sources provide diverse speech styles, recording conditions, and quality issues.

From expert judgments to training data. Because 634 expert-annotated recordings provide limited supervision for an audio-language model, we expand the corpus through reproducible aug-

mentations that raise the severity of selected axes while accounting for the original expert rating (Sec. 3.2). The resulting multi-task corpus contains approximately 79,925 Stage-1 SFT examples, 16,290 Stage-2 on-policy self-distillation examples, and 2,254 Stage-3 GRPO examples. We hold out complete annotators and their recordings for validation and testing; the split is detailed in Sec. 3.4.

Annotation protocol and consistency. We examine internal consistency of the collected data. The severity distribution covers the full scale, with 5,889 of 11,412 axis-level ratings labeled *none* (Figure 3). In addition, each annotator's axis-level severities are strongly associated with their independently assigned overall ratings, with a median within-annotator rank correlation of 0.92. Annotator selection, calibration, compensation, and review procedures are described in Appendix A.4.

3.2 Multi-axis controlled augmentation

Expert annotations are expensive, and 634 recordings provide limited training data for an audio-language model. We therefore expand ProQA through reproducible augmentations across ten axes with controllable audio parameters: noise, hum, reverberation, sibilance, tonal balance, loudness, bandwidth, distortion, dropouts, and handling pops. Each operation is applied at the model's 16 kHz input rate. We derive its four severity anchors from the parameter distribution observed in ProQA, using the lower and higher ends of the range, and the mean to define ordered severity points (Appendix A.7). Crucially, an augmentation increases the expert's existing severity rating rather than treating the source recording as clean:

$$s_a^{\text{final}} = \min\big(4,\ s_a^{\text{human}} + k_a\big), \tag{1}$$

where a denotes the quality axis, s_a^{human} is the engineer's severity rating for that axis, k_a is the number of augmentation steps, and s_a^{final} is the resulting severity. Ratings range from 0 (*none*) to 4 (*severe*). For example, a two-step augmentation raises *mild* reverberation to *significant*. This compounding rule matters because 31% of injections affect an axis that already contains a rated issue. Strong augmentations at severity 3 or 4 are applied to one axis at a time, while multi-axis combinations are limited to mild or moderate severity. We also exclude already-poor quality recordings according to their overall rating (Appendix A.7). In total, the procedure produces 15,524 augmented variants from 21,688 axis-level injections (Figure 3).

Updating annotations after augmentation. When an augmentation raises an axis's severity, the original assessment for that axis no longer describes the audio. We use Qwen3-4B-Instruct-2507 to update its audio-quality assessment, issue tags, and recommended fixes, conditioned on the axis, the new severity, and the engineer's original annotation. The model must preserve the substance and style of the expert annotation without mentioning the augmentation, its parameter values, or the explicit severity label. Annotations for unaffected axes and all annotations for original recordings remain unchanged. Appendix A.7 provides the complete generation and filtering constraints.

3.3 Task suite

We formulate the annotations as nine training tasks (Table 2). The three critique tasks generate increasingly broad assessments: Single-Axis evaluates one axis, Multi-Axis evaluates a specified subset, and Holistic assesses the complete recording. Their targets organize the output as audio quality assessment, issue tags, severity, and recommended fixes. The remaining tasks isolate severity estimation, issue detection, fix recommendation, and pairwise quality comparison.

Pairwise construction. A source recording and its augmented variant contain the same speech and differ only in the applied augmentation, giving a known quality ordering. We include both presentation orders to prevent position-based shortcuts. We also construct cross-source pairs, which require comparisons across different speakers and content, and label the pair type for separate evaluation.

Issue-detection options. We vary both the number of candidate tags and the number of correct options presented. Each example contains three to five options and may include all correct tags, a subset, or only one or two. This prevents the model from learning a fixed selection count or succeeding by selecting every option.

Table 2: Training examples for the nine tasks across the three training stages.

Task	**Stage 1** original	**Stage 2** original / aug.	**Stage 3** original / aug.	**Target output**
Single-Axis	6,678	924 / 2,316	—	assessment, tags, severity, and fixes for one axis
Multi-Axis	2,120	594 / 1,746	—	assessments and fixes for specified axes
Holistic	848	528 / 1,350	—	issues and fixes across all axes
Pairwise	30,000	990 / 3,000	194 / 508	preferred recording and the axes that differ
Severity (Blind)	13,156	330 / 690	— / 750	severity of one axis without issue tags
Severity (Issue-Given)	13,156	198 / 492	— / 500	severity of one axis given its issue tags
Issue Detection	20,000	330 / 690	— / 302	applicable issue tags selected from options
Fixes (Blind)	7,123	330 / 726	—	assessment and fixes without an issue hint
Fixes (Issue-Given)	7,123	330 / 726	—	recommended fixes given the issue tags
Total	**79,925**	**16,290**	**2,254**	

Stage 2 and Stage 3 report original / augmented examples. "—" indicates that the task is not used.

3.4 Training and model

AuditEar builds on Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), a mixture-of-experts omni model. We apply LoRA (Hu et al., 2022) ($r = 64$, $\alpha = 128$) to the language-model attention projections, audio encoder, and audio projector. The mixture-of-experts feed-forward layers remain frozen, leaving 120.5M trainable parameters.

Annotator-disjoint splits. Each recording is assessed by one engineer (Sec. 3.1), so holding out recordings alone would retain that engineer's rating style in training. We instead hold out entire annotators: three annotators comprising 102 recordings form the test set, and one comprising 22 recordings forms the validation set. They are selected to span different levels of rating strictness. The test set therefore measures generalisation to unseen recordings and annotators. We evaluate on a class-balanced subset so that the majority *none* class does not dominate the reported metrics. Appendix Figure 2 shows the severity distributions for training and evaluation.

Stage 1: Supervised Fine-Tuning. We jointly fine-tune the model on all nine tasks for one epoch using approximately $79{,}925$ examples from the original recordings (Table 2). This stage teaches the model the task formats and provides the initial policy for the subsequent stages. Augmented recordings and same-source comparison pairs are introduced only in Stages 2 and 3. As an ablation, we also continue supervised fine-tuning from the Stage-1 checkpoint using approximately $163{,}257$ original and augmented examples. This SFT-continuation baseline is compared with On-Policy Self-Distillation in Sec. 5.3.

Stage 2: On-Policy Self-Distillation. Stage 2 applies OPSD+RGSD (On-Policy Self-Distillation and Rubric-Guided Self Distillation) across all nine tasks. The student receives only the audio and task instruction. For the teacher pass, we use the same network with its LoRA adapters disabled, recovering the underlying pre-adaptation Qwen3-Omni backbone rather than loading a separate teacher checkpoint. The teacher receives privileged context unavailable to the student: the expert assessment under OPSD and, when RGSD is enabled, the per-axis severity rubric (Zhao et al., 2026; Rezaei et al., 2026). Distillation therefore transfers what the backbone produces with the expert reference in context into a student that must generate the assessment directly from the audio.

The student generates responses on-policy, and the teacher scores the same tokens. We minimise a generalised Jensen–Shannon divergence between their token distributions (Agarwal et al., 2024):

$$m \;=\; \beta\, p_T + (1-\beta)\, p_S, \qquad \mathcal{L} \;=\; \frac{1}{|R|} \sum_{t \in R} \Big[\, \beta\, \mathrm{KL}\big(p_{T,t} \,\|\, m_t\big) + (1-\beta)\, \mathrm{KL}\big(p_{S,t} \,\|\, m_t\big) \Big], \quad (2)$$

where $p_{S,t}$ and $p_{T,t}$ are the student's and teacher's token distributions at position t, and m_t is their interpolated distribution. The coefficient β controls the contribution of each model, $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence, and R is the set of response-token positions over which the loss is averaged. For efficiency, the distributions are restricted to the teacher's top-k tokens. Stage 1 provides an initial student policy whose on-policy responses are sufficiently close to the privileged teacher distribution for distillation. Stage 2 also retains approximately 30% original recordings

alongside the augmented examples, preserving the model's ability to recognize when no issue is present.

Stage 3: Reinforcement Learning with Verifiable Rewards. Self-distillation is effective for open-ended assessments but provides weaker supervision for tasks with directly verifiable answers, since the teacher receives the reference information as context. We therefore apply Group Relative Policy Optimisation (GRPO) to Issue Detection, Pairwise comparison and Severity (Blind Issue-Given) . Training begins from the Stage-2 model and remains KL-anchored to its frozen policy.

For each prompt, we sample G responses and assign a reward r_i to each response i. We standardize these rewards within the group to obtain the relative advantage $A_i = (r_i - \mu)/\sigma$, where μ and σ are the group mean and standard deviation. The per-response objective is

$$\mathcal{L}_{\mathrm{GRPO}} \;=\; -\,A_i\,\overline{\log \pi_\theta} \;+\; \beta_{\mathrm{KL}}\,\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big), \tag{3}$$

where π_θ is the trainable student policy and $\overline{\log \pi_\theta}$ is its mean token log-probability for the sampled response. The frozen Stage-2 policy π_{ref} serves as the reference, while β_{KL} controls the strength of the KL penalty that keeps the updated policy close to it. The advantage A_i increases the probability of responses that score above the group mean and decreases it for responses that score below the mean.

The three tasks use programmatic rewards:

$$r_{\mathrm{sev}} = 1 - \frac{|\hat{s} - s|}{4}, \qquad r_{\mathrm{tag}} = \mathrm{F1}\big(\hat{\mathcal{T}}, \mathcal{T}\big), \qquad r_{\mathrm{pair}} = \mathbf{1}[\hat{y} = y], \tag{4}$$

where $\hat{s}$ and s are the predicted and reference severity levels on the five-point ordinal scale. The sets $\hat{\mathcal{T}}$ and $\mathcal{T}$ contain the predicted and reference issue tags, respectively. Finally, $\hat{y}$ and y denote the predicted and reference winners in a pair, and $\mathbf{1}[\cdot]$ is the indicator function. Thus, r_{sev} gives partial credit for near-misses, r_{tag} measures overlap between predicted and reference tags, and r_{pair} gives a binary reward for selecting the correct recording. The pairwise target is derived from the data: the original recording over its augmented variant, the lower-severity variant, or the recording with the higher expert rating. Self-distillation therefore supervises the longer open-ended assessments, while GRPO focuses on short outputs with directly verifiable answers.

4 EXPERIMENTS

We compare AUDITEAR with strong audio-language baselines and its intermediate training stages on the annotator-disjoint test set. We first define the baselines and evaluation metrics, then summarize the training configuration. Tables 3 and 4 report the resulting performance.

4.1 BASELINES

We compare against four general-purpose audio-language models used zero-shot with the same task instructions: the Qwen3-Omni-30B-A3B-Instruct backbone (Xu et al., 2025), Audio-Flamingo-Next (Ghosh et al., 2026), Kimi-Audio-7B-Instruct (KimiTeam et al., 2025), and MiDashengLM-7B (Dinkel et al., 2025). We also trace the training progression from +**Human**, which is fine-tuned only on the original human critiques, through Stage 1 SFT, Stage 2 RGSD+OPSD, and the final AUDITEAR model after GRPO. A **text-only control** receives the instruction without audio to test whether performance can be explained by textual priors alone (Table 9).

4.2 EVALUATION PROTOCOL

We evaluate all systems on the same class-balanced subset of the annotator-disjoint test split ($5{,}652$ examples).

Deterministic metrics. For the Single-Axis task, severity is measured by exact accuracy and issue tags by set-F1. The two dedicated Severity tasks are also evaluated using exact accuracy, while Issue Detection uses micro-F1 over the selected tags. Pairwise comparison is measured by two-alternative forced-choice accuracy (2AFC). We additionally report *position consistency*, which measures whether the model selects the same recording when a pair is presented in both orders.

Table 3: Performance on the critique tasks.

Task · Metric	Qwen3 zero-shot	AF-Next zero-shot	Kimi-7B zero-shot	MiDasheng-7B zero-shot	+Human	Stage 1 SFT	Stage 2 RGSD+OPSD	AUDITEAR Stage 3 (GRPO)
Single-Axis · sev-exact	0.199	0.210	0.274	0.239	0.246	0.254	<u>0.278</u>	**0.281**
Single-Axis · tag-F1	0.405	0.379	0.366	0.423	0.410	0.300	<u>0.496</u>	**0.498**
Single-Axis · issue-F1	0.490	0.444	0.389	0.460	0.427	0.311	<u>0.497</u>	**0.509**
Single-Axis · fix-F1	0.421	0.326	0.232	0.201	0.333	0.207	<u>0.500</u>	**0.519**
Multi-Axis · issue-F1	0.417	0.334	0.188	0.258	0.293	0.340	<u>0.502</u>	**0.543**
Multi-Axis · fix-F1	0.292	0.221	0.108	0.120	0.209	0.204	<u>0.441</u>	**0.496**
Holistic · issue-F1	0.331	0.241	0.082	0.265	0.170	0.284	<u>0.490</u>	**0.528**
Holistic · fix-F1	0.213	0.099	0.044	0.131	0.103	0.176	<u>0.354</u>	**0.397**

Best values are in **bold**; second-best values are <u>underlined</u>.

Table 4: Performance on severity, issue detection, fixes, and pairwise comparison.

Task · Metric	Qwen3 zero-shot	AF-Next zero-shot	Kimi-7B zero-shot	MiDasheng-7B zero-shot	+Human	Stage 1 SFT	Stage 2 RGSD+OPSD	AUDITEAR Stage 3 (GRPO)
Severity (Blind) · acc	0.283	0.228	0.226	0.281	<u>0.285</u>	<u>0.285</u>	0.232	**0.319**
Severity (Issue-Given) · acc	0.456	0.348	0.439	0.412	0.507	<u>0.635</u>	0.529	**0.647**
Issue Detection · micro-F1	0.353	<u>0.430</u>	0.375	0.386	0.364	0.373	0.429	**0.438**
Fixes (Blind) · fix-F1	0.375	0.379	0.201	0.322	0.361	0.251	<u>0.462</u>	**0.476**
Fixes (Issue-Given) · fix-F1	0.386	0.344	0.284	0.381	0.434	0.378	<u>0.459</u>	**0.467**
Pairwise · 2AFC	0.485	0.475	0.508	0.462	0.507	<u>0.757</u>	0.735	**0.796**
Pairwise · pos-consist	0.271	0.099	0.226	0.081	0.249	<u>0.582</u>	0.526	**0.658**

Best values are in **bold**; second-best values are <u>underlined</u>.

Free-text metrics. For written assessments and fixes, Gemma-4-31B-it semantically matches predicted and reference items to compute *issue-F1* and *fix-F1*. Multi-Axis and Holistic outputs are matched separately for each axis and then aggregated, allowing equivalent wording to receive credit. The judge receives no audio and does not assign quality scores; it only identifies semantic matches between predicted and reference items.

4.3 TRAINING CONFIGURATION

Each training stage runs for one epoch. Stage 1 uses $79{,}925$ examples from the original recordings across all nine tasks (Table 2). Stage 2 uses $16{,}290$ examples across the same tasks, retaining approximately 30% original recordings among the augmented examples. Stage 3 uses a balanced set of $2{,}254$ examples from the verifiable Severity, Issue Detection, and Pairwise tasks. It samples $G = 8$ responses per prompt and $G = 16$ for Pairwise comparison, with a KL coefficient of $\beta_{\mathrm{KL}} = 0.1$ relative to the frozen Stage-2 policy. Complete optimisation, rollout, and hardware settings are provided in Appendix Table 7.

5 RESULTS AND ABLATIONS

5.1 MAIN RESULTS

Tables 3 and 4 compare the zero-shot audio-language models with each stage of our training pipeline. All systems are evaluated on the same class-balanced, annotator-disjoint set of $5{,}652$ examples.

General-purpose audio-language models perform poorly on professional quality assessment despite their broader audio capabilities. Audio-Flamingo-Next, Kimi-Audio-7B-Instruct, and MiDashengLM-7B reach only 0.099, 0.044, and 0.131 Holistic fix-F1, respectively, compared with 0.397 for AUDITEAR—relative improvements of 301%, 802%, and 203%. On Pairwise comparison, AUDITEAR's 0.796 accuracy improves relatively by 68% over Audio-Flamingo-Next (0.475), 57% over Kimi-Audio-7B-Instruct (0.508), and 72% over MiDashengLM-7B (0.462). AUDITEAR achieves the best result on every task, including 0.528 Holistic issue-F1 and 0.647 issue-given severity accuracy.

Table 5: Zero-shot transfer to QualiSpeech, measured by Pearson correlation.

Aspect	AUDITEAR	Best zero-shot	Best fine-tuned
Background noise	<u>0.631</u>	0.014	**0.696**
Distortion	<u>0.490</u>	0.056	**0.614**
Continuity	<u>0.459</u>	0.201	**0.535**
Overall quality	<u>0.628</u>	0.250	**0.660**

Best values are in **bold**; second-best values are <u>underlined</u>.

Table 6: Zero-shot transfer to the SpeechEval comparison split (2AFC).

Method	2AFC	Context
AUDITEAR (zero-shot, OOD)	<u>0.743</u>	ours
SQ-LLM (fine-tuned on SpeechEval)	**0.751**	in-domain
Untuned audio-language models	0.704	zero-shot
Majority-class predictor	0.565	chance

Best values are in **bold**; second-best values are <u>underlined</u>.

The training stages make complementary contributions. Stage 1 establishes the task formats and raises Pairwise 2AFC to 0.757. Stage 2 substantially improves open-ended coverage, lifting issue-, fix-, and tag-F1 across the critique tasks, but weakens some discriminative tasks. GRPO restores and further improves these capabilities while retaining the Stage-2 gains: the final AUDITEAR reaches 0.319 on blind severity, 0.647 on issue-given severity, and 0.658 position consistency.

5.2 Ablation: Audio grounding and per-axis performance

We test whether AUDITEAR relies on the recording by replacing the audio with noise, swapping in another recording, or removing it entirely (Table 9). Pairwise 2AFC falls from 0.796 to approximately chance (0.505, 0.511, and 0.500), while blind severity drops from 0.319 to 0.170–0.228. This collapse shows that the model's decisions depend on the audio rather than prompt priors.

We also separate the ten axes with augmentation coverage from the eight trained only on original recordings (Table 10 and Figure 1). Augmentation-covered axes perform better on issue identification (0.546 vs. 0.500) and recommended fixes (0.507 vs. 0.428), with the larger gain on fixes.

5.3 Ablation: SFT or self-distillation after Stage 1?

Starting from the same Stage-1 checkpoint, we compare continued SFT with RGSD+OPSD self-distillation. Continued SFT improves the discriminative tasks but sharply reduces open-ended coverage; for example, Holistic fix-F1 falls from 0.354 to 0.024. Self-distillation preserves the breadth of the assessments, after which GRPO improves severity and pairwise discrimination. Appendix Table 8 reports the complete comparison.

5.4 Generalisation to external benchmarks

We first evaluate AUDITEAR zero-shot on QualiSpeech (Wang et al., 2025b), mapping its quality aspects to our severity outputs. Table 5 shows that AUDITEAR exceeds the strongest published zero-shot audio-language model by approximately 3–45$\times$ in correlation and comes within 0.03–0.12 of systems fine-tuned on QualiSpeech.

SpeechEval (Wang et al., 2025a) provides a stricter out-of-domain test: it compares different text-to-speech and natural utterances, whereas our pairwise training is dominated by same-source comparisons and contains no synthetic-speech naturalness data. Still, AUDITEAR reaches 0.743 preference accuracy across four languages, outperforming untuned audio-language models (0.704) and approaching SQ-LLM fine-tuned on SpeechEval (0.751). Its 0.768 position consistency also indicates robust decisions under order reversal.

6 Conclusion, Limitations and Future Work

We introduced ProQA, an expert corpus of 634 speech recordings assessed across 18 production-quality axes, and AuditEar, an audio-language model trained to identify quality issues, estimate their severity, explain the audible evidence, and recommend fixes. Controlled augmentation expands the expert annotations while preserving their original severity context, and our three-stage training combines supervised fine-tuning, self-distillation, and verifiable rewards. AuditEar outperforms its zero-shot backbone and strong general-purpose audio-language models across all evaluated tasks, and generalises without benchmark-specific training to QualiSpeech and SpeechEval. These results show that structured, per-axis supervision can turn audio-quality assessment from scalar prediction into an actionable diagnosis.

Limitations and future work. ProQA contains only 634 English speech recordings, and collecting professional annotations is costly. Each recording is assessed by one engineer, preventing direct measurement of inter-annotator agreement, although we report evidence of within-annotator consistency. Controlled augmentation covers only ten of the 18 axes and approximates quality issues through reproducible signal transformations rather than naturally occurring failures. Our open-ended metrics also rely partly on an LLM judge. Future work should expand the corpus to additional languages, music, and multi-speaker audio; collect overlapping expert annotations; extend controlled augmentation to more axes and naturally occurring conditions; and evaluate AuditEar as a quality-control system and training reward for enhancement and generation models.

AI use statement

We disclose two uses of generative AI in this work, matching the categories selected in the submission form. **Generating synthetic datasets.** A large language model updates the per-axis assessment, issue tags, and recommended fixes for the controlled-augmentation variants so that the free-text supervision reflects the applied augmentation (Sec. 3.2). The underlying ordinal severity labels are determined by the reproducible augmentation rule rather than by the model. **Aiding and polishing writing.** Generative AI was used to improve the grammar, clarity, and readability of the manuscript. It was not used to develop the study's methodology, design its claims, or generate its scientific conclusions. The authors reviewed and verified all AI-assisted text and generated supervision and take full responsibility for the content of this work.

Ethics statement

The recordings in ProQA combine existing publicly released speech corpora, used under their original licences, with a curated collection of high-quality closed-source audio. We collected no new speech recordings. Dataset annotation was performed by paid professional audio engineers under the protocol described in Sec. 3.1 and Appendix A.4. Annotators were compensated at approximately $75 per hour under a standard work agreement. They assessed only technical production quality—including microphone capture, room acoustics, noise, and level—and were not asked to judge speakers, accents, or linguistic content. No personal or demographic information about the speakers enters the annotations or model targets.

Reproducibility statement

We document the data construction, training, and evaluation procedures throughout the paper and appendix. ProQA, its source corpora, and the professional annotation protocol are described in Secs. 3.1 and 3.1. The controlled-augmentation operations, parameter-to-severity mappings, and compounding rule are specified in Sec. 3.2, Appendix A.7, Table 13, and Eq. 1. The nine task formulations and their prompts are provided in Sec. 3.3 and Appendix A.9. The annotator-disjoint data splits are described in Sec. 3.4. Model architecture, adapted modules, training objectives, and complete hyperparameters are reported in Secs. 3.4 and 4.3 and Appendix Table 7. Finally, Sec. 4.2 defines the deterministic and model-based evaluation metrics.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13649.

Nishit Anand, Ashish Seth, Ramani Duraiswami, and Dinesh Manocha. TSPE: Task-specific prompt ensemble for improved zero-shot audio classification. In *ICASSP Workshop on Speech and Audio Language Models (SALMA)*, 2025.

Nishit Anand, Ashish Seth, Sreyan Ghosh, Dinesh Manocha, and Ramani Duraiswami. FIGMA: Towards fine-grained music retrieval. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026a. arXiv:2606.06615.

Nishit Anand, Jiaqi Su, Ke Chen, Yunyun Wang, Dinesh Manocha, Ramani Duraiswami, Rithesh Kumar, and Zeyu Jin. Para-llm: A unified approach to paralinguistic and acoustic speech understanding. In *Interspeech*, 2026b. arXiv:2609.22771.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pp. 4218–4222, 2020.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The T05 system for the VoiceMOS challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Yifei Cao, Changhao Jiang, Jiabao Zhuang, Jiajun Sun, Ming Zhang, Zhiheng Xi, Hui Li, Shihan Dou, Yuran Wang, Yunke Zhang, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. From scores to preferences: Redefining evaluation paradigm for speech quality reward modeling. In *Findings of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2510.00743.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. Audio large language models can be descriptive speech quality evaluators. *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2501.17202.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pp. 1086–1090, 2018. doi: 10.21437/Interspeech.2018-1929.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. Generalization ability of MOS prediction networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2022.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*, 2025.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*, 2025a.

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

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, et al. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. arXiv:2603.14145, 2026.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 33:4044–4054, 2025. doi: 10.1109/TASLPRO.2025.3612835.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. The VoiceMOS challenge 2022. In *Proc. Interspeech*, pp. 4536–4540, 2022.

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

Sonal Kumar, Simon Sedlacek, Vaibhavi Lokegaonkar, Fernando Lopez, Wenyi Yu, Nishit Anand, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2508.13992.

LibriVox. LibriVox: Free public domain audiobooks. `https://librivox.org`. Accessed: 2026-09-26.

Pranay Manocha, Adam Finkelstein, Richard Zhang, Nicholas J. Bryan, Gautham J. Mysore, and Zeyu Jin. A differentiable perceptual audio metric learned from just noticeable differences. In *Interspeech*, 2020. arXiv:2001.04460.

Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. arXiv:2102.05109.

Pranay Manocha, Zeyu Jin, and Adam Finkelstein. SQAPP: No-reference speech quality assessment via pairwise preference. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 891–895, 2022. doi: 10.1109/ICASSP43922.2022.9746615.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech*, 2021. arXiv:2104.09494.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pp. 2616–2620, 2017. doi: 10.21437/Interspeech.2017-950.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech 2023*, pp. 4823–4827, 2023. doi: 10.21437/Interspeech.2023-1905.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. arXiv:2110.01763.

MohammadHossein Rezaei, Anas Mahmoud, Zihao Wang, Utkarsh Tyagi, Advait Gosai, Razvan-Gabriel Dumitru, Aakash Sabharwal, Bing Liu, and Yunzhong He. Rubric-guided self-distillation: Post-training without rubric verifiers. *arXiv preprint arXiv:2606.12507*, 2026.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech 2024*, pp. 4873–4877, 2024. doi: 10.21437/Interspeech.2024-153.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)—a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 749–752, 2001.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech*, 2022. arXiv:2204.02152.

S. 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. *arXiv preprint arXiv:2410.19168*, 2024.

Ramaneswaran Selvakumar, Sonal Kumar, Hemant Kumar Giri, Nishit Anand, Ashish Seth, Sreyan Ghosh, and Dinesh Manocha. Do audio-language models understand linguistic variations? In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Short Papers*, 2025. arXiv:2410.16505.

Ashish Seth, Ramaneswaran Selvakumar, S. Sakshi, Sonal Kumar, Sreyan Ghosh, and Dinesh Manocha. EH-MAM: Easy-to-hard masked acoustic modeling for self-supervised speech representation learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6386–6400, 2024.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. EgoIllusion: Benchmarking hallucinations in egocentric video understanding. *arXiv preprint arXiv:2508.12687*, 2025.

Ashish Seth, Sonal Kumar, Ramaneswaran Selvakumar, Nishit Anand, Utkarsh Tyagi, Prem Seetharaman, Ramani Duraiswami, and Dinesh Manocha. Audio hallucination attacks: Probing the reliability of large audio language models. In *Interspeech*, 2026. arXiv:2603.29263.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Matteo Torcoli, Chih-Wei Wu, Sascha Dick, Phillip A. Williams, Mhd Modar Halimeh, William Wolcott, and Emanuël A. P. Habets. ODAQ: Open dataset of audio quality. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. arXiv:2401.00197.

James Traer and Josh H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. SpeechLLM-as-judges: Towards general and interpretable speech quality evaluation. *arXiv preprint arXiv:2510.14664*, 2025a.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 23588–23609, 2025b.

Yuanyuan Wang, Dongchao Yang, Yayue Deng, Zhiyong Wu, Yiwen Guo, Helen Meng, and Xixin Wu. UniSRM: A unified speech reward model for reasoning-based fine-grained assessment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2605.23261.

Table 7: Training configuration for AUDITEAR.

Setting	Stage 1 SFT	Stage 2 RGSD+OPSD	Stage 3 GRPO
Training examples	79,925	16,290	2,254
Tasks	All nine	All nine	Severity, Issue Detection, Pairwise
Data	Original	Original and augmented	Original and augmented
Original-data retention	100%	∼30%	Task-balanced
Epochs	1	1	1
Learning rate	5×10^{-5}	—	2×10^{-6}
Schedule	Cosine	—	—
Warmup	10%	—	—
Per-device batch size	1	—	—
Gradient accumulation	8	—	—
Maximum sequence length	4096	—	—
Student rollout budget	—	1024 tokens	16–256 tokens
Rollouts per prompt	—	—	$G = 8$; $G = 16$ for Pairwise
Sampling temperature	—	—	1.0
JSD coefficient	—	$\beta = 0.5$	—
Teacher support	—	top-$k = 64$	—
KL coefficient	—	—	$\beta_{\mathrm{KL}} = 0.1$
Reference policy	—	Pre-adaptation backbone	Frozen Stage-2 policy
Checkpoints per epoch	2	—	4
Precision	bf16	—	—
Gradient checkpointing	Enabled	—	—
Parallelism	Data parallel	—	—
Hardware	Up to 8× A100-80GB	—	—

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.18734.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

A APPENDIX

The appendix provides complete training settings, additional ablations and analyses, dataset statistics, augmentation details, and task examples.

A.1 TRAINING DETAILS

Table 7 provides the complete optimisation, rollout, and hardware configuration for the three training stages.

A.2 SFT-CONTINUATION VS. SELF-DISTILLATION: FULL ABLATION

SFT vs. RGSD+OPSD on top of Stage 1. The central training claim is that, once Stage 1 has taught the format, *how* one continues matters. We compare two ways of building on the *same* Stage-1 checkpoint: continuing with more supervised fine-tuning (an SFT-continuation baseline, A2) versus the Stage-2 RGSD+OPSD self-distillation. Self-distillation is the stronger continuation

Table 8: Ablation: how to continue from the same Stage-1 checkpoint. **A2** continues with more supervised fine-tuning (the SFT-continuation baseline); **Stage 2** is RGSD+OPSD self-distillation; **Stage 3** adds GRPO. Class-balanced held-out subset, Gemma-4-31B-it judge. More SFT (A2) improves the discriminative tasks (severity, pairwise) but collapses open-ended coverage (issue/fix/tag F1); self-distillation preserves coverage and GRPO recovers the discrimination on top. Best in each row in bold.

Task · Metric	**A2** SFT-cont.	**Stage 2** RGSD+OPSD	**Stage 3** GRPO
Single-Axis · sev-exact	**0.310**	0.278	0.281
Single-Axis · tag-F1	0.327	0.496	**0.498**
Single-Axis · issue-F1	0.418	0.497	**0.509**
Single-Axis · fix-F1	0.138	0.500	**0.519**
Multi-Axis · issue-F1	0.225	0.502	**0.543**
Multi-Axis · fix-F1	0.100	0.441	**0.496**
Holistic · issue-F1	0.063	0.490	**0.528**
Holistic · fix-F1	0.024	0.354	**0.397**
Severity (Blind) · acc	0.316	0.232	**0.319**
Severity (Issue-Given) · acc	0.645	0.529	**0.647**
Issue Detection · micro-F1	0.393	0.429	**0.438**
Fixes (Blind) · fix-F1	0.218	0.462	**0.476**
Fixes (Issue-Given) · fix-F1	0.347	0.459	**0.467**
Pairwise · 2AFC	**0.903**	0.735	0.796
Pairwise · pos-consist	**0.870**	0.526	0.658

for everything the assessor must *generate* (Table 8). Continuing with more supervised fine-tuning sharpens the *discriminative* skills — severity accuracy rises (Severity-Blind $0.232 \rightarrow 0.316$) and pairwise agreement jumps (2AFC $0.735 \rightarrow 0.903$) — but it *collapses* open-ended coverage: issue-, fix- and tag-F1 fall sharply across the Single-Axis, Multi-Axis and Holistic critiques and Issue Detection (Holistic fix-F1 $0.354 \rightarrow 0.024$, Multi-Axis fix-F1 $0.441 \rightarrow 0.100$), because plain SFT on a single target makes the model terse and under-report. In short, discrimination alone favors more SFT, but at the cost of the diagnostic coverage that is the point of the system. This is why we do not simply add more supervised fine-tuning: RGSD+OPSD preserves that coverage, and Stage 3 GRPO then recovers the discrimination from a checkable reward without the coverage loss. The reason self-distillation works here is specific to our *controlled-augmentation* data: because every degradation is injected with a known severity and every pairwise delta is exact in the injected severity, the teacher — the same network given the gold assessment and the per-axis severity rubric — is a reliable, on-distribution supervisor. Distilling from it on the student's *own* rollouts therefore teaches a calibrated critique rather than memorising one reference phrasing, which is all that plain SFT on a single target can do. Stage 3 GRPO then supplies what neither can: correctness on the *verifiable* tasks from a checkable reward (Eq. 4) rather than from a teacher that has already been shown the answer.

A.3 Audio-grounding controls and per-axis results

This appendix gives the full numbers behind the two analyses summarised in Sec. 5: the audio-grounding controls (Table 9), which confirm the model answers from the audio and not the prompt, and the per-axis breakdown of the final model (Table 10), which separates the axes with synthetic coverage from those supported only by real recordings.

A.4 Annotator panel and calibration

ProQA was annotated by 22 professional audio engineers (16 male, 6 female), hired online after a review of their portfolios. Every annotator had at least five years of professional mixing and mastering experience, with credits spanning film, television, news broadcast and online video production. Before annotating, each engineer completed a calibration pass consisting of a written guidelines document, which described the data-collection goal, the questionnaire format and what each field required, together with eight anchor recordings labelled at low and high quality to align their use of the severity scale. This onboarding took approximately four hours per annotator. Annotators were compensated at approximately $\$75$ per hour, well above minimum wage, under a standard work

Table 9: Audio-grounding controls. Deterministic, judge-independent metrics for AUDITEAR on the real recordings versus three corrupted-input conditions: the audio replaced by noise, swapped for a different recording (question unchanged), and removed entirely. The text prompt is identical across conditions. Every metric falls to its chance floor once the audio is altered.

Metric	Real audio	Noise	Swap	No audio	Chance
Severity (Blind) · exact-acc	**0.319**	0.170	0.226	0.228	0.20 / 0.22[†]
Issue Detection · micro-F1	**0.438**	0.321	0.369	0.396	$\sim$0.40[‡]
Pairwise · 2AFC	**0.796**	0.505	0.511	0.500	0.500

[†] uniform / majority-class floor. [‡] random-selection floor given the option count.

Table 10: Per-axis results for the final model (AUDITEAR, GRPO), on the balanced test subset, separated into the 10 axes with synthetic coverage and the 8 supported only by the real recordings. Metrics are severity exact-accuracy and issue / fix micro-F1. The covered axes are stronger at naming the issue and its fix.

Group	Axis	sev-exact	issue-F1	fix-F1
Degradable	Signal-chain noise: hiss, hum, buzz	0.315	0.590	0.626
	Stationary noise	0.357	0.583	0.558
	Reverb	0.263	0.596	0.665
	Sibilance & harshness	0.321	0.628	0.668
	Overall tonal balance & EQ	0.275	0.545	0.477
	Loudness & dynamic range	0.355	0.593	0.562
	Frequency response limitations	0.224	0.495	0.371
	Distortion	0.485	0.360	0.286
	Dropouts / glitches / artefacts	0.392	0.421	0.302
	Handling & pops	0.207	0.574	0.412
	Group mean	**0.314**	**0.546**	**0.507**
Real-only	Dynamic noise	0.591	0.239	0.458
	Mic quality & clarity	0.314	0.602	0.472
	Mic placement	0.286	0.468	0.319
	Capture consistency over time	0.393	0.400	0.413
	Flutter echoes / resonances	0.382	0.497	0.338
	Near-surface reflections	0.332	0.510	0.571
	Processing issues (DSP & AI)	0.443	0.268	0.402
	Voice directness	0.321	0.546	0.438
	Group mean	**0.367**	**0.500**	**0.428**

agreement, and were located worldwide. Each recording was assigned to a single engineer, and reliability comes from who annotates and from iterative review rather than from redundant labelling (Sec. 3.1).

A.5 ADDITIONAL STATISTICS AND PIPELINE FIGURES

This appendix collects supporting statistics for PROQA (Table 11) and full-size versions of the pipeline and distribution figures referenced from the main text.

A.6 THE SCORED AXES

Table 12 lists all 18 scored axes, grouped as they appear in the questionnaire, together with the number of recordings on which each carries a rated issue and whether the axis admits a synthetic degradation.

A.7 THE DEGRADATION ENGINE

Per-axis mechanism. All operators run at 16 kHz. *Stationary noise*: additive white Gaussian noise to a target SNR. *Signal-chain noise*: summed $60/120/180$ Hz sinusoids (mains hum) to a target

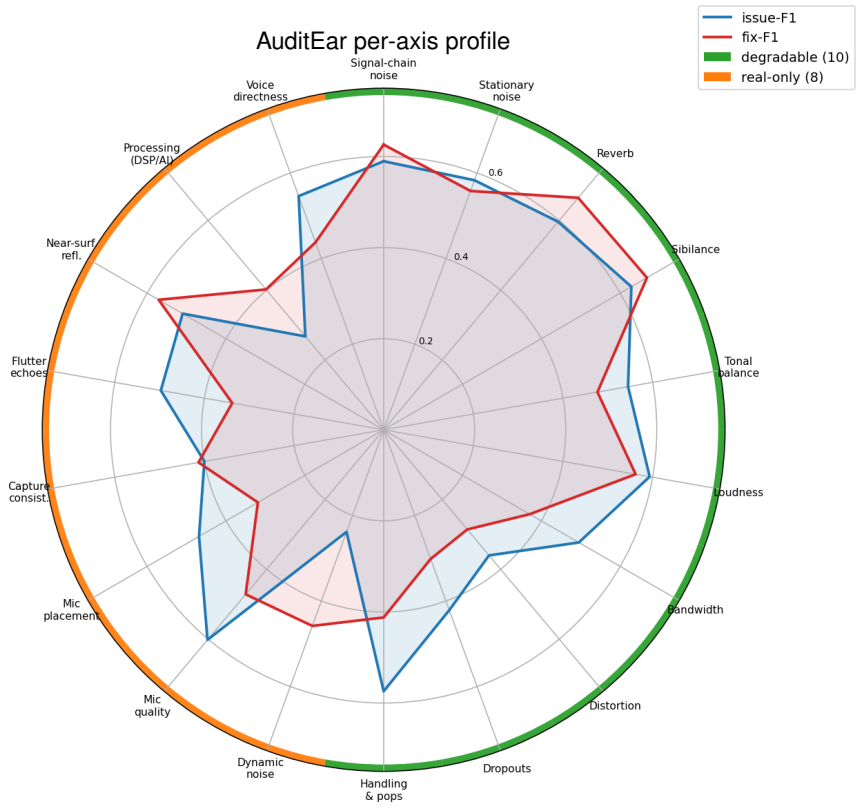

Figure 1: Per-axis profile of the final model (AUDITEAR, GRPO) across all 18 scored axes: issue-F1 (blue) and fix-F1 (red). The coloured outer arc marks each axis's group — green for the 10 axes with synthetic coverage, orange for the 8 supported only by real recordings. The same numbers appear in Table 10.

Table 11: Core statistics of PROQA. Severity counts are over annotated axes; an axis may be rated *none*, which is an informative judgement rather than a missing label.

Statistic	**Value**
Expert-annotated recordings	634
Credentialed annotators	22
Annotated, scored and evaluated axes per recording	18
Scored axis instances	11,412
carrying an issue (severity $\geq$ mild)	5,523
Free-text assessments written by annotators	10,141
Free-text fixes written by annotators	8,823
Severity: none / mild / moderate / significant / severe	5,889 / 2,619 / 1,656 / 787 / 461
Overall rating 1 / 2 / 3 / 4 / 5	38 / 191 / 92 / 251 / 62
Synthetically degraded variants	15,524
Degradation injections	21,688
compounding onto an existing rated issue	6,797 (31%)
Degradable axes	10 of 18
Stage-1 SFT examples	79,925
Stage-2 / Stage-3 examples	16,290 / 2,254

SNR. *Reverberation*: convolution with a real room from the MIT-IR survey (Traer & McDermott, 2016) whose RT60 is nearest the target, RMS-matched. *Sibilance*: an RBJ high-shelf boost above 5 kHz. *Tonal balance*: a progressive high-frequency roll-off above 1 kHz. *Bandwidth*: a sixth-

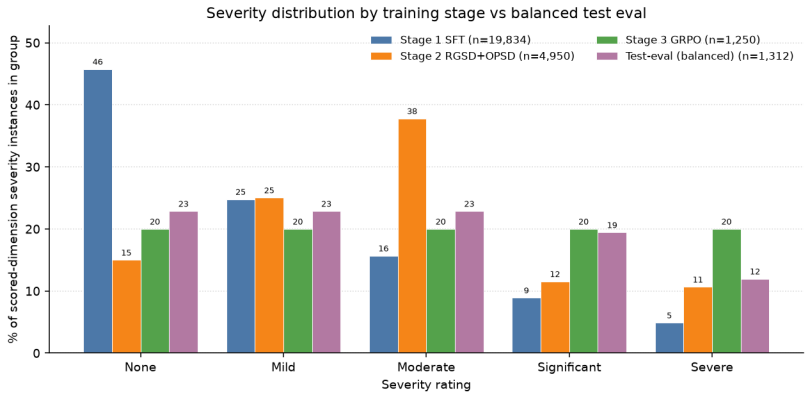

Figure 2: Severity distribution across the three training stages and the evaluation set (percentage of severity-bearing instances within each group). Stage 1 (human labels) is front-loaded on *none* (45.8%); the compounding augmentation in Stage 2 shifts mass toward the middle and roughly doubles *severe* (4.9% → 10.6%); Stage 3 GRPO is sampled uniformly across severities. The held-out test set is class-balanced (near-uniform), so the reported metrics are not dominated by the *none* majority of the raw corpus.

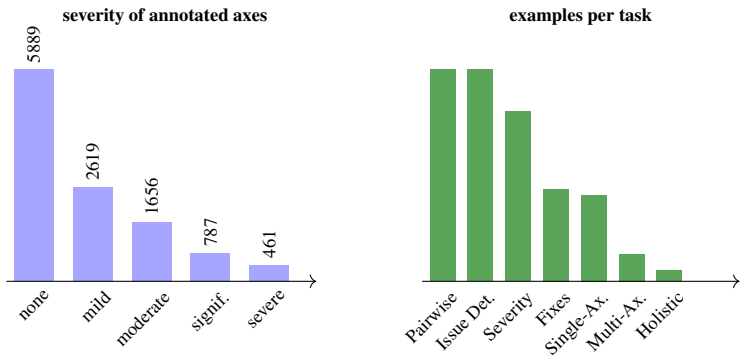

Figure 3: Corpus distributions. **Left:** most scored axes are rated *none* — 5,889 of 11,412 — which is the expected shape when annotators judge each axis on its own merits rather than condemning every axis of a recording that sounds bad overall. **Right:** examples per task in the Stage-1 SFT corpus; Pairwise and Issue Detection dominate.

order Butterworth low-pass. *Distortion*: hard clipping after a drive gain, RMS-matched. *Loudness*: attenuation to a target ITU-R BS.1770 loudness reduction. *Dropouts*: zeroing of 40 ms segments to a target zeroed fraction. *Handling pops*: a low-frequency thump plus a broadband click per pop, spread across the clip. Per-severity anchor values are in Table 13.

Level management. Every clip is first attenuated by a fixed -6 dB and only over-full-scale peaks are clipped, so speech level is constant across severities and only the injected energy rises into the headroom; this preserves each recording's natural loudness, itself a rated axis.

Compounding. An injection adds to the annotator's own severity, $s_a^{\text{final}} = \min(4, s_a^{\text{human}} + k_a)$. For the four axes with a reliable acoustic estimator (noise, loudness, bandwidth, sibilance) we *measure* the recording and drive it until the metric reaches the anchor for the final severity (grounded); the other six are additive. Axes already rated *severe* are skipped.

Combination and coverage. A severity-3/4 injection is applied to a single axis only; combinations of two or three axes are held to severities 1–2. The scheme is gated on the overall rating: recordings rated 4–5 admit up to three mild axes or one axis at any severity, rating-3 one or two mild axes, and rating-1–2 none. Every eligible recording receives all four severities on all ten axes; layered variants

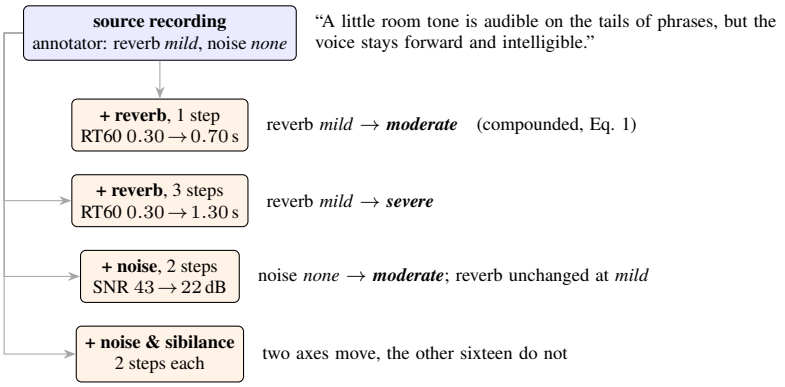

Figure 4: One recording and four of its variants. Each variant changes a named set of axes by a known number of severity steps, and the severity *compounds* onto the annotator's own rating rather than replacing it — the source is already *mild* on reverberation, so a one-step injection makes it *moderate* and a three-step injection makes it *severe*. Axes that were not touched keep their original rating, which is what lets a pair of these recordings supervise "this axis changed and those did not". Because both members of such a pair are the same speech, the difference in the injected severity is known rather than assumed.

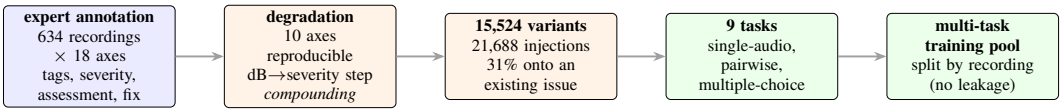

Figure 5: From expert annotation to training corpus. Each recording is annotated on 18 scored axes with four label types. Ten degradation axes, one per degradable axis, admit a synthetic degradation whose strength maps to a severity step by a fixed, published rule; the injected severity is added to the annotator's own rating for that axis, so 31% of injections land on an axis that already carried a rated issue. The resulting variants support nine tasks, including pairwise comparisons whose quality difference is known exactly. Splits are by recording, never by example: one recording contributes on the order of several hundred examples, so an example-level split would place the same audio and the same annotator's prose on both sides.

are sampled on top. Each variant is generated from a deterministic per-$(\text{clip}, \text{axis}, \text{severity}, \text{variant})$ seed, so the audio reproduces exactly.

Text regeneration. A regenerated assessment must satisfy four constraints. It describes the recording as it is and never says anything was added, injected or processed. It keeps the earlier note's voice and invents no frequency, level or tool name absent from it. It does not state the severity word, which would let the model read the label off its own input. And it does not mention a parameter setting. Generations that assert a labelled problem is absent, or drift out of English, are discarded rather than repaired.

A.8 Deriving severity anchors from the corpus

We derive the augmentation anchors from the distribution of the each axis in ProQA. For each axis, we measure the parameter across the original recordings and use the lower and upper ends of its observed range, together with the corpus mean, to define four ordered severity levels. For axes with a reliable acoustic estimator—noise, loudness, bandwidth, and sibilance—the augmentation directly targets the corresponding anchor. For the remaining axes, whose estimators are weak or confounded, we derive the operating points from the parameter range of the augmentation itself.

A.9 Task prompts and worked examples

Each task pairs a fixed instruction with the audio; targets follow the issue-first order (assessment, tags, severity, fix).

Table 12: The 18 scored axes the model produces, the number of recordings on which each carries an issue (severity $\geq$ mild, of 634), and whether a synthetic degradation exists for it.

Group	Axis	recordings w/ issue	degradable
Capture	Mic quality & clarity	455	✗
	Mic placement	319	✗
	Capture consistency over time	230	✗
	Handling & pops	322	✓
Signal fidelity	Frequency response limitations	408	✓
	Signal-chain noise: hiss, hum, buzz	336	✓
	Dropouts / glitches / artefacts	262	✓
Room / space	Reverb	302	✓
	Voice directness	272	✗
	Flutter echoes / resonances	271	✗
	Near-surface reflections	245	✗
Noise	Stationary noise	368	✓
	Dynamic noise	184	✗
Tonal balance	Overall tonal balance & EQ	504	✓
	Sibilance & harshness	275	✓
Level / dynamics	Distortion	134	✓
	Loudness & dynamic range	428	✓
	Processing issues (DSP & AI)	208	✗

Table 13: Parameter values for the four augmentation severity levels. The operating points are derived from the parameter ranges observed in PROQA; all augmentations are applied at the model's 16 kHz input rate.

Axis	Parameter	sev 1	sev 2	sev 3	sev 4
Stationary noise	SNR (dB)	29	22	15	8
Signal-chain hum	SNR (dB)	27	20	14	10
Reverberation	RT60 (s)	0.30	0.70	1.00	1.30
Sibilance	HF-shelf boost (dB)	6	10	15	20
Tonal balance	HF loss by 8 kHz (dB)	8	14	18	24
Loudness	LUFS reduction (LU)	5	10	15	20
Bandwidth	lowpass cutoff (Hz)	5500	4500	3200	2100
Distortion	clip drive ($\times$)	3	4.5	6.5	12
Dropouts	fraction zeroed	0.02	0.03	0.05	0.08
Plosives	burst level	0.8	1.3	1.9	2.5

- Single-Axis — *"... assess one axis: Mic quality & clarity ..."* $\rightarrow$ *[Mic quality & clarity] Assessment: ...frequencies above ~8 kHz are not captured ... Issue tags: Band-limited. Severity: moderate. Fixes: ... use a mic with full frequency response.*
- Multi-Axis — *"... assess these axes: Frequency Response; ... Tonal balance; ..."* $\rightarrow$ per-axis assessments for the named subset only.
- Holistic — *"... assess ALL axes ..."* $\rightarrow$ the full per-axis critique for the recording.
- Pairwise (two clips) — *"... Which has better overall audio quality, and why?"* $\rightarrow$ *A and B are the same recording; B adds plosives (mild). A is higher quality. Fix for B: high-pass ~80 Hz and de-plosive.*
- Severity — *"... rate the severity of Mic quality & clarity: none / mild / moderate / significant / severe."* $\rightarrow$ *moderate*. (A Blind variant gives no hint; an Issue-Given variant supplies the issue tags.)
- Issue Detection — *"... select all options that apply ... Muffled; Clear; Band-limited; ..."* $\rightarrow$ *Band-limited*.
- Fixes (Blind) — *"... What is your assessment, and how would you fix it?"* $\rightarrow$ assessment + remedy for the axis.

- Fixes (Issue-Given) — *"… issues: Band-limited, severity moderate. How would you fix it?"* → *… use a direct feed / a full-frequency-response microphone.*